Biophysical Chemistry, 44 (1992) 61–70
Elsevier Science Publishers B.V., Amsterdam

On the base sequences of the promoters in transcription initiation

T. Shinoda

National Chemical Laboratory for Industry, Tsukuba Research Center, Tsukuba, Ibaraki 305 (Japan)

(Received 13 August 1991; accepted in revised form 12 March 1992)

Abstract

The base sequence of a specific DNA region identified as the promoter is investigated by means of the quantity $\tilde{S}_r$ corresponding to "superdelocalizability" of oxygen ion of each phosphate for the ten DNA dimer units ($[XY/Y'X']^{2-}$) and ($[XY/Y'X']^{2-} + H^+$)-complexes. A mechanism is proposed of how RNA polymerase can recognize its transcription site (phosphate), and is applied to the *Escherichia coli* promoters, lacUV5, recAp, rrnEpl, and rrnEp2. The result explains fairly well the character of the promoters experimentally found.

Keywords: Mechanism of transcription initiation; Base sequences of promoters; *Escherichia coli* promoters

1. Introduction

In many studies on DNA, it has been assumed that the chemical reactivities of all phosphates are equivalent. This assumption, however, has not been made out of conviction so far. We previously attempted to investigate in detail the electronic structure of the ten DNA dimer-units in both A and B conformations by DV-X_α cluster calculations [1]. The result was that the level structures of the O2, O3 of each phosphate, differ remarkably depending on DNA base sequences and that these differences enable us to easily recognize the type of stacked base pair in the $5' \rightarrow 3'$ direction. There are eight non-degenerate levels just below the Fermi level, which are made up mainly of the O2 and O3 $2p$ orbitals, as described in our previous paper [1]. From the level structures, the quantity $\tilde{S}_r$ corresponding to "superdelocalizability" [2], which has been used for explaining the reactivities in the frontier electron method, was computed for each phosphate. The results are summarized in Table 1 (taken from Ref. [1]). DNA dimer-unit is briefly represented as $XY/Y'X'$ (for atomic site labeling, see Fig. 1a), where XY denotes a stacked pair with a base X followed by a base Y in the $5' \rightarrow 3'$ direction; X' and Y' are the hydrogen-bonded base pairs of Y and X, respectively. The phosphate of the XY strand is designated as phosphate-1 (P_1) and that of the $X'Y'$ strand as phosphate-2 (P_2). The numbering of the O2, O3 atoms is common standard [3]. As shown in Table 1, the ten DNA dimer-units can be grouped into two classes: One for whose phosphates $\tilde{S}_r(P_1) \neq \tilde{S}_r(P_2)$, and the other for whose phosphates $\tilde{S}_r(P_1) = \tilde{S}_r(P_2)$. For the first group of dimer-units, we

Correspondence to: Dr. T. Shinoda, National Chemical Laboratory for Industry, Tsukuba Research Center, Tsukuba, Ibaraki 305 (Japan)

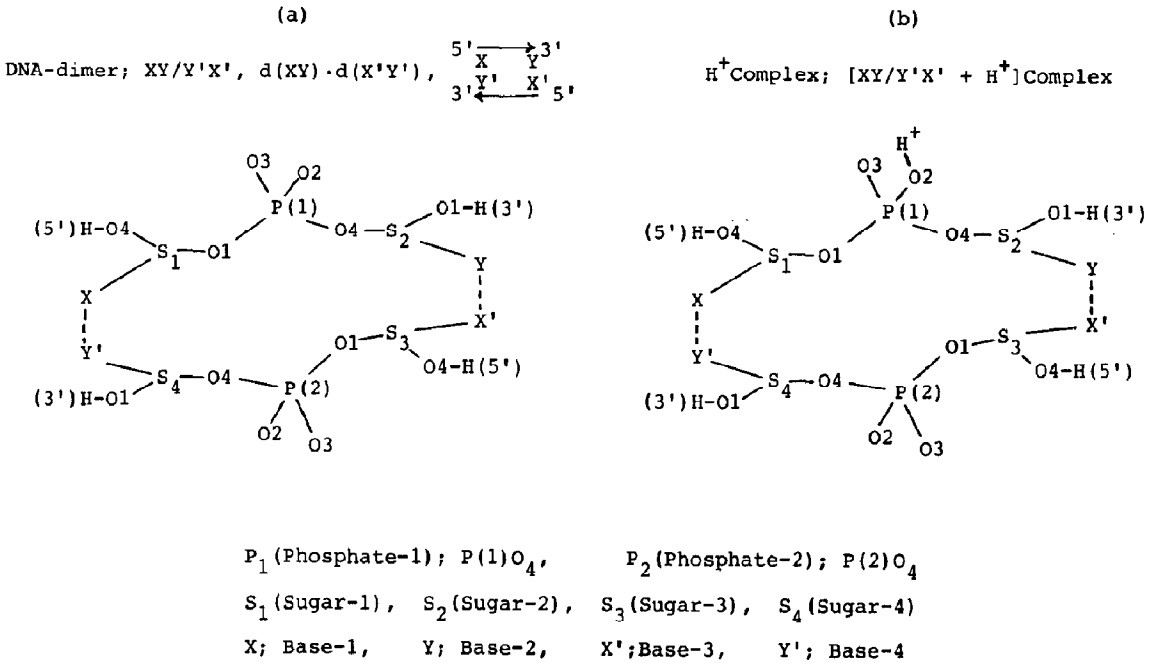

Fig. 1. Structure of DNA dimer-unit, (XX/Y'X') (a), and [XY/Y'X' + H^+]-complex (b).

Table 1

Quantity $\tilde{S}_r$ corresponding to "superdelocalizability" (a measure of chemical reactivity) of O2 and O3 of each phosphate in the ten DNA dimer-units (XY/Y'X') for both A and B forms and in the six [(XY/Y'X')+H^+]-complexes for the B form (in parentheses)

Dimer-unit	Phosphate-1(P_1)		Phosphate-2(P_2)	
	O2	O3	O2	O3
AA/TT(A)	174.8	188.4	15.326	2.036
TA/AT(A)	7.964	6.025	7.964	6.025
AT/TA(A)	4.911	3.877	4.911	3.877
GG/CC(A)	3434	1116	4.300	2.633
CG/GC(A)	5.519	4.780	5.519	4.780
GC/CG(A)	5.277	4.275	5.277	4.275
AG/TC(A)	105.0	26.225	3.476	3.442
TG/AC(A)	14.706	5.654	3.915	4.034
AC/TG(A)	7.423	3.449	4.901	4.576
TC/AG(A)	5.780	2.958	19.731	13.847
AA/TT(B)	7.884	4.659	4.863 (9.101)	2.936 (5.952)
TA/AT(B)	6.510	4.455	6.510	4.455
AT/TA(B)	4.605	3.492	4.605	3.492
GG/CC(B)	26.414	10.478	2.846 (8.676)	1.936 (4.203)
CG/GC(B)	4.373	3.690	4.373	3.690
GC/CG(B)	4.034	3.535	4.034	3.535
AG/TC(B)	33.658	5.567	2.837 (5.880)	1.969 (4.710)
TG/AC(B)	22.576	4.527	3.283 (39.644)	2.295 (40.484)
AC/TG(B)	6.084	2.885	4.375 (134.0)	2.628 (142.0)
TC/AG(B)	6.314 (7.021)	2.932 (4.122)	11.512	4.771

computed the $\tilde{S}_r$ value of a phosphate not coordinating in the $[XY/Y'X' + H^+]$-complex, where H^+ coordinates to the phosphate with a larger $\tilde{S}_r$ value is shown in Fig. 1(b), to investigate the stability after coordination. The obtained values for B-form units are indicated between parentheses in Table 1. The details shall be described elsewhere [4].

In this paper we apply these results to base sequences of the promoters in transcription initi-

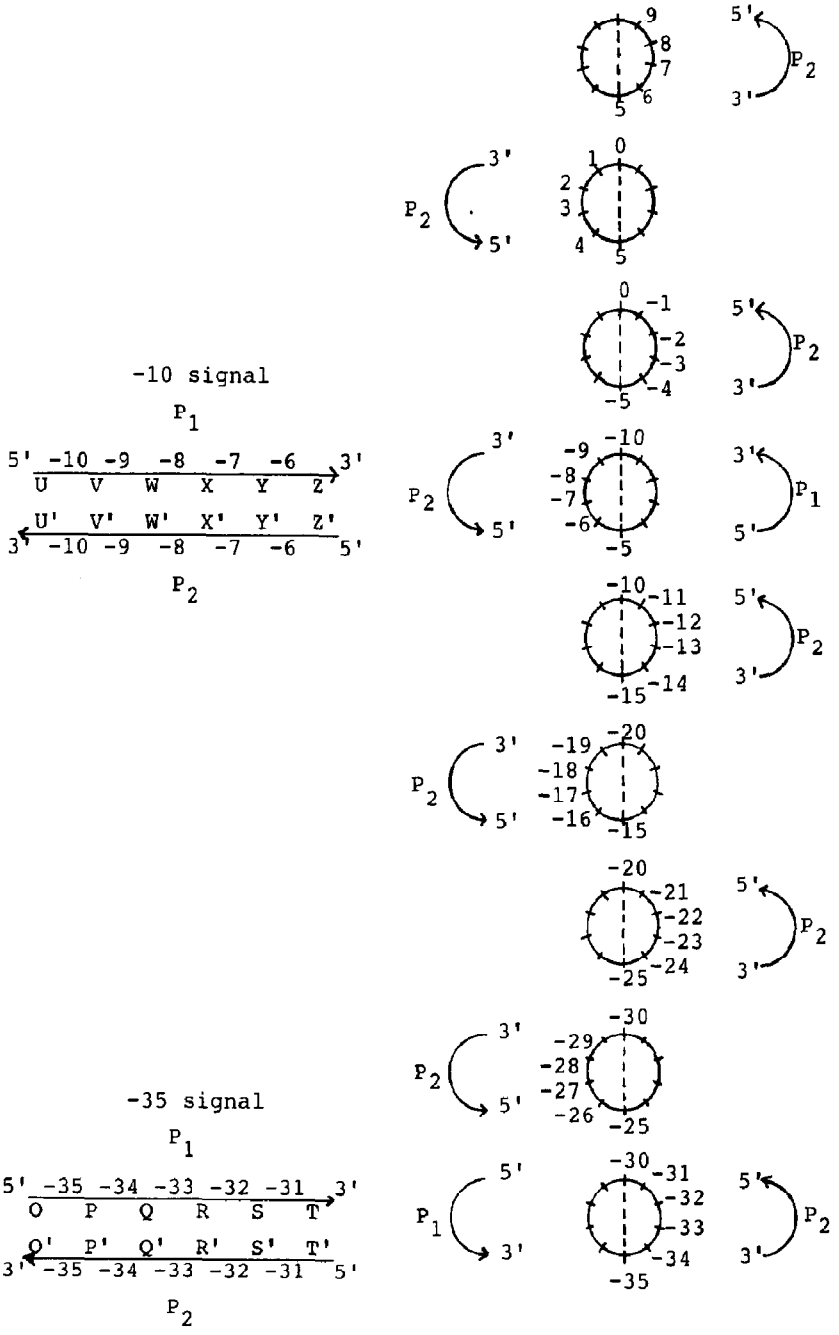

Fig. 2. Schematic representation of phosphates of a promoter.

ation, and propose a mechanism of how RNA polymerases recognize the transcription sites (phosphates). A reasonable interpretation is given for the base sequences of TATAAT and TTGACA, which have been believed experimentally to be the most ideal sequence for the −10 and −35 signals, respectively.

2. Proposition of a mechanism

From Table 1, it is seen that a positive ion (or part) does not coordinate to the O3 but rather to the O2 ($\tilde{S}_r(O2) > \tilde{S}_r(O3)$). Using the $\tilde{S}_r$ values of the O2 for the DNA dimer-units ($[XY/Y'X']^{2-}$) in the B-form, we investigated in detail the base sequences of specific DNA regions (−10 and −35 signals) in various *E. coli* promoters. Taking into account our findings, we propose a mechanism of how the RNA polymerase binds to the phosphate of the transcription site, shown schematically in Fig. 2. In Fig. 2, circles represent sectional planes perpendicular to the axis of DNA double helix and display the arrangement of the phosphates in one cycle (one unit) of the helix. The numbers in Fig. 2 indicate phosphate in the 2-positions (P_2) of the helix that is transcribed. At first the RNA polymerase coordinate to the −35 signal in an asymmetric distribution (Process-1), because the base sequence of the −35 signal contains characteristically more than one dimer unit, where the $\tilde{S}_r$ value of one phosphate of a pair (P_1, P_2) is very large and that of the other phosphate is very small. This asymmetric group of RNA polymerases slides to the −10 signal along the phosphates which lie in parallel to the axis of the DNA double helix, and arranges in a higher symmetric and more stable distribution at the −10 signal (Process-2). The base sequence of the −10 signal, such as TATAAT, consists characteristically of the dimer units with $\tilde{S}_r(P_1) = \tilde{S}_r(P_2)$ and a unit with $\tilde{S}_r(P_1) \neq \tilde{S}_r(P_2)$ for each pair. The latter is necessary for the group of RNA polymerases to recognize the direction of transcription. If the base sequence of the −10 signal is, for example, TATATA, it is impossible for the group of RNA polymerases to recognize the direction of transcription. The $\tilde{S}_r$ values of the dimer units with $\tilde{S}_r(P_1) = \tilde{S}_r(P_2)$ are better to be large for strong promoters. The $\tilde{S}_r$ value of the CG and GC units is smaller than that of the TA and AT units, as shown in Table 1. Therefore, we may say that the base sequence of TATAAT gives the best −10 signal for the strong promoters. Secondly, the higher symmetric group of RNA polymerases on the P_2 side coordinate to each located at the 0, 1, 2, 3, and 4 positions. On the other hand, the RNA polymerases on the P_1 side slide again half a cycle and coordinate to each P_2 located at the 5, 6, 7, 8, and 9 positions (Process-3). In this process, the RNA polymerase that coordinates to the dimer unit with $\tilde{S}_r(P_1) \neq S_r(P_2)$, becomes free and instead the RNA polymerase that coordinates to the TA or AT dimer unit on the promoter except the −10 sequence, slides and carries out Process-3. It is necessary for the strong promoter that the position of the TA or AT dimer unit on the circumference of the helix, indicated by the −35 signal in Fig. 2, is just consistent with that of the dimer unit with $\tilde{S}_r(P_1) \neq \tilde{S}_r(P_2)$, the −10 signal. Finally, each RNA polymerase starts to bind to each phosphate of the transcription site from 0 to 100 positions of the P_2 (ten turns of the helix), since one RNA polymerase corresponds to about 100 hydrogen bonded base pairs. In this process (Process-4), the first RNA polymerase docks at the 0-position and binds to the 100 transcription sites. The group of the other nine RNA polymerase slides together with the first and continue to one-by-one bind to the P_2 of the remaining transcription sites until all positions are occupied. This results in a thousand transcription sites (P_2) perfectly bonded with ten RNA polymerases. Since the higher symmetry of the electrostatic fields around the P ions can be retained during these processes, such a collective moving of RNA polymerases as above mentioned may be considered to occur spontaneously.

3. Base sequences of TTGACA and TATAAT

The base sequence of TTGACA is believed to be the most ideal −35 sequence. As mentioned in Section 2, the dominant role of the −35 signal is considered to have RNA polymerases collect

and coordinate around promoters as soon as possible (Process-1). Therefore, the base sequence of the −35 signal should preferably consist of DNA dimer-units with large values of $\tilde{S}_r$. However, as can be seen from Table 1, for many dimer-units (such as AG/TC, GC/CC and TG/AC) $\tilde{S}_r(P_1) > \tilde{S}_r(P_2)$. Even if a dimer-unit can promote RNA polymerase to coordinate to a P_1-site, its RNA polymerase may be freed when another RNA polymerase does not coordinate to the complementary P_2-site for the sake of stability. We must also take into account the $\tilde{S}_r(P_2)$ value after coordination to P_1. The six dimer-units for which $\tilde{S}_r(P_1) \neq \tilde{S}_r(P_2)$, can be investigated by using the $\tilde{S}_r$ values indicated in parentheses of Table 1.

From Table 1, it is found that the $\tilde{S}_r(P_2)$ of the TG/AC-unit steeply increases, whereas those of the AG/TC and GC/CC increase only slightly. On the other hand, for the AC/TG-unit, although the very large increase of $\tilde{S}_r(P_2)$ upon coordination, the $\tilde{S}_r(P_1)$ itself is a small value as compared with those of the AG/TC, GG/CC, TG/AC. As a result, we may conclude that the TG/AC-unit is the best candidate for strong −35 signals among the ten dimer-units tested. The scheme for TTGACA, which contains two TG/AC-units, is displayed in Fig. 3(a). The $\tilde{S}_r$ values for the $[XY/Y'X']^{2-}$ and $([XY/Y'X']^{2-} + H^+)$-complexes are indicated by the thick and the thin line, respectively. The length of each line is proportional to the $\tilde{S}_r$ value. This feature also corresponds well to the strength of the promoters.

It is explained in Section 2 that the base sequence of TATAAT [1] yields the best −10 signal

[1] In the strict sense of the word, this is a sequence which consists of two TA, two AT and one AA/TT dimer-unit.

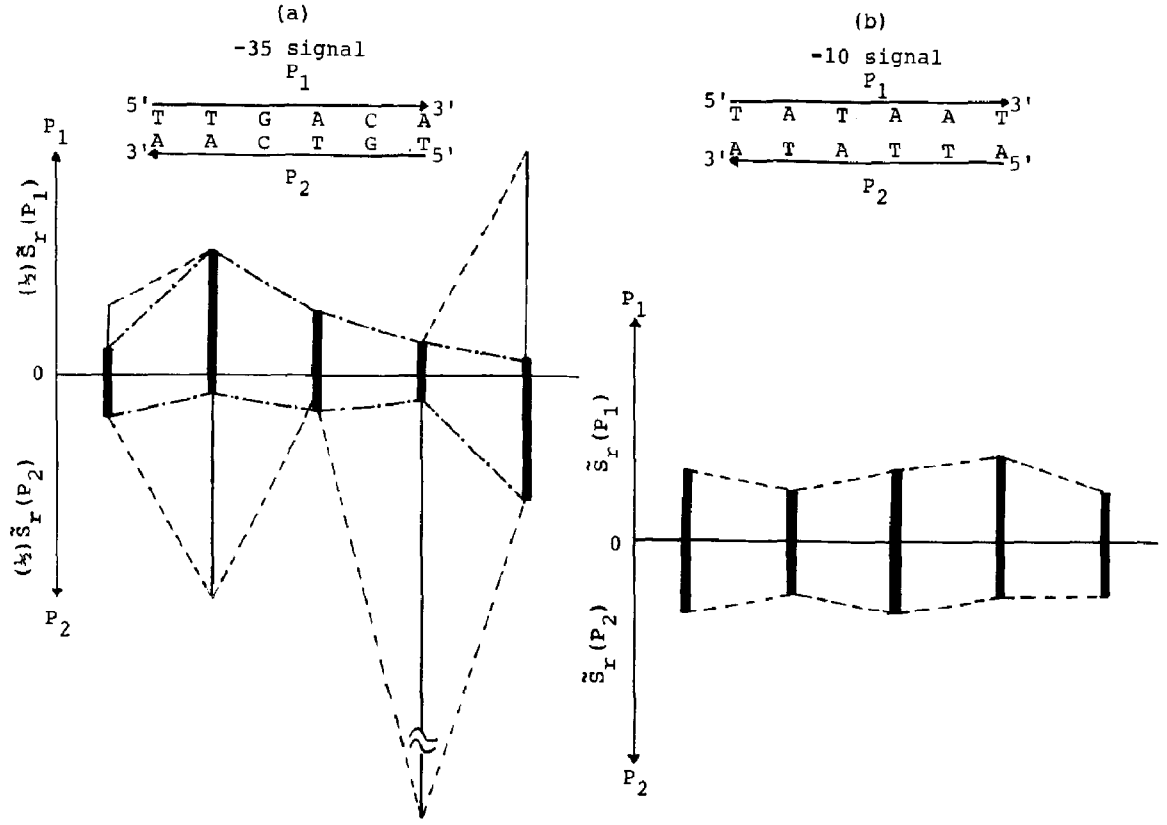

Fig. 3. Scheme of $\tilde{S}_r$ value of each phosphate in (a) TTGACA (−35 signal) and (b) TATAAT (−10 signal).

for the strong promoters. Therefore, even if a CGCG-box docks in the promoter region, the TATAAT sequence can be determined unambiguously as the −10 signal when there is a TATAAT sequence present. The feature is well displayed in Fig. 3(b) by using the $\tilde{S}_r$ values for the $[XY/Y'X']^{2-}$ complex given in Table 1.

4. Application to *E. coli* promoters

We attempted to apply the method to *E. coli* promoters, lacUV5, recAp, rrnEpl, rrnEp2, and to explain the character of the promoters experimentally found by Kajitani and Ishihama [5]. They explained the obtained experimental results by using the so-called Parameters I and II, which were proposed by themselves. Reaction was assumed to take place in two consecutive steps. The two Parameters I, II were designated as the promoter strength of each step. However, the steps are neither more realistic or fine than the processes presently proposed in Section 2.

We employ the base sequences of each promoter which were given in Fig. 5 of Ref. [6]. To detect a correlation between the −10 and the −35 signals, the $\tilde{S}_r$ value of each phosphate pair (P_1, P_2) for the four promoters have been plotted on the circumferences of both signals in Fig. 4 in a manner as described in Section 2. The $\tilde{S}_r$ value of each dimer-unit used is that of the O2 atom of each phosphate in the ten DNA dimer-units $(XY/Y'X')$ of the B-form, as given in Table 1. The length of line on the circumference is proportionate to the $\tilde{S}_r$ value. The lines of the dimer-unit with $\tilde{S}_r(P_1) = \tilde{S}_r(P_2)$ are marked by the thick line. From Fig. 4, it is found that all four promoters are equivalent on the conditions of strong promoters as for the −10 signal, but are different in the −35 signal from each other. The −35 signal of lacUV5 contains the TA-unit at the same position on the circumference as the AA-unit of the −10 signal, whereas those of the other promoters contain the dimer-units with $\tilde{S}_r(P_1) \neq \tilde{S}_r(P_2)$ and no dimer-unit with $\tilde{S}_r(P_1) = \tilde{S}_r(P_2)$ at the position of the −10 signal. Therefore, it suggests that lacUV5 is the highest in the rate of Process-3 among these promoters. All the −35 signals of the four promoters fit equally well with the conditions of Process-1, since there is a

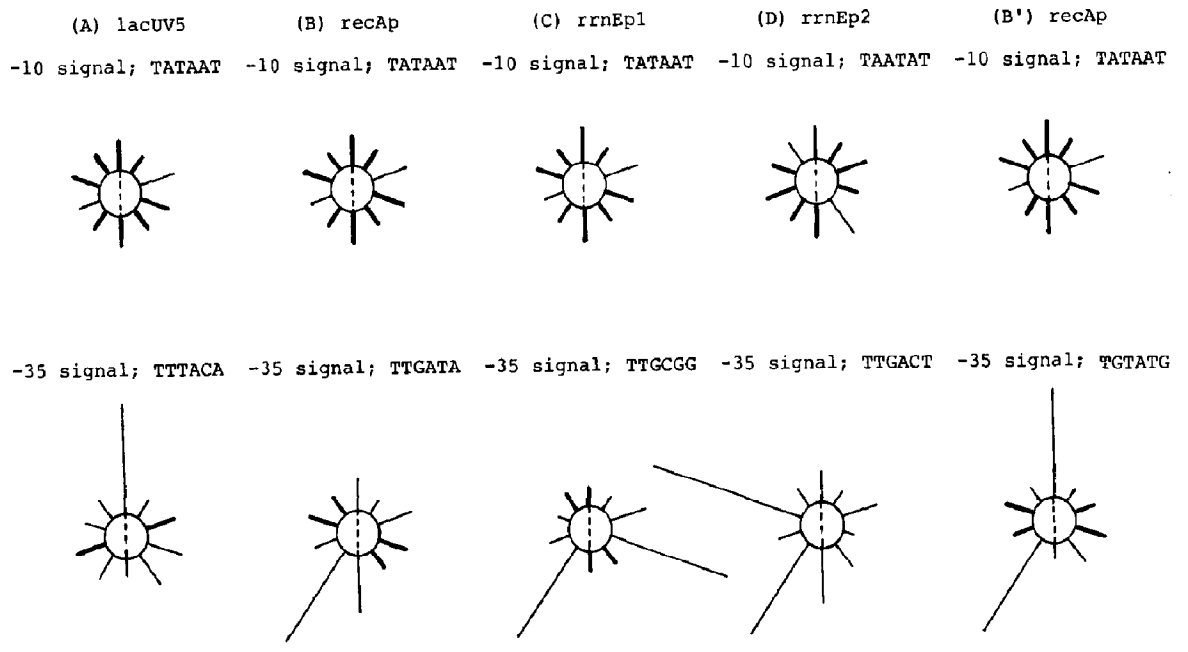

Fig. 4. $\tilde{S}_r$ value of each phosphate in the −10 and the −35 signals for four *E. coli* promoters: (A) lacUV5; (B) recAp; (C) rrnEp1; (D) rrnEp2. Shown in Fig. 4(B') is that for the −35 sequence which is proposed to be the more reasonable one of recAp.

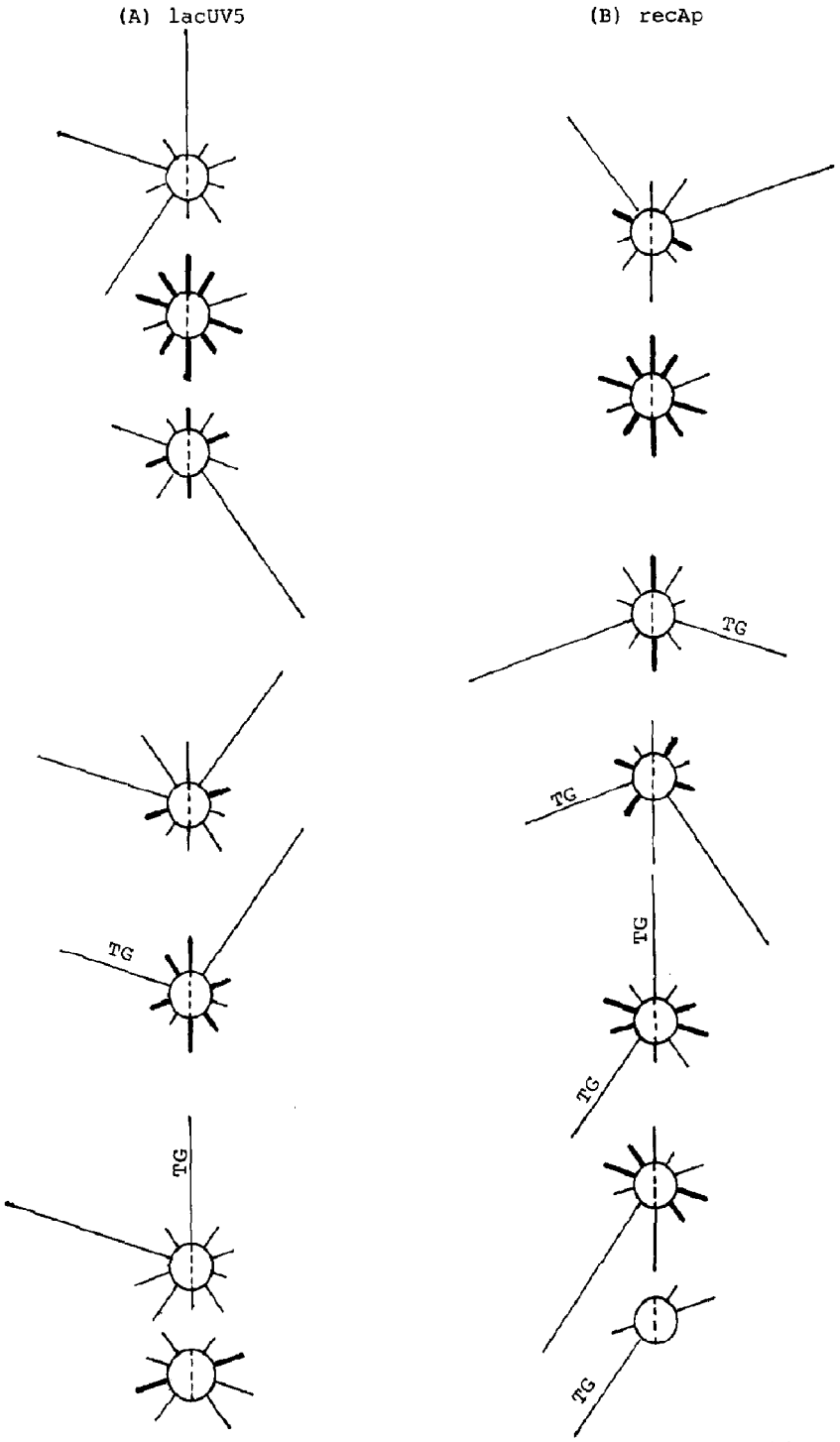

Fig. 5. $\tilde{S}_r$ value of each phosphate in the whole promoter region for four *E. coli* promoters: (A) lacUV5; (B) recAp; (C) rrnEp1; (D) rrnEp2. The style and method are much the same as shown in Figs. 2 and 4.

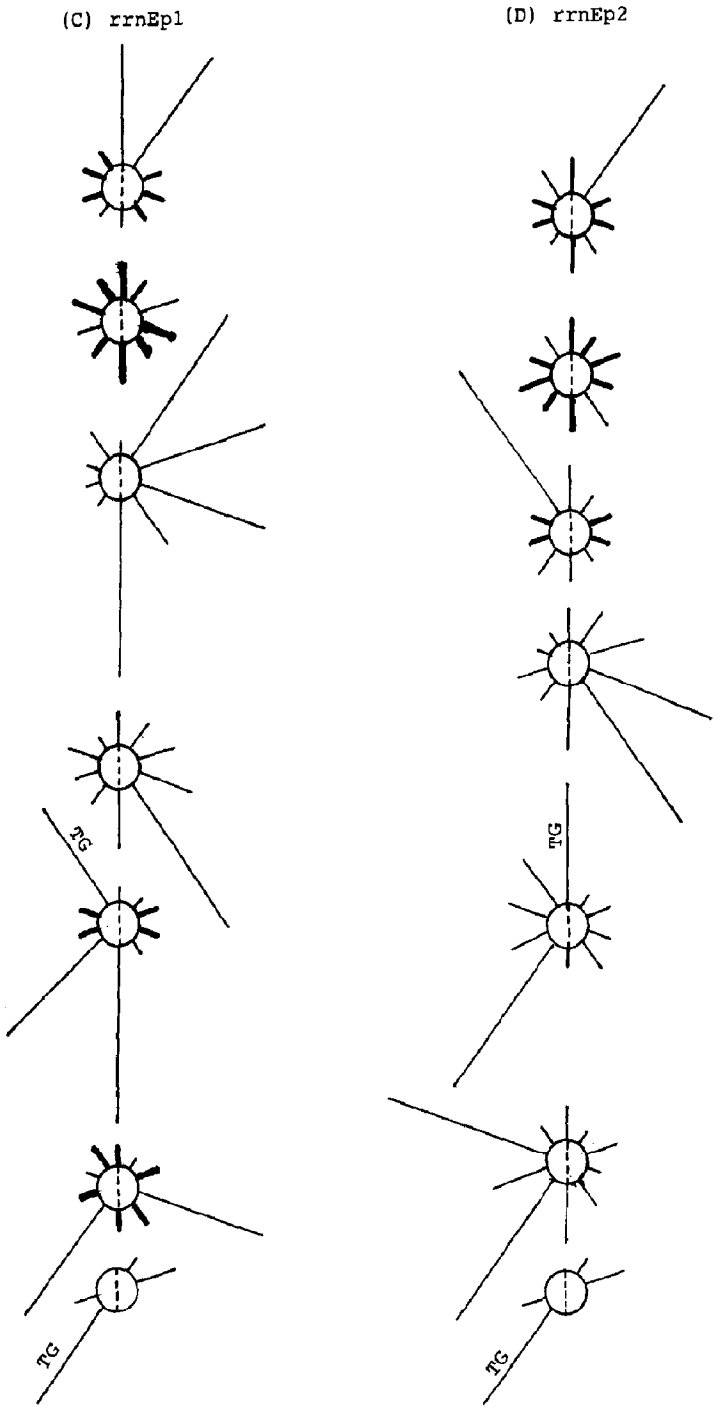

Fig. 5. (continued).

TG/AC-unit in the −35 sequence. In Process-1, a very asymmetric distribution of RNA polymerases collected may be released at once. Taking into account both $\tilde{S}_r$-values for the [XY/Y′X′] and the [XY/Y′X′ + H⁺] = complexes, given in Table 1, the order of collection of RNA polymerases may be AG/TC > GG/CC > TG/AC, whereas that of the stability of coordinated RNA polymerases in ranked TG/AC > GG/CC ≧ AG/TC. The investigation of all of the base sequences by taking into account those between the −10 and −35 signals may be necessary for an interpretation of the difference in the promoter strength among recAp, rrnEpl, and rrnEp2. Figure 5 schematically shows the result by the scheme as displayed in Fig. 2 together with that of lacUV5. In Fig. 5(A), one can find the TA-unit at the −32 position firstly as slided from the −7 position (AA/TT-unit of the −10 signal of lacUV5) to the −35 region. For recAp the −32 position of the −35 signal corresponded to the −7 position is the TT/AA-unit as can be seen in Fig. 5(B). If one slides down from the −7 position, the AT-unit can be found first at the −22 position. In the case of recAp, it is more reasonable to consider TGTATG (as shown in Fig. 4B′) as the −35 signal of recAp than TTGATA, which has been determined experimentally in Ref. [6]. The $\tilde{S}_r$ value of AT-unit is smaller than that of the TA-unit. Therefore, the order of promoter strength is lacUV5 > recAp. The fact, reported in Ref. [5], that Parameter I of promoter strength is very large for recAp might be explained as follows. The number of TG/AC-units in the base sequences of a specific DNA region that is identified as the promoter, is five for recAp and only two for the other promoters (Fig. 5) [2]. As mentioned in Section 3, the TG/AC-unit is the best one for stable correction of RNA polymerases among the ten dimer-units. Parameter I might be associated with this feature. After sliding down from the −7 position, shown in Fig. 5(C), the CG- and GC-units are found at the −22 and −27 positions, respectively. In the case of rrnEp1, the RNA polymerase coordinated to the CG-unit at −22 position might carry out Process-3. As can be seen from Table 1, the $\tilde{S}_r$ value of the CG-unit is smaller than that of the AT-unit. As a result, the order of promoter strength is lacUV5 > recAp > rrnEp1. For rrnEp2, no DNA dimer-unit with $\tilde{S}_r(P_1) = \tilde{S}_r(P_2)$ can be found as slided down from the −9 position (AA/TT-unit of the −10 signal of rrnEp2) to the −32 position (the last position is identified as the promoter of rrnEp2) in Fig. 5(D). This suggests that rrnEp2 is the slowest in rate among the four promoters of Process-3. As seen from Fig. 4(D), the −35 signal (TTGACT) of rrnEp2 possesses a TG/AC-unit. Therefore, it is easy for the −35 signal to stably correct RNA polymerases (Process-1). However, it is difficult to carry out Process-3, since the −35 signal possesses no dimer-unit with $\tilde{S}_r(P_1) = \tilde{S}_r(P_2)$. In this case, it can be assumed that Processes-3 and 4 are carried out by compensating RNA polymerase coordinated to the other units (GA-unit (−24 position), GT-unit (−29 position) or AA/TT itself of the −10 signal for loss of the unit in additional processes. This feature is reflected on the experimental result [5] that as for the strength of rrnEp2, Parameter I is very small, whereas Parameter II is large.

5. Conclusions

On the basis of the interesting results obtained by previous calculations [1], i.e. the electronic states of the O2 and O3 in phosphates differ drastically from each other and might play a crucial role as recognition sites in various reaction processes concerning DNA, we propose a mechanism of how RNA polymerases recognize the transcription sites (phosphates). By employing the $\tilde{S}_r$ value of each phosphate for the ten DNA dimer-units (XX/Y′X′) and the six [XY/Y′X′ + H⁺]-complexes, a reasonable interpretation can be given why the base sequences of TATAAT and TTGACA have been believed experimentally to be the most ideal sequence for the −10 and −35 signals, respectively. From each character of phosphates for the ten dimer-units, the dimer-

[2] When the −35 signal of recAp is TGTATG (Fig. 4B′), the number of TG/AC-unit in the base sequence of −35 signal is two for recAp, whereas it is one for the other promoters.

units with $\tilde{S}_r(P_1) = \tilde{S}_r(P_2)$, TA/AT, AT/TA, CG/GC, GC/CG, might play the role of the −10 signal and the order of promoter strength is to be TA/AT > AT/TA > CG/GC > GC/CG. The dimer-units with $\tilde{S}_r(P_1) \geqq \tilde{S}_r(P_2)$, AA/TT, GA/CT, might make it possible to recognize the direction of transcription in the −10 sequence. The TG/AC which enables stable coordination, is an important unit for the −35 signal. On the other hand, for the AG/TC and GG/CC it is easy to collect RNA polymerases in the promoter region but difficult to result stable coordination.

As the result of an application to *E. coli* promoters, lacUV5, recAp, rrnEp1, rrnEp2, we can conclude that the order of promoter strength is lacUV5 > recAp > rrnEp1 > rrnEp2.

Our investigation of the base sequence of a specific DNA region identified as the promoter indicates that it is necessary for biochemical and biophysical workers to re-consider all of the base sequences in the promoter region without focusing their attention on only the −10 and −35 sequences. We wonder whether the region of the −10 and −35 signals can be determined perfectly by experimental methods such as that proposed for the −35 signal of recAp. We can say that it is now necessary to determine experimentally perfect base sequence of various promoters, because only a small change in base sequence of the promoter will result in a different character of the promoter in a similar fashion as with a mutation from a change in a base sequence on DNA. Our finding suggests that a modification of the promoter in genetic engineering by means of trial and error may prove fruitless and might be accompanied with dangers beyond one's supposition.

Acknowledgements

I am grateful to Dr. Akira Ishihama of the National Institute of Genetics and Dr. Hirotsugu Miyashiro of the Research Institute for Traditional Sino-Japanese Medicines, Toyama Medical and Pharmaceutical University for valuable comments.

References

1 T. Shinoda, N. Shima and M. Tsukuda, J. Theor. Biol. 151 (1991) 433.
2 K. Fukui, T. Yonezawa and C. Nagata, Bull. Chem. Soc. Japan 27 (1954) 423.
3 G.D. Fasman, in: Handbook of Biochemistry and Molecular Biology (CRC Press, Cleveland, OH, 1976) vol. 2, p. 411.
4 T. Shinoda, to be published.
5 M. Kajitani and A. Ishihama, Nucleic Acids Res. 11 (1983) 3873.
6 A. Ishihama, Chem. Today 11 (1983) 50.