STATISTICAL EXAMINATION OF DISSIMILARITY OF PROMOTER AND PROMOTER-LIKE SEQUENCES IN ESCHERICHIA COLI

SHIGEHICO KANAYA and YOSHIHIRO KUDO

Department of Electrical and Information Engineering Faculty of Yamagata University, Yonezawa, Yamagata, 992, Japan

Abstract -Specificity of Escherichia coli σ^{70} -promoter consensus-like sequences is examined by the χ^2 -statistics. The promoter consensus bases are serially designated as T_1 , T_2 , G_3 , A_4 , C_5 , A_6 , T_7 , A_8 , T_9 , A_{10} , A_{11} and T_{12} by neglecting a spacer bases between A_6 and T_7 . Occurrences of highly conserved sequences of the consensus are investigated with 75 transcription units in E.coli. About four-fifths of transcription units involving all of the polycistronic units in E.coli genome are repressed in both strands. A remarkable difference is detected on the frequencies of T_1 , T_7 and T_{12} between in promoters and coding regions. That is, whereas for promoter region they are higher than the critical frequency, for coding region they are lower, especially for 17-base spacers, than the critical one. The facts that occurrences of highly conserved sequences are repressed in transcription units and that χ -distribution of the three T's (T_1 , T_7 and T_{12}) are complementary between in promoters and coding regions suggest that transcription initiations within transcription units are repressed and the three T's of promoter sequences constitute a transcription trigger.

Key words Escherichia coli σ^{70} -promoter χ^2 -statistics

In prokaryotes, arrangements of base sequences in a DNA strand is one of factors controlling gene expression. In fact, upstream of transcription sites of Escherichia coli promoters a consensus sequence consisting of two 6-base sequences, 5'TTGACA 3' (the -35 region) and 5'TATAAT 3' (the -10 region) has been detected[1-3], and similarities of sequences corresponding to it (consensus-like sequences) have a marked influence on the transcription strength by σ^{70} RNA polymerase of E.coli[1-4]. In order to examine the specificity of the consensus sequence in the genome, we examine (1) repression of occurrences of the consensus-like sequences in transcription units and (2) difference of base substitution in the consensus sequence between in transcription units and in promoters.

METHOD

Analysis I: Repression of occurrences of the consensus-like sequences in transcription units

For TS and NTS which stand for transcribed and non-transcribed DNA strands, respectively, all pairs of two 6-base sequences, (b_{11} to b_{16}) and (b_{21} to b_{26}), separated by s bases as a spacer between $i-11$ -s and $t+11$ -s (i and t are positions of transcription initiation and termination bases, respectively), are selected and compared with the consensus sequence, base by base. The each frequency of sequences, $f_s(k)$, is counted, where k is the number of matched bases. The χ^2 -statistic, χ_s^2 is described as

$$\chi_s^2 = \sum_k \{ [f_s(k) - T_s(k)]^2 / T_s(k) \}$$

where $T_s(k)$ are the theoretical frequency of sequences on k -base matched to the consensus. Since the χ^2 -value also can be expressed as $\chi_s^2 = \sum_k [\chi_s(k)]^2$ χ -value is defined as

$$\chi_s(k) = [f_s(k) - T_s(k)] / \sqrt{T_s(k)}.$$

Each transcription unit may be assigned indices K_s 's defined by a set of $\chi_s(k)$ values:

$$K_s = \frac{\sum_{\chi_s(k) \leq 0} \{k \cdot \chi_s(k)\}}{\sum_{\chi_s(k) \leq 0} \chi_s(k)} - \frac{\sum_{\chi_s(k) > 0} \{k \cdot \chi_s(k)\}}{\sum_{\chi_s(k) > 0} \chi_s(k)}$$

A K_s indicates relationship between two gravities of sets of positive and negative $\chi_s(k)$'s, and can be an index representing the degree of repression of consensus-like sequences in transcription units: The more negative a K_s of a transcription unit is, the more strongly, consensus-like sequences with spacers of the s value are repressed in the transcription unit.

Analysis II: Difference of base substitution in the consensus between in transcription units and promoters

The promoter consensus sequence in *E. coli* are serially designated as $T_1, T_2, G_3, A_4, C_5, A_6, T_7, A_8, T_9, A_{10}, A_{11}, T_{12}$ by neglecting s -base spacer between A_6 and T_7 . As representatives of promoters and promoter-like non-promoters, PR (PRomoter) and PS (PSeudopromoter) are independently defined, respectively. First of all any 12-base sequence, on NTS, virtually made of two 6-base ones separated by 16- to 18-base spacer, is matched base by base with the consensus sequence, and ranked A to L according to the matched numbers, twelve (exact) to one, respectively. Rank A-D is synthesized rank of the highest down to D. PS is any of them in coding regions in genome. PR is defined on the basis of each transcription initiation site by the following procedure: All of the 12-base sequences of the highest rank separated to four to eight bases upstream from the transcription initiation site is searched for ones with 17-base spacers. If found, PR is anyone of only sequences with 17-base spacers, and the others are abandoned. If not, PR is anyone of them only with 16- or 18-base spacers. The conserved pattern for each of PR and PS of bases at twelve positions corresponding to the consensus sequences are examined by the χ^2 -statistic, χ^2_{bs} .

$$\chi^2_{bs} = \sum_k \left[\frac{\{f_i(x) - T_i(x)\}^2}{T_i(x)} \right]$$

where $f_i(x)$ and $T_i(x)$ are empirical and theoretical frequencies of base x at the position i , respectively. And $\chi^2_{bs}(x)$ are also defined as

$$\chi^2_{bs}(x) = \frac{\{f_i(x) - T_i(x)\}}{\sqrt{T_i(x)}}$$

RESULTS

From GenBank (Release 67.0) were extracted 75 transcription units, whose both (initiation and termination) sites are known, consisting of 6 regulatory (attenuators and leader peptides), 52 monocistronic and 17 polycistronic units. Table 1 classifies transcription units. About three-fourths of the monocistronic units and all of the polycistronic units, after all, four-fifths of the units have negative values of K_s 's.

The χ^2 -statistics is applied to the PRs and the PSs in rank A to D. The numbers of the PRs and the PSs of ranks A to D are 1, 6, 9 and 40, and 0, 0, 24 and 542, respectively. Table 2 summarizes of conserved features for PRs and PSs. The χ^2 -values at the positions of 1, 6, 7 and 12 for the PRs of Rank A-D, are larger than a critical value (7.81; level of significance=0.05). On the PRs, the frequencies are higher for T_1, T_7 and T_{12} , and lower for A_6 , than critical one (42.48). On the PSs, those are higher for T_2, A_6, T_9 and A_{11} and lower for $T_1, C_5, T_7, A_8, A_{10}$ and T_{12} than critical one (427.79).

TABLE 2. Summary of conserved features for PRs and PSs

Spacing	Rank	-35 Region						-10 Region					
		T1	T2	G3	A4	C5	A6	T7	A8	T9	A10	A11	T12
< Promoter >													
16-18	A-D	↑					↓ T G	↑					↑
16-18	D	↑					↓ T G	↑					
16	D					↓ A G							
17	D												
18	D			↓ T G						↓ G C			
<Pseudopromoter>													
16-18	A-D	↓ A C	↑			↓ A T	↑	↓ A G	↓ T	↑	↓ T G C	↑	↓ A C
16-18	D	↓ A C	↑			↓ A T	↑	↓ A G	↓ T	↑	↓ T G C	↑	↓ A C
16	D								↑	↓ A	↓ T G C	↑	↑
17	D	↓ A C	↑			↓ A T		↓ A G	↓ T G	↑		↑	↓ A C
18	D		↑			↓ A T		↓ A G					

'↑' and '↓' represent the frequencies of the consensus bases which are significantly larger and smaller than the theoretical ones (level of significance = 0.05). Letters (A, T, G, C) are bases with significantly high frequencies instead of the consensus bases.

DISCUSSION

The fact that Ks's of most of monocistronic and polycistronic units (three-fourths of the former units and all of the latter) are negative suggests that consensus-like sequences are suppressed strongly in the former units, and more strongly in the latter. Although the higher frequencies of T₁, T₇ and T₁₂ for PRs than the critical one (42.48) is essentially the same as reported [1], it is very interesting that for PSs, especially with 17-base spacers, corresponding ones are lower than the critical one (42.79).

These results shown in Tables 1 and 2 not only suggest that the three T's (T₁, T₇ and T₁₂) constitute a transcription trigger, but also support (1) that the χ^2 -statistics is an excellent means to specify the genetic signals in the vicinity of the transcription sites, and (2) that difference of conserved patterns of the 12-base sequences between PRs and PSs is a powerful key to determine the transcription initiation signals.

Since only about 30% of whole genome has been sequenced even in *E. coli* [5], this kind of investigation must use as much amount of information as possible. The present approach would be the more suitable for the complicated object because of its multifaceted viewpoint.

REFERENCES

1. U. Siebenlist, R. Simpson and W. Gibelt, *Cell*, 20, 269 (1980).
2. D. Howley and W. McClure, *Nucleic Acids Res.* 11, 2237 (1983).
3. W. Roznikoff, D. Sigel, D. Cowing and C. Gross *Ann. Rev. Genet.* 19, 355 (1985).
4. M. Kobayashi, K. Nagata and A. Ishihama, *Nucleic Acids Res.* 18, 7367 (1990).
5. D. Daniels and F. Blattner, *Nature*, 325, 831 (1987).