Journal of Zhejiang University SCIENCE B
ISSN 1673-1581 (Print); ISSN 1862-1783 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Regularities in the *E. coli* promoters composition in connection with the DNA strands interaction and promoter activity

BEREZHNOY Andrey Yu[1], SHCKORBATOV Yuriy G.[†2], HISANORI Kiryu[3]

([1]*National Scientific Center, Kharkov Physical-Technical Institute, Kharkov 61108, Ukraine*)

([2]*Institute of Biology, Kharkov National University, Maidan Svobody 4, Kharkov 61077, Ukraine*)

([3]*Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Koto-ku Tokyo 135-0064, Japan*)

[†]E-mail: Yury.G.Shckorbatov@univer.kharkov.ua

Received May 11, 2006; revision accepted Sept. 26, 2006

Abstract: The energy of interaction between DNA strands in promoters is of great functional importance. Visualization of the energy of DNA strands distribution in promoter sequences was achieved. The separation of promoters in groups by their energetic properties enables evaluation of the dependence of promoter strength on the energetic properties. The analysis of groups (clusters) of promoters distributed by the energy of DNA strands interaction in −55, −35, −10 and +6 sequences indicates their connection with the transcriptional activity.

Key words: DNA sequence, Promoter strength, DNA chains interaction energy, DNA sequences classification
doi:10.1631/jzus.2006.B0969 **Document code:** A **CLC number:** Q75

INTRODUCTION

It is well known that the physical properties, namely DNA strands interaction in promoter are connected with their activity. As it was established the two promoter regions −35 and −10 are the most important for the promoter functioning, all bases in the −35 (TTGACA) and −10 (TATAAT) hexamers were highly conserved (Harley and Reynolds, 1987; Hawley and McClure, 1983). The functional importance of these sequences was proved in many investigations (Record *et al*., 1996). The open complex of RNAP and DNA includes the melted DNA in the region from −10 to +3 regardless of presence or absence of −35 element (Dombroski, 1997). RNA polymerase holoenzyme has been observed to bind single-stranded oligodeoxynucleotides bearing the −10 sequence of the nontemplate strand (Roberts and Roberts, 1996). These sites interact with different RNAP sigma-factor regions (Huang *et al*., 1997; Roberts and Roberts, 1996). If temperature affects the ability of the *E. coli* RNAP to transcribe the −10 sequence of the $nifA1_{mut3}$ promoter, then the melting of TATGGT may be an important element. Studies with *Escherichia coli* (Eco) RNAP indicated that upon RNAP binding to promoter DNA, an unstable closed complex (RP_c) is formed, followed by at least two additional intermediates (I_1 and I_2) before the initiation-competent, stable open complex (RP_o) is formed (Craig *et al*., 1998; Saecker *et al*., 2002), in which a 14-bp region of the promoter DNA, including the start site of transcription, has been melted.

$$R+P\leftrightarrow RP_c\leftrightarrow I_1\leftrightarrow I_2\leftrightarrow RP_o$$

In addition to the promoter DNA, the RNAP is thought to undergo conformational changes as well (Craig *et al*., 1998; Saecker *et al*., 2002).

In our previous work the graphic method for presentation of the energetic properties of promoter sequences was elaborated. Four promoter groups with their energetic properties were revealed. The present work is dedicated to more detailed analysis of different sequences in promoter using the promoter sites

distribution by their similarity.

MATERIALS AND METHODS

DNA sequences

We obtained 106 *Escherichia coli* promoter sequences using σ^{70} subunit from the Regulon database (©2004, CIFN/UNAM all Rights Reserved. RegulonDB DataBase V 4.0, 02-FEB-05) thanks to courtesy of Regulon database administration (Table 1). All promoter sequences were transcribed with the aid of σ^{70}. Promoter strength (the promoter ability to initiate transcription) was measured with the help of fluorescent labelling method in microarray experiments on the total transcripts of *E. coli*. Promoter strength was determined in arbitrary units reflecting the fluorescence intensity (Kanehisa *et al*., 2004; Mori *et al*., 2000). Data for promoter strength was obtained from KEGG EXPRESSION database (http://www.genome.jp/kegg/expression/) containing microarray data obtained by the Japanese research community. Orientation of promoter sequence in genome was determined as forward or reverse depending on the gene position in the genome. As far as we know forward and reverse orientation is not connected with gene functioning. The number of sequences analyzed was 106.

Computer analysis

We suggest the notion of DNA-sequence energy that is determined as a sum of energy of interaction of each nucleotide pair in promoter divided by nucleotide number. The energy of interaction of complementary nucleotides pair was determined (Kudritskaya and Danilov, 1976). The energy of AT-pair was determined as −29.33 kJ/mol and the energy of GC-pair was determined as −70.35 kJ/mol. We

Table 1 Promoters list

Promoter number	Promoter name	Promoter number	Promoter name	Promoter number	Promoter name	Promoter number	Promoter name
1	accA	28	dcuDp	55	manA	82	rho
2	accB	29	dinGp	56	menAp	83	rimL
3	accD	30	Div	57	metB	84	rnh
4	adk	31	dppA	58	metYp2	85	rplJ
5	alaS	32	drpA	59	mhpR	86	rplK
6	ampC	33	Efpp	60	mraZp	87	rplT
7	arsR	34	fimBp1	61	nanAp	88	rpoB
8	asnBp	35	folA	62	narUp1	89	rpoN
9	aspC	36	Frrp	63	nohAp	90	rpsJ
10	astCp1	37	ftsJp1	64	otsB	91	rsdp2
11	atpI	38	fxsAp	65	paaXp	92	sbcB
12	bcp	39	galRp	66	pcnBp	93	serB
13	btuB	40	gcvR	67	Pdx	94	smp
14	cdh	41	glnS	68	pfkB	95	spc
15	cedAp	42	Gnd	69	Pgi	96	str
16	cfap1	43	hemHp	70	Phe	97	sufAp
17	clpAp1	44	hemN	71	pheS	98	tesA
18	cmk	45	hepAp	72	pntA	99	thrA
19	corA	46	hisA	73	ppc	100	tufB
20	creA	47	hisB	74	pthp	101	ung
21	cusCp	48	Hiss	75	ptr	102	upp
22	cusRp	49	hscB	76	purA	103	xseBp
23	cutA	50	katE	77	putA	104	ybjCp
24	cysE	51	lacI	78	pyrEp2	105	yeiKp
25	dapA	52	Lep	79	pyrF	106	yhcA
26	dapB	53	Lpp	80	relA		
27	dapD	54	lysP	81	rep		

determined the mean energy of interaction in complementary nucleotides by means of "sliding-window" method. The mean energy of interaction was attributed to the nucleotide pair in the center of the 10 nucleotides window which shifted from left to right with the Step 1 nucleotide.

The process of division on clusters consists of the following stages:

1. We obtained the mean sequence of all promoter sequences sets. The procedure includes determination of the most frequently occurring nucleotide on each position of the sequence by simply comparing nucleotides in all of the analyzed sequences. Such a resulting sequence we call "mean sequence".

2. We determined the quantity of nucleotides differences of each analyzed sequence with the "mean sequence".

3. We associated the quantity of differences of each sequence and put all the analyzed sequences in order according to the quantity of their differences from the "mean sequence".

4. We found the difference between the maximal and minimal number of differences from the mean sequence.

5. We divided the number obtained by 9, which is the clusters size.

6. We consequently assigned to each cluster all the sequences whose differences are within the limits of cluster. If the quantity of elements in every cluster was less than 3 we repeated the division on clusters once more with the number of clusters decreased by one. The resulting number of clusters for different sequences sets was different and equaled 5 to 4.

Computer programs for obtaining energy distribution according to nucleotide position and for arrangement of sequences into groups were elaborated on by Berezhnoy and Shckorbatov (2005).

RESULTS

In our previous work (Berezhnoy and Shckorbatov, 2005), the nucleotide interaction's energy distribution as analyzed by our promoters looks as shown in Fig.1 indicating that the energy minima are connected with the four areas (−55, −35, −10 and +6). The sequences located at −35, −10 and +6 positions are often analyzed (Lewin, 2004). We propose additionally to analyze −55 sequence (10 nucleotides). We analyzed in the current article the nucleotide interaction energy distribution in all of these areas and in the whole promoter. We determined the mean (consensus) distribution of the nucleotide interaction energy via promoter. With cluster analysis using method for determining the Euclidian distances, we sorted all 106 promoters on four clusters. The results obtained are presented in Table 2. All the clusters significantly differ in the energy distribution through the promoter length. We determined the energy distribution throughout the promoters, −35 sequences, −10 sequences and +6 sequences. We determined five groups in these sequences based on the similarity of energy distribution in the above sequences. The dependences of mean strength of promoters in each cluster on the mean energy of the same cluster are presented in the Fig.2. The points in these figures represent the clusters of corresponding promoter sequences. Our data indicate the absence of linear dependence between these parameters. The functions describing these dependences are presented in Table 3. As one can see from the Fig.2, the dependences of mean strength in clusters on the energy of DNA chains interaction in the whole promoter, −35 sequence and +6 sequence are very similar and have parabolic shape. It means that the strongest promoters have mean energy values and that the most energetic promoters have minimal strength. Dependence of the mean clusters energy on the energy of DNA chain interaction in −55 sequence has relatively more complex character revealing two maxima of promoter strength. The promoters arranged by the energy of

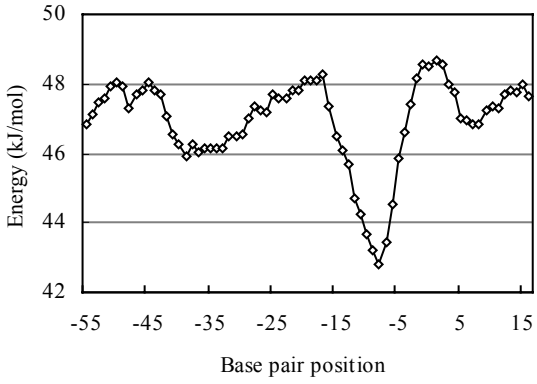

Fig.1 Dependence of mean energy of interaction between complementary DNA nucleotides on the mean window center position in promoter sequences

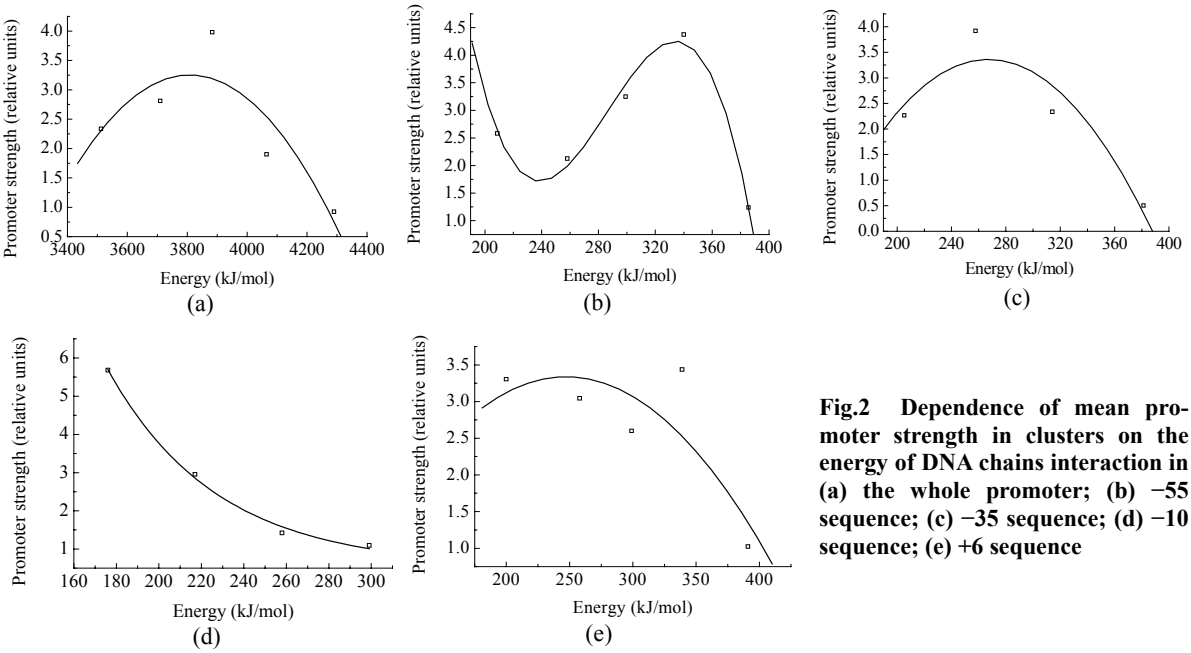

Fig.2 Dependence of mean promoter strength in clusters on the energy of DNA chains interaction in (a) the whole promoter; (b) −55 sequence; (c) −35 sequence; (d) −10 sequence; (e) +6 sequence

Table 2 Cluster's composition in the whole promoter and in its parts (−55, −35, −10 and +6 sequences)

Promoter site	Cluster's number	Numbers of promoters in cluster
Whole promoter	1	2, 10, 16, 17, 20, 28, 40, 41, 42, 53, 59, 61, 72, 76, 83, 99, 104
	2	7, 8, 9, 11, 12, 21, 26, 29, 32, 34, 36, 37, 38, 43, 46, 50, 55, 56, 57, 62, 66, 68, 69, 71, 75, 77, 80, 84, 86, 87, 91, 93, 100, 101, 102, 105
	3	3, 4, 5, 13, 14, 22, 24, 25, 27, 30, 31, 33, 35, 39, 44, 47, 52, 54, 58, 67, 70, 73, 74, 85, 88, 89, 90, 92, 94, 95, 96, 97, 103
	4	1, 6, 15, 18, 19, 45, 48, 51, 60, 63, 64, 65, 78, 79, 81, 82, 106
	5	23, 49, 98
−55 sequence	1	8, 11, 16, 22, 25, 26, 28, 34, 41, 43, 46, 51, 59, 61, 62, 69, 72, 91, 93, 100, 101, 102, 104, 105, 106
	2	12, 13, 17, 18, 21, 29, 30, 33, 36, 38, 42, 50, 57, 58, 67, 68, 71, 73, 74, 75, 77
	3	2, 4, 6, 7, 15, 19, 20, 27, 31, 35, 40, 44, 45, 49, 52, 53, 54, 55, 56, 64, 76, 78, 80, 83, 84, 86, 94, 95, 96, 97, 98, 103
	4	3, 5, 9, 24, 32, 37, 39, 48, 60, 63, 66, 70, 81, 82, 85, 88, 89, 90, 92
	5	1, 10, 14, 23, 47, 65, 79, 87, 99
−35 sequence	1	2, 3, 8, 10, 11, 14, 16, 28, 32, 36, 37, 38, 40, 42, 44, 47, 59, 61, 62, 78, 79, 80, 88, 92, 101, 105
	2	1, 5, 6, 7, 9, 12, 18, 20, 21, 23, 25, 27, 33, 39, 41, 51, 53, 55, 56, 58, 60, 65, 66, 69, 70, 71, 72, 74, 76, 77, 84, 85, 87, 94, 96, 97, 99, 100, 103, 104
	3	4, 13, 15, 17, 19, 22, 24, 29, 30, 31, 34, 35, 43, 45, 46, 48, 49, 50, 52, 54, 57, 63, 64, 67, 68, 73, 75, 81, 82, 83, 86, 89, 90, 91, 93, 95, 98, 102, 106
	4	26
−10 sequence	1	10, 11, 18, 33, 34, 36, 37, 55, 56, 58, 59, 61, 85, 90, 95, 96, 98, 102
	2	1, 3, 5, 6, 15, 16, 17, 19, 20, 22, 23, 27, 28, 32, 35, 38, 42, 43, 44, 46, 47, 48, 49, 50, 52, 53, 54, 60, 63, 64, 65, 66, 67, 68, 69, 71, 72, 73, 74, 75, 76, 78, 79, 80, 81, 82, 83, 86, 87, 88, 92, 94, 97, 100
	3	2, 4, 12, 13, 21, 25, 26, 29, 30, 31, 39, 40, 41, 45, 51, 57, 70, 77, 84, 91, 99, 101, 103, 104, 105, 106
	4	7, 8, 9, 14, 24, 62, 89, 93
+6 sequence	1	1, 4, 11, 12, 16, 21, 30, 32, 34, 35, 50, 54, 55, 56, 61, 66, 70, 86, 91, 96, 102, 103
	2	2, 5, 8, 10, 18, 19, 26, 27, 33, 37, 41, 42, 43, 45, 46, 52, 57, 58, 59, 60, 63, 65, 67, 68, 69, 73, 74, 77, 82, 83, 84, 85, 89, 95, 104
	3	3, 6, 9, 14, 20, 22, 25, 28, 38, 39, 47, 51, 53, 64, 71, 72, 76, 79, 88, 90, 92, 93, 94, 97, 106
	4	7, 13, 17, 23, 24, 36, 48, 75, 78, 80, 81, 87, 98, 99, 100, 101
	5	15, 29, 31, 40, 44, 49, 62, 105

Table 3 Characteristics of promoter strength regression on the energy of DNA chains interaction in different promoter sites

Promoter site	Promoter strength regression dependence on the promoter site energy	Regression quotient
Whole promoter	$y=-152.5+0.343x-1.89\times10^{-4}x^2$	0.876
−55 sequence	$y=-128-5.82x+8.76\times10^{-2}x^2-4.27\times10^{-4}x^3$	0.991
−35 sequence	$y=-13.4+0.525x-4.11\times10^{-3}x^2$	0.942
−10 sequence	$y=31-0.866x+6.26\times10^{-3}x^2$	1.000
+6 sequence	$y=-2.53+0.198x-1.68\times10^{-3}x^2$	0.813

−10 sequence have exponential dependence of their strength on the energies of −10 sequence.

CONCLUSIONS

1. We improved the original methods of computer analysis (Berezhnoy and Shckorbatov, 2005) enabling visualization of the energy distribution in promoter sequences and separation of the promoters in groups by their energetic properties, thus enabling us to judge the dependence of promoter strength on their energetic properties.

2. We revealed that −55 sequence composition is connected with promoter transcriptional activity.

3. The analysis of groups (clusters) of promoters distributed by the energy of DNA strands interaction in −55, −35, −10 and +6 sequences indicates their connection with the transcriptional activity.

4. The dependence of mean promoter strength in clusters has at least one maximum in the zone of mean energies of clusters in the cases of −55, −35 and +6 sequences but in the case of −10 sequence the dependence is monotonous.

References

Berezhnoy, A.Y., Shckorbatov, Y.G., 2005. Dependence of *E. coli* promoter strength and physical parameters upon the nucleotide sequence. *J. Zhejiang University Sci. B*, **6**(11): 1063-1068. [doi:10.1631/jzus.2005.B1063]

Craig, M.L., Tsodikov, O.V., McQuade, K.L., Schlax, P.E.Jr, Capp, M.W., Saecker, R.M., Record, M.T.Jr, 1998. DNA footprints of the two kinetically significant intermediates in formation of an RNA polymerase-promoter open complex: evidence that interactions with start site and downstream DNA induce sequential conformational changes in polymerase and DNA. *J. Mol. Biol.*, **283**(4): 741-756. [doi:10.1006/jmbi.1998.2129]

Dombroski, A.J., 1997. Recognition of the −10 promoter sequence by a partial polypeptide of σ^{70} in vitro. *J. Biol. Chem.*, **272**:3487-3494.

Harley, C.B., Reynolds, R.P., 1987. Analysis of *E. coli* promoter sequences. *Nucleic Acids Res.*, **15**:2343-2361.

Hawley, D.K., McClure, W.R., 1983. Compilation and analysis of *Escherichia coli* promoter sequences. *Nucleic Acids Res.*, **11**:2237-2255.

Huang, X., Lopez de Saro, F.J., Helmann, J.D., 1997. Sigma factor mutations affecting the sequence-selective interaction of RNA polymerase with −10 region single-stranded DNA. *Nucleic Acids Res.*, **25**(13):2603-2609. [doi:10.1093/nar/25.13.2603]

Kanehisa, M., Goto, S., Kawashima, S., Kuno, Y., Hattori, M., 2004. The KEGG resources for deciphering the genome. *Nucleic Acids Res.*, **32**(90001):D277-D280. [doi:10.1093/nar/gkh063]

Kudritskaya, Z.G., Danilov, V.I., 1976. Quantum mechanical study of bases interactions in various associates in atomic dipole approximation. *J. Theor. Biol.*, **59**(2):303-318. [doi:10.1016/0022-5193(76)90172-7]

Lewin, B., 2004. Genes-VIII. Pearson Prenice Hall, New York.

Mori, H., Isono, K., Horiuchi, T., Miki, T., 2000. Functional genomics of *Escherichia coli* in Japan. *Res. Microbiol.*, **151**(2):121-128. [doi:10.1016/S0923-2508(00)00119-4]

Record, M.T.Jr, Reznikoff, W.S., Craig, M.L., McQuade, K.L., Schlaz, P.J., 1996. *Escherichia coli* RNA Polymerase ($E\sigma^{70}$), Promoters, and the Kinetics of the Steps of Transcription Initiation. *In*: Neidhardt, F.C., Curtiss III, R., Ingraham, J.L., Lin, E.C.C., Low, K.R., Magasanik, B., Reznikoff, W.S., Riley, M., Schaechter, M., Umbarger, H.E. (Eds.), *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology, 2nd Ed. ASM Press, Washington DC, p.792-820.

Roberts, C.W., Roberts, J.W., 1996. Base-specific recognition of the nontemplate strand of promoter DNA by *E. coli* RNA polymerase. *Cell*, **86**(3):495-501. [doi:10.1016/S0092-8674(00)80122-1]

Saecker, R.M., Tsodikov, O.V., McQuade, K.L., Schlax, P.E.Jr, Capp, M.W., Record, M.T.Jr, 2002. Kinetic studies and structural models of the association of *E. coli* sigma (70) RNA polymerase with the lambdaP (R) promoter: large scale conformational changes in forming the kinetically significant intermediates. *J. Mol. Biol.*, **319**(3): 649-671. [doi:10.1016/S0022-2836(02)00293-0]