Mathematical Biology and Bioinformatics. 2011. V. 6. №1. P. t1-t13.
URL: http://www.matbio.org/2011/Ozoline2011(6_t1).pdf.
Original Russian text
Kiselev S.S., Ozoline O.N., 2011, published in Matematicheskaya biologiya i bioinformatika, 2011, V. 6. №1. P. 39-52.
URL: http://www.matbio.org/2011/Ozoline2011(6_39).pdf.

========== PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ==========
========== "MATHEMATICAL BIOLOGY AND BIOINFORMATICS"==========

UDC: 579.252

Structure-specific modules as indicators of promoter DNA in bacterial genomes

Kiselev S.S., Ozoline O.N.*

Institute of Cell Biophysics, Russian Academy of Sciences, Pushchino, Moscow region, 142290, Russia

Abstract. A unified version of promoter-search software, exploiting evolutionary conservation in the structural organization of bacterial transcription machinery is suggested (PlatPromU). In contrast with the initial algorithm PlatProm, optimized for recognition of σ^D-dependent promoters in the genome of *Escherichia coli* (*E.coli*), modified version does not use weight matrices, reflecting the occurrence frequency of consensus base pairs within −10 and −35 elements. Its predicting potentiality was assessed by the ability to recognize the known promoters of *Corynebacterium glutamicum* (*C.glutamicum*) — evolutionarily distant from *E.coli* microorganism. «*Sensitivity*» of PlatPromU appeared to be comparable with that of specialized program (PlatPromC), adapted for recognition of *C.glutamicum* regulatory sites, and higher than predicting potentiality of initial algorithm PlatProm. Unified program, modeling only structural and conformational features of promoter DNA, may, therefore, be recommended as a tool for preliminary mapping of regulatory sites in genomes with unknown context of specific elements.

Key words: *unified promoter search algorithm, the bacterial genome annotation.*

1. INTRODUCTION

A computer-based search for promoters has now becoming an important instrument for genome annotation. With equal efficiency it can reveal transcription initiation sites for genes encoding proteins, rRNA and tRNA, as well as promoters, controlling synthesis of untranslated, antisense and alternative RNA-products [1, 2]. However, most promoter finders explore sequence motifs specifically recognized by σ-subunits of RNA polymerase as the main indicators of promoter DNA. In spite of apparent evolution stability of the transcription machinery, the context of these motifs varies noticeably both for promoters of different bacteria and for promoters recognized by different types of σ-factors within one and the same microorganism [3]. That means that algorithms searching promoters of a particular type should be specifically adapted for the context of their conservative elements. This necessity in a specific adaptation essentially restrains the usage of computational approaches as a valid annotation tool.

In this study we suggest the approach opening an opportunity to overcome this limitation using algorithm PlatProm, initially adapted for σ^D-dependent promoters of *E.coli*. In addition to consensus hexanucleotides, forming specific contacts with σ-subunit of RNA polymerase, PlatProm takes into account conformational features of the promoter DNA. Due to evolution stability of RNA polymerases and universal structural organization of transcription machinery in different microorganisms, these features may be invariant or very similar in different

* ozoline@icb.psn.ru

promoter types. The contribution of structure-specific elements in the PlatProm scores constitutes approximately 50%. Full version of this software recognizes 85.5% of *E.coli* known promoters with p<0.0038 reliability. If only structure-specific modules in the promoter DNA are taken into account, "*sensitivity*" of the program decreases, but still remains rather high — 62.4%. Discriminative potentiality of structural modules provides thereby a chance to detect promoters without any contribution from the sequence-specific modules.

The genome of *Corynebacterium glutamicum* was used to test predictive potentiality of the unified program. In this genome 160 promoters of the main σ-factor – SigA (a counterpart of σ^D in *E.coli*) were mapped by experimental techniques. This is enough to develop new version PlatPromC, specifically adapted to corresponding promoters. The ability of unified version PlatPromU to find the regulatory sites of *C.glutamicum* was compared with that of specified version (PlatPromC), as well as with predictive potentiality of PlatProm algorithm. For this reason a new method of determining threshold levels, which ensure selection of transcriptional signals with equal statistical significance in the outputs of different search programs, was developed. The scores exceeding this level for 3, 4 and 5 standard deviations (StD) were considered as significant. Adapted program PlatPromC was most effective at the first level of reliability, but more stringent criteria made unified program more efficient, thus opening a way to suggest it for annotation procedures.

2. METHODS

2.1. Genome of *C.glutamicum* and promoters

The nucleotide sequence of the genomic DNA of *C.glutamicum* ATCC 13032 (NC_003450 in NCBI database [4]) was used for scanning and corresponding gene map for subsequent analysis. This genome is 3309401 base pairs (bp) long and contains 53% of G/C bp. Coordinates of the transcription start points for known promoters of *C.glutamicum* were taken from the original papers (table 1). Nucleotide sequences of promoters were obtained using auxiliary program DNA Tools (developed by A.A. Deev).

Table 1. Coordinates of the experimentally mapped transcription initiation points for *sigA*-dependent promoters of *Corynebacterium glutamicum*

Prom.	Start point and direction	R[a]	Prom.	Start point and direction	R[a]	Prom.	Start point and direction	R[a]
cg0042	29965 (–)	5	*narK*	1253952 (–)	19	P-45[b]	2346400 (–)	7
cg0043	29995 (+)	5	*atp1*	1271835 (+)	20	*glnA*	2348721 (+)	7
citH	70350/2/3[c] (–)	6	*atp2*	1272131 (+)	10	*thrC*	2355050 (–)	7
P-45 [b]	194354 (+)	7	*ssuD1*	1283324 (+)	18	*aceE*	2379862 (+)	34
gltB	195199 (+)	7	*pfkA*	1315055 (+)	16	*aecD*	2444605 (+)	14
dccT	239837 (+)	8	*rbsR*	1316264 (+)	19	*rbsK2*	2463200 (+)	25
leuA	268136 (–)	7	*lysE*	1328945 (–)	7	*aceB-P3*	2470325 (–)	11
orfMP	269124 (–)	7	*lysG*	1329000 (+)	7	*aceB-P2*	2470608/10 (11) (–)	11 (7)
askP1	269333 (–)	7	*ilvB*	1337840 (+)	7	*aceA*	2470630 (+)	7
askP2	270071 (+)	7	*ilvC*	1340628 (+)	7	*mdh*	2523282 (–)	35
lrp	276754 (–)	7	*leuB*	1353454 (+)	7	*pcaHG*	2541084 (–)	36
brnF	276829 (+)	7	*ltbR*	1380259 (–)	20	*clpP1*	2556624 (–)	32
brnE	277614 (+)	7	*leuC*	1380380 (+)	20	*metB*	2591526 (–)	14
glxR	307582 (–)	9	*ptsG*	1422959 (61,62) (+)	23, 24, 16	*malE1*	2608051 (–)	37
ushA	343576 (+)	10	*uriR*	1432678 (–)	25	*gntK-P2*	2630572 (+)	38
lpdA	387692 (+)	7	*ugpA*	1450890 (+)	10	*gntK-P1*	2630620 (+)	39

t2

ramB	392208 (–)	9	*metH*	1591237 (–)	14	*cg2782*	2674805 (+)	12
sdhC	392690 (+)	11	*acn*	1626169/72 (+)	26	*gpm*	2690077 (–)	16
cg0527	471013 (–)	12	*acnR*	1629247 (+)	11	*cg2810*	2699615 (–)	40
secE	496793 (+)	7	*sufR*1	1653617 (–)	27	*ramA*	2721299 (–)	41
P-13	597651 (+)	7	*amt*	1676679 (–)	7	*sucC*	2726673 (–)	11
groESL	610252 (+)	13	*pgk-P1*	1682462 (–)	7	*pstS*	2737620 (–)	10
P-2	632028 (–)	7	*pgk-P2*	1682499 (–)	16	*nucH*	2753958 (+)	10
metX	666353 (–)	14	*gapA*	1683809 (–)	7	*dctA*	2759320 (–)	42
metY	667809 (–)	14	*metK*	1700445 (–)	14	*phoR*	2774859 (–)	10
metY2	667832 (–)	14	*cg1935*	1813663 (+)	28	*pqo*	2778550 (–)	43
mdhB	676145 (–)	11	P-10	1868922 (–)	7	*cgl2611*	2778968 (+)	44
icd	680075 (–)	11	*sigA*	2011495 (+)	7	*thrE*	2790923 (+)	7
cg0771	684976 (–)	12	*divS-P1*	2036434 (–)	29	*cg2911*	2796866 (+)	5
pyc	705155 (+)	7	*divS-P2*	2036503 (–)	29	*ptsS*	2811869 (–)	24
cg0794	711644 (–)	5	*lexA*	2036607 (+)	29	*clpC*	2846977 (–)	32
cg0795	711669 (+)	5	*sugR*	2037767 (+)	30	*porH*	2888411 (–)	45
cg0922	850279 (–)	12	*ptsI-P2*	2041349 (–)	24	*groEL2*	2890687 (–)	13
gltA-P2	877479 (+)	15	*ptsI-P1*	2041415/7 (–)	24	*pta2*	2938094 (–)	46
gltA-P1	877715(7)[d] (+)	15, 7	*fruR-P1*	2041435/6 (8) (+)	24, 30	*pta1*	2937982 (–)	46
P-1A	939686 (+)	7	*fruR-P2*	2041602/5 (+)	24	P-22A	2944795 (–)	7
rpf2	963782 (+)	7	*cgl1934*	2041640 (+)	31	*fda*	2955421 (–)	7
gapB	993092 (+)	16	*ptsH*	2045635 (+)	30	*ald*	2981791 (–)	47
P-34	1034563 (+)	7	*ptsH-P1*	2045660 (+)	24	*dnaK*	2986507 (–)	13
eno	1034879 (+)	16	*ptsH-P2*	2045680 (+)	24	*adhA*	2996912 (–)	48
P-64	1045560 (+)	7	*clgR*	2069968 (–)	32	*cysI*	3005214 (–)	49
glyA	1050560 (+)	17	*dapA*	2080183 (–)	7	*fpr2*	3005440 (+)	49
fum	1063654 (–)	11	*dapB2*	2081925 (–)	7	*tctC*	3012908 (–)	6
ssuI	1063936 (+)	18	*dapB1*	2081974 (–)	7	P-45[b]	3033754 (+)	7
seuA	1066071 (+)	18	*mqo*	2115532 (–)	11	*pckA*	3053929 (–)	16
ssuD2	1069959 (+)	18	*gdh*	2196368 (–)	7	*gntP*	3108088 (+)	39
P-75	1102054 (+)	7	*ilvA*	2246172 (–)	7	*ldhA*	3113479 (83) (–)	30, 35
pgm	1107515 (+)	16	*ftsZ1*	2280258 (–)	33	*cgl2816*	3118211 (+)	50
orf3-aroP	1155750 (–)	7	*ftsZ2*	2280457 (–)	33	*cg3327*	3201755 (–)	12
odhA	1176370 (–)	11	*ftsZ3*	2280503 (–)	33	*malE*	3208210 (+)	16
metE	1190662 (–)	14	*ftsZ4*	2280648 (–)	33	*trp*	3233129 (+)	7
argS	1238270 (+)	7	*ftsZ5*	2280729 (–)	33	*cg3372*	3248349 (+)	40
hom	1242420 (+)	7	*metF*	2299526 (–)	14			
thrB	1243843 (+)	7	*sucB*	2339224 (+)	11			

«a» – reference to literary source, «b» – promoter P-45 present in a three copies, «c» – multiple start points, «d» – start points given in different sources.

2.2. Design of PlatProm weight matrices

Correspondence of genomic sequences to conservative hexanucleotides –35 and –10, forming the specific contacts with the σ-subunit of RNA polymerase, was estimated using position weight matrices (PWM). Their occurrence frequencies (weights) were calculated the same way as proposed by Hertz and Stormo [51]. Each of these matrices possesses 24 parameters k_{ij}, determined as:

$$k_{ij} = \ln(f_{ij} / n_j), \quad (1)$$

t3

where i – nucleotide position in the element, j – particular nucleotide (A, C, G или T), f_{ij} – occurrence frequency of nucleotide j in position i, n_j – normalization coefficient, reflecting the occurrence frequency of j in analyzed genome.

The correspondence of analyzing nucleotide sequences to the consensus elements is estimated as a sum of contributions given by all pairs located in regions of expected disposition of conserved modules:

$$K_c = \sum_{i}^{12} \sum_{j}^{4} k_{ij} ,$$

where k_{ij} – weight of the nucleotide, located in the analyzed position and calculated by the formula (1), or 0 (the possibility of summation is provided for degenerative alphabets).

The variations in spacer (S) length between –35 and –10 elements (allowed range $14 \le S \le 21$) and distance (D) between –10 element and the transcription start point (allowed range $2 \le D \le 9$) were taken into account using weight matrices, reflecting occurrence frequencies of different distances in the training set of promoters:

$$K_{S(D)} = \ln (N_{S(D)} / N_{17(6)}),$$

where $N_{S(D)}$ – the number of promoters with corresponding S and D, $N_{17(6)}$ – the number of promoters with optimal S (17 bp) and D (6 bp).

Any deviation from the optimal values of S or D decreased the total score on the value of $K_{S(D)}$. Since the number of promoters with very long or very short positional distances is rather small, their real $K_{S(D)}$ gave very high negative contributions, which impaired alignment in respect to conservative hexanucleotides. That is why, the value of K_S for all promoters with spacer length > 18 bp was set equal to K_S, calculated for promoters with S = 18 bp, while for promoters with S < 16 bp K_S was equated to promoters with S = 16. Dependence on D was reduced by the same manner. For promoters with D = 4, 5, 7 or 8, values of K_D were calculated according to occurrence frequency of corresponding promoters in the compilation. K_D for promoters with $D \le 3$ was set equal to K_D of promoters with D = 4, while K_D of promoters with c D = 9 was equated to promoters with D = 8.

The computation of the optimal PWM was performed by the method of successive iterations. PWMs of the first step were calculated manually on the basis of 30 promoters, which conservative hexanucleotides were found experimentally by genetic techniques. Modules identified by PlatProm as a consensus hexanucleotides in the promoters of learning compilation (308 non-homologous and non-overlapping sequences), were used by the program to generate refined PWMs (first iteration). These PWMs were used in the next step and so on up to the full stabilization of occurrence frequencies in subsequent steps.

Peculiarities in the nucleotide sequence nearby transcription start points were accounted by one-dimensional weight matrix, which parameters (k_{di}) reflect the occurrence frequency of 16 dinucleotides in the position –1:

$$k_{di} = \ln (f_{di} / n_{di}),$$

Where di – particular dinucleotide, f_{di} – occurrence frequency of di in position –1 in the promoter compilation, n_{di} – occurrence frequency of di in genome.

The presence of functionally important dinucleotide TG in the 5'-flanking region of –10 element (k_{TG}) was accounted by the same way.

Along with elements listed above, PlatProm takes into account specific conformational features of promoter DNA, as well as the modules, favoring transcription complex formation and its transition to the productive initiation [1–3, 52–56]. These elements include:

- regular distribution of polyA(T)-tracts, which interact with RNA polymerase α-subunits or stabilize the transcription complex by a properly induced bend;

- flexible YR-dinucleotides (Y=C=T, R=A=G), which support adaptive isomerization of the DNA helix upon interaction with RNA polymerase or regulatory proteins;
- periodic distribution of mixed A/T-tracts, hypothetically participating in RNA polymerase sliding along the DNA;
- direct and inverted repeats as a putative targets for interaction with transcription factors;
- other dominant motives previously revealed for *E.coli* promoters by cluster analysis [56].

Table 2. Cascade weight matrix, reflecting the heightened occurrence frequency in the presence of flexible steps within –55/–52 promoter region

Position	Sequence module	Normalized logarithm of the occurrence frequency in promoters
–53	ACAT	1,56
–56	CACA	0,90
–52	CAT	0,89
–56	TCAT	0,70
–55	CAT	0,61
–57	ACAC	0,44
–57	ACA	0,18
All elements absent		–0,036

Sequence elements were considered as a promoter-specific, if their occurrence frequency in the particular position of the promoter DNA (the range of analyzed area: –250/+150 according to the start point of transcription) was at least 5 standard deviations (StD) higher than background level. All of them are taken into account by 60 *cascade* matrices (exemplified in table 2), which differ from usual PWMs by containing frequency weights for only dominating motifs (calculated as normalized natural logarithm of occurrence frequency for a particular sequence element in the fixed promoter position). If scrutinized sequence possesses several overlapping motifs (for instance, $ACAT_{-53}$ and CAT_{-52}, table 2), the contribution to the total score gives only that one, which weight in the promoter compilation is higher (ACAT). An absence of all promoter-specific motifs in a particular sequence is penalized by the negative contribution quantified as a logarithm of the portion of such promoters in the training set.

The total score was calculated as a sum of contributions given by all weight matrices, balanced by such a way, so as an overall contribution of the cascade matrices appeared to be ~50%. For this reason, the contribution of the every cascade matrix was normalized per the relative information content of corresponding promoter region and that of the last base pair in the −35 element (the least conservative base pair). Information content was calculated by the algorithm, suggested in [57].

2.3. Estimation of the statistically significant threshold level

Two sets of sequences were previously used to estimate the background level and StD, typical for non-promoter DNAs [1]. The first of them (CS1) was composed of 273 fragments of coding sequences taken from convergent *E.coli* genes longer than 700 bp, which are separated by at least 50 bp intergenic space. The probability of coming across a functional promoters within such genes is minimal. The second set contained 400 random sequences with the same AT/GC-content as in the studied genome. Each of these two sets has certain advantages and limitations. An advantage of the first compilation is its biological authenticity but already annotated genome is required to collect natural sequences and available number may be less than required for statistical analysis. Any size set of random sequences can be

easily obtained computationally but generated fragments with certain probability will contain promoter-like sequences, which non-random distribution in genomes is controlled by evolution. In this work we suggest the novel method of threshold levels determination. Its essence is to select natural sequences, which have minimal probability to function as promoter DNA. For this purpose the mean value of score (F) and StD was estimated for 1000 bp long fragments (average length of a gene) using the sliding window mode of calculations. Then the genome was partitioned per the segments of equal length, and position with minimal F ($F_{\min}$) was found within each of them (exemplified in fig. 1). The mean value of $F_{\min}$ across the whole genome ($\overline{F}$) was considered the background level, while the mean value of corresponding StDs characterized variability of the PlatProm scores in non-promoter regions.

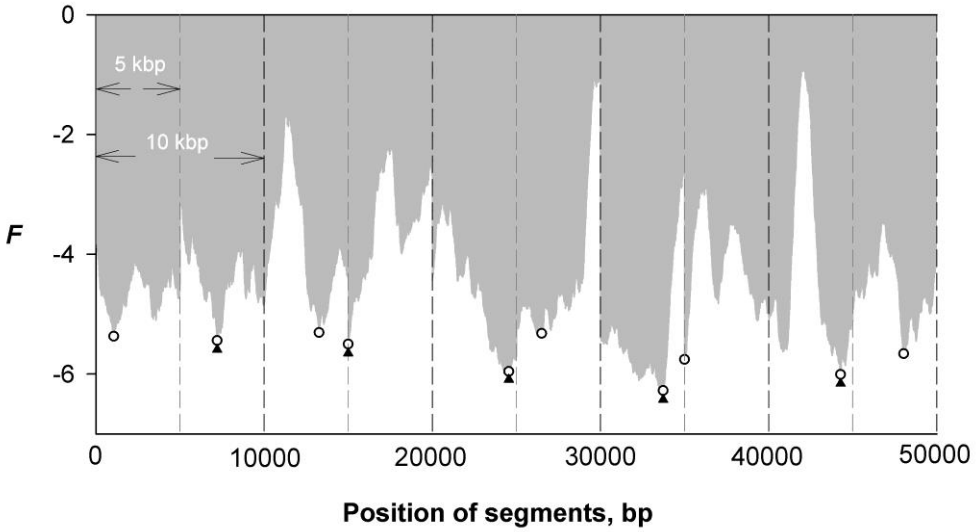

Figure 1. Searching for non-promoter regions in the first 50000 bp of the *E.coli* MG1655 genome (NC_000913 in [4]). Grey bars represent an average score within 1000 bp fragments. Local minima ($F_{\min}$), revealed within 5000 and 10000 bp segments are marked by circles and triangles, respectively.

The values of StD did not show essential dependence on the partitioning density, while the value of $F_{\min}$ tend to decrease with increased length of the segments (fig. 1, table 3). To select an optimal density, the values of $\overline{F}$ given by different segmentations were compared with the background level quantified on the basis of CS1. The closest values were found in the case, when genome of *E.coli* was partitioned by 5000 bp segments, so such values of $\overline{F}$ were used to calculate the three thresholds (L) after scanning of the *C.glutamicum* genome by different PlatProm versions:

$$L_n = \overline{F} + n\text{StD},\ n = 3,\ 4 \text{ and } 5.$$

Table 3. Dependence of $\overline{F}$ and StD on the segmentation density

Segment length (bp)	$\overline{F}$	StD	Second threshold level (L_2)
Escherichia coli K12 MG1655 (PlatProm)			
5000	–5,41	3,27	7,67
10000	–5,76	3,24	7,2
20000	–6,04	3,21	6,8
50000	–6,35	3,20	6,45
Corynebacterium glutamicum ATCC 13032 (PlatPromC)			
5000	–4,14	2,63	6,37
10000	–4,40	2,61	6,04
20000	–4,62	2,60	5,78
50000	–4,91	2,59	5,44

2.4. Adaptation of the PlatProm PWMs to the context of *C.glutamicum* −35 and −10 modules

Assuming invariant structural organization in the bacterial transcription machinery, only PWMs of the PlatProm were adapted to the context of *C.glutamicum* promoters. Initially these matrices were designed using training set composed of 308 known *E.coli* promoters (see above), while "*sensitivity*" of the program was estimated using test compilation, composed of 290 non-overlapping and non-homologous *E.coli* promoters, absent in the training set [1, 2, 52]. But restricted amount of experimentally mapped promoters in *C.glutamicum* genome (a total of 160 sequences of which 3 identical) not allowed to compose two independent compilations. Thus the strategy of alternate targets was used to test the quality of specialized program. In this case each known promoter in turn was considered as a testing sample, while remaining 157 sequences were used to quantify PWMs. But through-genome scan with a purpose to determine the background level was made by a specialized version of the program (PlatPromC), which was designed based on all 158 promoters.

2.5. Unification of PlatProm

The design of the unified algorithm is a complex multistep program, which implementation may require accounting of additional factors or, perhaps, further simplification of the PlatProm scoring system. As a first step in this study we validated predictive capacity of the program PlatPromU using only cascade matrices of PlatProm.

2.6. Criteria used to estimate predictive capacity of computer algorithms

"*Sensitivity*" of the programs used was estimated as a percentage of promoters identified at different levels of reliability. Scores, exceeding the background level by 3, 4 or 5 StDs ($p < 0.0014$, $p < 0.00004$ and 0.000001, respectively) were considered as significant. Promoters was considered as recognized, if predicted point of the transcription initiation laid in the range ±5 bp nearby the experimentally mapped start. Since experimental identification of the RNAs 5'-ends is associated with some inaccuracy, the promoter was considered as accurately recognized by computer program, if the position of predicted start coincided or was located within 2 bp region nearby experimental point.

3. RESULTS

Fig. 2 exemplifies the distribution of significant scores in front of *C.glutamicum* gene *phoR*, encoding phosphate regulon sensor kinase-phosphotransferase. The transcription start point of this gene is located 44 bp upstream of the initiating codon ATG [10].

All three algorithms (PlatPromC, PlatProm and PlatPromU) revealed promoter-like site alongside of *phoR*. But *E.coli*-specific algorithm (PlatProm) overlooking the real transcription initiation point (position −44) offers the position −75 as the most probable start (middle plot in fig. 2). Specialized program PlatPromC accurately identifies the real start (red bar in the top plot of fig. 2), while also predicts transcription initiation at position −74. The score in this position exceeds the background level by 4.99 StD, which corresponds to $p < 0.000001$. That means that only 7 promoter-like signals with the same amplitude may be found in the genome of *C.glutamicum* by chance. The probability for this signal to be a false positive is, therefore, very low, especially as unified program (bottom plot in fig. 2) predicted this additional start (and the real point of transcription initiation) with very high reliability. Most probably that means that *phoR* expression may be controlled by two tandem promoters.

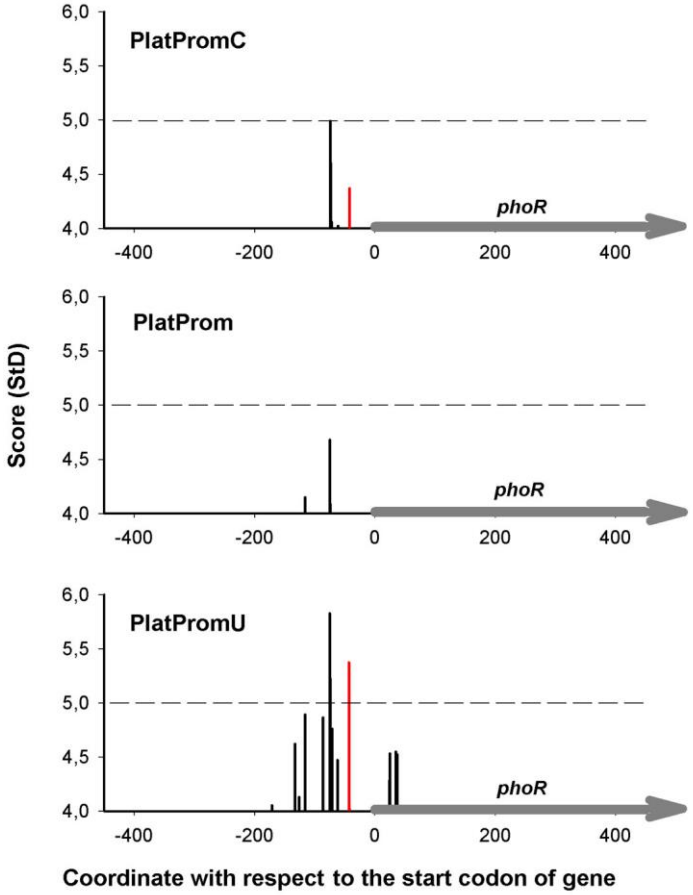

Figure 2. Transcription start points predicted by three algorithms for the *C.glutamicum* gene *phoR* (gray arrow). Red colour marks the real transcription start point. The X-axis corresponds to the second level of reliability ($\overline{F}$ + 4 StD), dashed line delineates the third level ($\overline{F}$ + 5 StD).

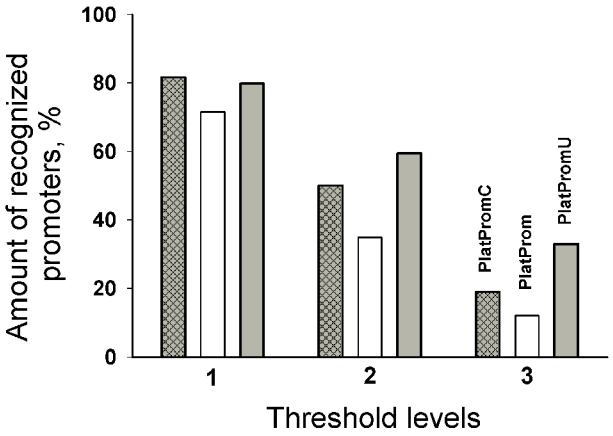

Figure 3. An ability of PlatPromC, PlatProm and PlatPromU (indicated on the plot) to recognize *C.glutamicum* promoters at different levels of reliability. Signals, exceeding the background level ($\overline{F}$) by 3, 4 and 5 StDs were considered significant at threshold levels 1, 2 and 3, respectively.

Fig. 3 demonstrates the summary of comparative analysis. At the first level ($p < 0.0014$) adapted program PlatPromC (hatched bars) was most effective. It identifies 81.6% promoters, i.e. as much as PlatProm at the same cut off level (81%, data not shown) recognizes in the test compilation, composed of *E.coli* promoters. That means that cascade matrices of PlatPromC, which remain «tuned» to the conformational features of *E.coli* promoters, are equally discriminative for promoters of *C.glutamicum*. Their use in combination with *E.coli*-specific PWMs significantly decreased predictive capacity of the program (white bars). This is

consistent with the generally accepted view, which assumes relevance of the specific adaptation for the promoter search algorithms. «*Sensitivity*» of the unified program, operating with only cascade matrices (grey bars), at the first cut off level was almost the same as «*sensitivity*» of adapted algorithm (79.7%), while at higher levels even exceeded it. That means that the structure-specific modules in the promoter DNA, scored by the cascade matrices, indeed, may be used as efficient indicators.

DISCUSSION

The study presents the first version of the unified software with ability to point out transcription initiation sites within the genomes with poorly characterized or completely unknown regulatory elements. For this purpose a simplified version of the PlatProm was used. PWMs, estimating a correspondence of nucleotide sequences to the consensus −10 and −35 elements of *E.coli* σ^D-dependent promoters, in the simplified version were switched off. Performance of the program was tested using promoters of *C.glutamicum*. This Gram-pozitive microorganism belongs to the phylum *Actinobacteria*, while Gram-negative *E.coli* is a Gammaproteobacteria. A predictive capacity of unified version was, therefore, evaluated in the strict conditions of heterologous genetic system. The data obtained, with no doubt, indicate that the structural features of promoter DNA may be used to point out transcription initiation sites. However, a majority of currently used structure-specific modules are enriched by A/T-pairs. Thus, it is not yet clear, how efficiently the unified program will be able to find promoters in the genomes with heightened and, conversely, lowered GC-content.

The analysis of the data obtained revealed higher tendency of transcription signals, found by the unified program, to form extensive clusters, than that of specialized algorithms (fig. 2). The presence of clustered transcription signals nearby the real promoters has long been known and discussed [1, 58, 59]. It has been speculated that overlapping promoter-like sites are used by cell transcription machinery to increase local concentration of RNA polymerase hereabout transcribed genomic regions [58, 59]. Along with this phenomenon, in the genome of *E.coli* genome we discovered abnormally long ($\geq$ 300 bp) «promoter islands» [1]. Efficiently interacting with RNA polymerase they exhibited paradoxically low transcription activity [1]. It is not excluded that these novel structural elements perform some kind of specific biological role not necessarily associated with RNA synthesis. Comprehensive comparative screening of the «promoter islands» by unified and specific algorithms in different genomes may be useful in order to accept or reject an assumption on the involvement of these new genomic elements in the structural remodeling of chromosomal DNA.

The work was supported by Russian Foundation for Basic Research (grant 10-04-01218).

REFERENCES

1. Shavkunov K.S., Masulis I.S., Tutukina M.N., Deev A.A., Ozoline O.N. Gains and unexpected lessons in genome-scale promoter mapping. *Nucleic Acids Res.* 2009. V. 37. P. 4419–4431.
2. Ozoline O.N., Deev A.A. Predicting antisense RNAs in the genomes of *Escherichia coli* and *Salmonella typhimurium* using promoter-search algorithm PlatProm. *J. Bioinf. Comput. Biol.* 2006. V. 4. P. 443–454.
3. Ozoline O.N., Purtov Yu.A., Brok-Volchanski A.S., Deev A.A., Lukyanov V.I. Specifity of DNA-protein interactions within transcription complexes of *Escherichia coli*. *Molecular biology.* 2004. V. 38. P. 663–673.
4. URL: ftp://ftp.ncbi.nih.gov/genomes/Bacteria/.
5. Schroder J., Jochmann N., Rodionov D.A., Tauch A. The Zur regulon of *Corynebacterium glutamicum* ATCC 13032. *BMC Genomics.* 2010. V. 11. Article No. 12.

6. Brocker M., Schaffer S., Mack C., Bott M. Citrate utilization by *Corynebacterium glutamicum* is controlled by the CitAB two-component system through positive regulation of the citrate transport genes *citH* and *tctCBA*. *J. Bacteriol.* 2009. V. 191. P. 3869–3880.
7. Patek M., Nesvera J., Guyonvarch A., Reyes O., Leblon G. Promoters of *Corynebacterium glutamicum*. *J. Biotechnol.* 2003. V. 104. P. 311–323.
8. Youn J.-W., Jolkver E., Kramer R., Marin K., Wendisch V.F. Identification and characterization of the dicarboxylate uptake system DccT in *Corynebacterium glutamicum*. *J. Bacteriol.* 2008. V. 190. P. 6458–6466.
9. Jungwirth B., Emer D., Brune I., Hansmeier N., Puhler A., Eikmanns B.J., Tauch A. Triple transcriptional control of the resuscitation promoting factor 2 (*rpf2*) gene of *Corynebacterium glutamicum* by the regulators of acetate metabolism RamA and RamB and the cAMP-dependent regulator GlxR. *FEMS Microbiol. Lett.* 2008. V. 281. P. 190–197.
10. Kocan M., Schaffer S., Ishige T., Sorger-Herrmann U., Wendisch V.F., Bott M. Two-component systems of *Corynebacterium glutamicum*: deletion analysis and involvement of the PhoS-PhoR system in the phosphate starvation response. *J. Bacteriol.* 2006. V. 188. P. 724–732.
11. Han S.O., Inui M., Yukawa H. Transcription of *Corynebacterium glutamicum* genes involved in tricarboxylic acid cycle and glyoxylate cycle. *J. Mol. Microbiol. Biotechnol.* 2008. V. 15. P. 264–276.
12. Brune I., Werner H., Huser A.T., Kalinowski J., Puhler A., Tauch A. The DtxR protein acting as dual transcriptional regulator directs a global regulatory network involved in iron metabolism of *Corynebacterium glutamicum*. *BMC Genomics.* 2006. V. 7. Article No. 21.
13. Barreiro C., Gonzalez-Lavado E., Patek M., Martin J.-F. Transcriptional analysis of the *groES-groEL1*, *groEL2*, and *dnaK* genes in *Corynebacterium glutamicum*: characterization of heat shock-induced promoters. *J. Bacteriol.* 2004. V. 186. P. 413–417.
14. Suda M., Teramoto H., Imamiya T., Inui M., Yukawa H. Transcriptional regulation of *Corynebacterium glutamicum* methionine biosynthesis genes in response to methionine supplementation under oxygen deprivation. *Appl. Microbiol. Biotechnol.* 2008. V. 81. P. 505–513.
15. Van Ooyen J., Emer D., Bussmann M., Botta M., Eikmanns B.J., Eggeling L. Citrate synthase in *Corynebacterium glutamicum* is encoded by two *gltA* transcripts which are controlled by RamA, RamB, and GlxR. *J. Biotechnol.* 2011. (in press).
16. Han S.O., Inui M., Yukawa H. Expression of *Corynebacterium glutamicum* glycolytic genes varies with carbon source and growth phase. *Microbiology.* 2007. V. 153. P. 2190–2202.
17. Schweitzer J.-E., Stolz M., Diesveld R., Etterich H., Eggeling L. The serine hydroxymethyltransferase gene *glyA* in *Corynebacterium glutamicum* is controlled by GlyR. *J. Biotechnol.* 2009. V. 139. P. 214–221.
18. Koch D.J., Ruckert C., Albersmeier A., Huser A.T., Tauch A., Puhler A., Kalinowski J. The transcriptional regulator SsuR activates expression of the *Corynebacterium glutamicum* sulphonate utilization genes in the absence of sulphate. *Mol. Microbiol.* 2005. V. 58. P. 480–494.
19. Nishimura T., Vertes A.A., Shinoda Y., Inui M., Yukawa H. Anaerobic growth of *Corynebacterium glutamicum* using nitrate as a terminal electron acceptor. *Appl. Microbiol. Biotechnol.* 2007. V. 75. P. 889–897.
20. Barruiso-Iglesias M., Barreiro C., Flechoso F., Martin J.F. Transcriptional analysis of the F_0F_1 ATPase operon of *Corynebacterium glutamicum* ATCC 13032 reveals strong induction by alkaline pH. *Microbiology.* 2006. V. 152. P. 11–21.

21. Nentwich S.S., Brinkrolf K., Gaigalat L., Huser A.T., Rey D.A., Mohrbach T., Marin K., Puhler A., Tauch A., Kalinowski J. Characterization of the LacI-type transcriptional repressor RbsR controlling ribose transport in *Corynebacterium glutamicum* ATCC 13032. *Microbiology.* 2009. V. 155. P. 150–164.
22. Brune I., Jochmann N., Brinkrolf K., Huser A.T., Gerstmeir R., Eikmanns B.J., Kalinowski J., Puhler A., Tauch A. The IclR-type transcriptional repressor LtbR regulates the expression of leucine and tryptophan biosynthesis genes in the amino acid producer *Corynebacterium glutamicum*. *J. Bacteriol.* 2007. V. 189. P. 2720–2733.
23. Engels V., Wendisch V.F. The DeoR-type regulator SugR represses expression of *ptsG* in *Corynebacterium glutamicum*. *J. Bacteriol.* 2007. V. 189. P. 2955–2966.
24. Tanaka Y., Okai N., Teramoto H., Inui M., Yukawa H. Regulation of expression of phosphoenolpyruvate:carbohydrate phosphotransferase system (PTS) genes in *Corynebacterium glutamicum* R. *Microbiology.* 2008. V. 154. P. 264–274.
25. Brinkrolf K., Ploger S., Solle S., Brune I., Nentwich S.S., Huser A.T., Kalinowski J., Puhler A., Tauch A. The LacI/GalR family transcriptional regulator UriR negatively controls uridine utilization of *Corynebacterium glutamicum* by binding to catabolite-responsive element (*cre*)-like sequences. *Microbiology.* 2008. V. 154. P. 1068–1081.
26. Krug A., Wendisch V.F., Bott M. Identification of AcnR, a TetR-type repressor of the aconitase gene *acn* in *Corynebacterium glutamicum*. *J. Biol. Chem.* 2005. V. 280. P. 585–595.
27. Nakunst D., Larisch C., Huser A.T., Tauch A., Puhler A., Kalinowski J. The extracytoplasmic function-type sigma factor SigM of *Corynebacterium glutamicum* ATCC 13032 is involved in transcription of disulfide stress-related genes. *J. Bacteriol.* 2007. V. 189. P. 4696–4707.
28. Zemanova M., Kaderabkova P., Patek M., Knoppova M., Silar R., Nesvera J. Chromosomally encoded small antisense RNA in *Corynebacterium glutamicum*. *FEMS Microbiol. Lett.* 2008. V. 279. P. 195–201.
29. Jochmann N., Kurze A.-K., Czaja L.F., Brinkrolf K., Brune I., Huser A.T., Hansmeier N., Puhler A., Borovok I., Tauch A. Genetic makeup of the *Corynebacterium glutamicum* LexA regulon deduced from comparative transcriptomics and in vitro DNA band shift assays. *Microbiology.* 2009. V. 155. P. 1459–1477.
30. Dietrich C., Nato A., Bost B., Le Marechal P., Guyonvarch A. Regulation of *ldh* expression during biotin-limited growth of *Corynebacterium glutamicum*. *Microbiology.* 2009. V. 155. P. 1360–1375.
31. Gao Y.-G., Suzuki H., Itou H., Zhou Y., Tanaka Y., Wachi M., Watanabe N., Tanaka I., Yao M. Structural and functional characterization of the LldR from *Corynebacterium glutamicum*: a transcriptional repressor involved in L-lactate and sugar utilization. *Nucl. Acids Res.* 2008. V. 36. P. 7110–7123.
32. Engels S., Schweitzer J.-E., Ludwig C., Bott M., Schaffe S. *clpC* and *clpP1P2* gene expression in *Corynebacterium glutamicum* is controlled by a regulatory network involving the transcriptional regulators ClgR and HspR as well as the ECF sigma factor σ^H. *Mol. Microbiol.* 2004. V. 52. P. 285–302.
33. Letek M., Ordonez E., Fiuza M., Honrubia-Marcos P., Vaquera J., Gil J.A., Mateos L.M. Characterization of the promoter region of *ftsZ* from *Corynebacterium glutamicum* and controlled overexpression of FtsZ. *Int. Microbiol.* 2007. V. 10. P. 271–282.
34. Schreiner M.E., Fiur D., Holatko J., Patek M., Eikmanns B.J. E1 enzyme of the pyruvate dehydrogenase complex in *Corynebacterium glutamicum*: molecular analysis of the gene and phylogenetic aspects. *J. Bacteriol.* 2005. V. 187. P. 6005–6018.
35. Inui M., Suda M., Okino S., Nonaka H., Puskas L.G., Vertes A.A., Yukawa H. Transcriptional profiling of *Corynebacterium glutamicum* metabolism during organic

acid production under oxygen deprivation conditions. *Microbiology.* 2007. V. 153. P. 2491–2504.

36. Zhao K.X., Huang Y., Chen X., Wang N.X., Liu S.J. PcaO positively regulates *pcaHG* of the beta-ketoadipate pathway in *Corynebacterium glutamicum*. *J. Bacteriol.* 2010. V. 192. P. 1565–1572.
37. Okibe N., Suzuki N., Inui M., Yukawa H. Isolation, evaluation and use of two strong, carbon source-inducible promoters from *Corynebacterium glutamicum*. *Lett. Appl. Microbiol.* 2010. V. 50. P. 173–178.
38. Frunzke J., Engels V., Hasenbein S., Gatgens C., Bott M. Co-ordinated regulation of gluconate catabolism and glucose uptake in *Corynebacterium glutamicum* by two functionally equivalent transcriptional regulators, GntR1 and GntR2. *Mol. Microbiol.* 2008. V. 67. P. 305–322.
39. Letek M., Valbuena N., Ramos A., Ordonez E., Gil J.A., Mateos L.M. Characterization and use of catabolite-repressed promoters from gluconate genes in *Corynebacterium glutamicum*. *J. Bacteriol.* 2006. V. 188. P. 409–423.
40. Ruckert C., Milse J., Albersmeier A., Koch D.J., Puhler A., Kalinowski J. The dual transcriptional regulator CysR in *Corynebacterium glutamicum* ATCC 13032 controls a subset of genes of the McbR regulon in response to the availability of sulphide acceptor molecules. *BMC Genomics.* 2008. V. 9. Article No. 483.
41. Cramer A., Eikmanns B.J. RamA, the transcriptional regulator of acetate metabolism in *Corynebacterium glutamicum*, is subject to negative autoregulation. *J. Mol. Microbiol. Biotechnol.* 2007. V. 12. P. 51–59.
42. Youn J.-W., Jolkver E., Kramer R., Marin K., Wendisch V.F. Characterization of the dicarboxylate transporter DctA in *Corynebacterium glutamicum*. *J. Bacteriol.* 2009. V. 191. P. 5480–5488.
43. Schreiner M.E., Riedel C., Holatko J., Patek M., Eikmanns B.J. Pyruvate:quinone oxidoreductase in *Corynebacterium glutamicum*: molecular analysis of the *pqo* gene, significance of the enzyme, and phylogenetic aspects. *J. Bacteriol.* 2006. V. 188. P. 1341–1350.
44. Itou H., Okada U., Suzuki H., Yao M., Wachi M., Watanabe N., Tanaka I. The CGL2612 protein from *Corynebacterium glutamicum* is a drug resistance-related transcriptional repressor: structural and functional analysis of a newly identified transcription factor from genomic DNA analysis. *J. Biol. Chem.* 2005. V. 280. P. 38711–38719.
45. Barth E., Barcelo M.A, Klackta C., Benz R. Reconstitution experiments and gene deletions reveal the existence of two-component major cell wall channels in the genus *Corynebacterium*. *J. Bacteriol.* 2010. V. 192. P. 786–800.
46. Gerstmeir R., Wendisch V.F., Schnicke S., Ruan H., Farwick M., Reinscheid D., Eikmanns B.J. Acetate metabolism and its regulation in *Corynebacterium glutamicum*. *J. Biotechnol.* 2003. V. 104. P. 99–122.
47. Auchter M., Arndt A., Eikmanns B.J. Dual transcriptional control of the acetaldehyde dehydrogenase gene *ald* of *Corynebacterium glutamicum* by RamA and RamB. *J. Biotechnol.* 2009. V. 140. P. 84–91.
48. Arndt A., Eikmanns B.J. The alcohol dehydrogenase gene *adhA* in *Corynebacterium glutamicum* is subject to carbon catabolite repression. *J. Bacteriol.* 2007. V. 189. P. 7408–7416.
49. Ruckert C., Koch D.J., Rey D.A., Albersmeier A., Mormann S., Puhler A., Kalinowski J. Functional genomics and expression analysis of the *Corynebacterium glutamicum fpr2-cysIXHDNYZ* gene cluster involved in assimilatory sulphate reduction. *BMC Genomics.* 2005. V. 6. Article No. 121.

50. Georgi T., Engels V., Wendisch V.F. Regulation of L-lactate utilization by the FadR-type regulator LldR of *Corynebacterium glutamicum*. *J. Bacteriol.* 2008. V. 190. P. 963–971.
51. Hertz G.Z., Stormo G.D. *Escherichia coli* promoter sequences: analysis and prediction. *Methods in Enzymology*. 1996. V. 273. P. 30–42.
52. Brok-Volchanski A.S., Masulis I.S., Shavkunov K.S., Lukyanov V.I., Purtov Yu.A., Kostyanicina E.G., Deev A.A., Ozoline O.N. Predicting sRNA genes in the genome of *E.coli* by the promoter-search algorithm PlatProm. In: *Bioinformatics of Genome Regulation and Structure II.* Eds. Kolchanov N., Hofestaedt R., Milanesi L. New York: Springer, 2006. P. 11–20.
53. Ozoline O.N., Deev A.A., Arkhipova M.V., Chasov V.V., Travers A. Proximal transcribed regions of bacterial promoters have non-random distribution of A/T-tracts. *Nucl. Acids Res.* 1999. V. 27. P. 4768–4774.
54. Ozoline O.N., Deev A.A., Trifonov E.N. DNA bendability — a novel feature in *E.coli* promoter recognition. *J. Biomol. Struct. Dynam.* 1999. V. 16. P. 825–831.
55. Chasov V.V., Deev A.A., Masulis I.S., Ozoline O.N. Distribution and functional significance of A/T-tracts in promoter sequences of *Escherichia coli*. *Molecular biology.* 2002. V. 36. P. 537–542.
56. Ozoline O.N., Deev A.A., Arkhipova M.V. Noncanonical sequence elements in the promoter structure. Cluster analysis of promoters recognized by *Escherichia coli* RNA polymerase. *Nucleic Acids Res.* 1997. V. 25. P. 4703–4709.
57. Schneider T.D., Stormo G.D., Gold L., Ehrenfeucht A. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* 1986. V. 188. P. 415–431.
58. Huerta A.M., Collado-Vides J. Sigma70 promoters in *Escherichia coli*: specific transcription in dense regions of overlapping promoter-like signals. *J. Mol. Biol.* 2003. V. 333. P. 261–278.
59. Huerta A., Francino M.P., Morett E., Collado-Vides J. Selection for unequal densities of s70 promoter-like signals in different regions of large bacterial genomes. *PLoS Genetics*. 2006. V. 2. Article No. e185.

Received January 21, 2011.
Published February 03, 2011.