Математическая биология и биоинформатика. 2011. Т. 6. № 1. С. 39-52.
URL: http://www.matbio.org/2011/Ozoline2011(6_39).pdf

==========МАТЕРИАЛЫ III МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ==========
==========«МАТЕМАТИЧЕСКАЯ БИОЛОГИЯ И БИОИНФОРМАТИКА»==========

УДК 579.252

Структурообразующие модули как индикаторы промоторной ДНК в бактериальных геномах

Киселев С.С., Озолинь О.Н.*

Институт биофизики клетки, Российская академия наук, Пущино, Московская область, 142290, Россия

Аннотация. Предложена унифицированная версия компьютерной программы поиска промоторов PlatPromU, базирующаяся на эволюционной консервативности структурной организации бактериального аппарата транскрипции. В отличие от исходного алгоритма PlatProm, оптимизированного для узнавания σ^D-зависимых промоторов *Escherichia coli* (*E.coli*), новая версия не использует весовых матриц, отражающих частоту присутствия консервативных пар в элементах −10 и −35. Её предсказательный потенциал оценивали по способности распознавать известные промоторы *Corynebacterium glutamicum* (*C.glutamicum*) — эволюционно удалённого от *E.coli* микроорганизма. Оказалось, что «чувствительность» PlatPromU сопоставима с предсказательным потенциалом специализированной программы (PlatPromC), адаптированной к узнаванию регуляторных участков *C.glutamicum*, и выше, чем у исходного алгоритма PlatProm. Это значит, что унифицированная программа, моделирующая только структурно-конформационные свойства промоторной ДНК, может быть рекомендована в качестве инструмента для предварительного поиска регуляторных участков в геномах с неизвестным контекстом специфических элементов.

Ключевые слова: *универсальный алгоритм поиска промоторов, аннотация бактериальных геномов.*

1. ВВЕДЕНИЕ

Компьютерный поиск промоторов в настоящее время становится важным инструментом аннотации геномов, так как позволяет обнаруживать не только регуляторные участки для генов, кодирующих белки, рибосомные и транспортные РНК, но и места инициации синтеза нетранслируемых, антисмысловых и альтернативных РНК-продуктов [1, 2]. При этом большинство алгоритмов поиска промоторов в качестве основных индикаторов промоторной ДНК используют мотивы нуклеотидной последовательности, специфически распознаваемые σ-субъединицами РНК-полимеразы. Несмотря на очевидную эволюционную стабильность аппарата транскрипции, контекст этих мотивов весьма вариабелен как для промоторов разных бактерий, так и для промоторов одного и того же микроорганизма, предназначенных для взаимодействия с разными σ-факторами [3]. Это означает, что для поиска промоторов конкретного типа компьютерные алгоритмы нужно «настраивать» на контекст их консервативных модулей. Подобная необходимость в специфической адаптации является главным препятствием, ограничивающим использование

* ozoline@icb.psn.ru

компьютерного поиска промоторов в качестве полноценного инструмента аннотации нуклеотидных последовательностей.

В данной работе предложен подход, открывающий возможность преодолеть это препятствие с использованием алгоритма PlatProm, первоначально адаптированного для σ^D-зависимых промоторов *E.coli*. Кроме консенсусных гексануклеотидов, образующих специфические контакты с σ-субъединицей РНК-полимеразы, эта программа учитывает структурно-конформационные свойства промоторной ДНК. Из-за консервативности РНК-полимераз и структурной организации транскрипционного аппарата эти свойства должны быть инвариантными, или очень похожими в разных промоторах. Вклад структурообразующих элементов в общий показатель промотор-подобия PlatProm составляет около 50%. Полноценная версия программы распознаёт 85.5% известных промоторов *E.coli* с достоверностью $p < 0.0038$. Если учитывать только структурообразующие элементы в промоторной ДНК, чувствительность программы на этом уровне уменьшается, но остаётся высокой — 62.4%. Дискриминационный потенциал структурных модулей предоставляет, следовательно, шанс обнаружить промоторы без учёта специфических элементов.

Для строгой количественной оценки предсказательного потенциала унифицированной программы был использован геном *Corynebacterium glutamicum*. В нём экспериментальными методами было картировано 160 промоторов основного σ-фактора SigA (аналог σ^D *E.coli*), что позволило получить новую версию PlatPromC, специфически адаптированную к соответствующим промоторам. Способность унифицированной версии PlatPromU находить регуляторные участки *C.glutamicum* сравнивали с предсказательным потенциалом исходной (PlatProm) и адаптированной (PlatPromC) версий. При этом для корректного сопоставления результатов сканирования был разработан новый метод определения пороговых значений показателей промотор-подобия (скоров), обеспечивающих селекцию статистически равнозначимых сигналов при использовании разных программ поиска. Значимыми считали скоры, превышающие фоновое значение на 3, 4 и 5 стандартных отклонений (StD). На первом уровне достоверности наиболее эффективной оказалась адаптированная программа PlatPromC, но на двух других больше всего промоторов было обнаружено унифицированной программой, что свидетельствует о перспективности её использования в качестве эффективного инструмента аннотации нуклеотидных последовательностей.

2. МЕТОДЫ

2.1. Геном *C.glutamicum* и промоторы

Для сканирования и анализа была использована нуклеотидная последовательность генома *C.glutamicum* ATCC 13032 (NC_003450 в NCBI [4]) и соответствующая генная карта. Длина этого генома — 3309401 нуклеотидных пар (н.п.), содержание G/C-пар — 53%. Координаты стартовых точек транскрипции для известных промоторов *C.glutamicum* были взяты из оригинальных статей (табл. 1). Нуклеотидные последовательности промоторов были получены с использованием вспомогательной программы DNA Tools (разработана А.А. Деевым).

Таблица 1. Координаты экспериментально определённых точек инициации транскрипции для *sigA*-зависимых промоторов *Corynebacterium glutamicum*

Промот.	Точка старта и направл.	И[а]	Промот.	Точка старта и направл.	И[а]	Промот.	Точка старта и направл.	И[а]
cg0042	29965 (–)	5	*narK*	1253952 (–)	19	P-45[б]	2346400 (–)	7
cg0043	29995 (+)	5	*atp1*	1271835 (+)	20	*glnA*	2348721 (+)	7

citH	70350/2/3[в] (–)	6	*atp2*	1272131 (+)	10	*thrC*	2355050 (–)	7
P-45[б]	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
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)[г] (+)	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[б]	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			

«а» - ссылка на литературный источник, «б» - промотор P-45 присутствует в трёх копиях, «в» - множественные точки старта, «г» - точки старта, приводимые в разных источниках.

2.2. Стратегия расчёта весовых матриц PlatProm

Для оценки соответствия геномных последовательностей консервативным гексануклеотидам –35 и –10, формирующим специфические контакты с σ-субъединицей РНК-полимеразы, были использованы позиционные весовые матрицы (PWM), частотные веса в которых были рассчитаны так же, как предложено Гертцем и Стормо [51]. Каждая из матриц содержит 24 параметра k_{ij}, определяемые по формуле:

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

i – номер позиции в элементе,
j – конкретный нуклеотид (A, C, G или T),
f_{ij} – частота встречаемости нуклеотида j в позиции i,
n_j – нормировочный коэффициент, отражающий частоту присутствия j в анализируемом геноме.

Степень соответствия анализируемой последовательности консенсусным элементам определятся как сумма вкладов всех пар, расположенных в области потенциального присутствия консервативных модулей:

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

k_{ij} – вес присутствующего в соответствующей позиции анализируемой области нуклеотида, рассчитанный по формуле (1), или 0 (возможность суммирования предусмотрена для вырожденных алфавитов).

Вариации в размере спейсера (S) между элементами –35 и –10 (разрешённый диапазон $14 \le S \le 21$) и расстояния (D) от элемента –10 до стартовой точки транскрипции (разрешённый диапазон $2 \le D \le 9$) учитывали с помощью весовых матриц, отражающих частоты их встречаемости разных длин в обучающем наборе промоторов:

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

$N_{S(D)}$ – число промоторов, имеющих соответствующие S и D,

$N_{17(6)}$ – число промоторов с оптимальным S (17 н.п.) и D (6 н.п.).

При этом любое отклонение от оптимальных значений S и D приводило к снижению общего скора на величину $K_{S(D)}$. Поскольку число промоторов с очень длинными и очень короткими позиционными расстояниями мало, использование для них реальных $K_{S(D)}$ приводило к большим негативным вкладам, что сопровождалось искажённым выравниванием по консервативным гексануклеотидам. Поэтому K_S для всех промоторов с длиной спейсера >18 н.п. принимали равным K_S, определённому для промоторов со спейсером 18 н.п. В случае короткого спейсера (≤16 н.п.) использовали K_S, рассчитанное для промоторов с S = 16. Аналогичным образом снижали зависимость от D. Для промоторов, имеющих D = 4, 5, 7 или 8, значения K_D рассчитывали в соответствии с частотой встречаемости таких промоторов в компиляции. Для промоторов с D ≤ 3 использовали K_D, рассчитанное для промоторов с D = 4. K_D для промоторов с D = 9 принимали равным K_D, определённому для промоторов с D = 8.

Расчёт оптимальных позиционных весовых матриц проводили методом последовательных итераций. На первом этапе были использованы весовые матрицы, рассчитанные вручную для 30 промоторов, положение консервативных гексануклеотидов в которых было установлено генетическими методами. Модули, идентифицированные PlatProm как консенсусные гексануклеотиды в промоторах обучающей компиляции (308 негомологичных и неперекрывающихся последовательностей), использовались программой для вычисления уточнённых PWM (первая итерация). Эти PWM использовались в следующем раунде и т.д. до полной стабилизации частотных весов матриц.

Для учёта особенностей нуклеотидной последовательности вблизи стартовых точек транскрипции использовали одномерную весовую матрицу, параметры которой (k_{di}) отражают частоту встречаемости в позиции –1 каждого из 16 динуклеотидов:

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

di – конкретный динуклеотид,

f_{di} – частота встречаемости динуклеотида *di* в позиции –1 в промоторах из компиляции,

n_{di} – частота встречаемости динуклеотида *di* в геноме.

Точно так же учитывали присутствие функционально-значимого динуклеотида TG, фланкирующего 5'-конец элемента –10 (k_{TG}).

Таблица 2. Каскадная весовая матрица, отражающая повышенную частоту присутствия гибких звеньев в участке –55/–52

Позиция	Мотив	Нормированный логарифм частоты встречаемости в промоторах
-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
Отсутствуют все элементы		-0,036

Помимо этого, PlatProm учитывает особые конформационные свойства промоторной ДНК, а также модули, способствующие формированию транскрипционного комплекса и переходу его к продуктивной инициации [1–3, 52–56]. Эти элементы включают в себя:

- регулярное распределение polyA(T)-треков, взаимодействующих с α-субъединицами РНК-полимеразы или стабилизирующих транскрипционный комплекс за счёт образуемых ими анизотропных изгибов;
- гибкие YR-динуклеотиды (Y=C=T, R=A=G), способствующие адаптивной изомеризации ДНК при взаимодействии с РНК-полимеразой и регуляторными белками;
- периодическое распределение смешанных A/T-треков, предположительно принимающих участие в скольжении РНК-полимеразы по ДНК;
- прямые и инвертированные повторы, как вероятные участки взаимодействия с факторами транскрипции;
- другие мотивы, доминирующие в разных участках промоторной ДНК, выявленные для промоторов *E.coli* при помощи кластерного анализа [56].

Типичными для промоторов элементами считали такие мотивы нуклеотидной последовательности, частота присутствия которых в конкретном участке промоторной

ДНК (анализировали диапазон –250/+150 относительно точки старта) превышала фоновый уровень, по крайней мере, на 5 стандартных отклонений (StD). Все они учитываются при помощи 60 *каскадных* матриц (см. пример в табл. 2), которые отличаются от обычных PWM тем, что содержат частотные веса только для доминирующих в конкретном участке мотивов (вычисляются как нормированный натуральный логарифм частоты присутствия того или иного элемента в конкретной позиции промотора). При наличии в анализируемой последовательности нескольких перекрывающихся мотивов (например $ACAT_{-53}$ и CAT_{-52} в табл. 2) вклад в общий показатель промотор-подобия дает тот из них, «вес» которого в промоторной компиляции выше (ACAT). При отсутствии всех характерных для промоторов мотивов в конкретной последовательности назначается отрицательный вклад. Он вычисляется как логарифм доли таких промоторов в обучающем наборе.

Суммарный показатель промотор-подобия вычислялся как сумма вкладов всех весовых матриц, причём система расчёта была оптимизирована таким образом, что общий вклад каскадных матриц составляет приблизительно 50%. Для этого вклады каждой из каскадных матриц нормировали на относительное информационное содержание соответствующего участка промотора и последней пары в элементе −35 (наименее консервативная пара). Информационное содержание рассчитывали по алгоритму, предложенному в работе [57].

2.3. Определение статистически значимого порогового значения

Ранее для оценки фонового уровня и характерного для непромоторных ДНК StD было использовано два набора нуклеотидных последовательностей [1]. Один из них (CS1) состоял из 273 фрагментов кодирующих последовательностей конвергентно транскрибируемых генов *E.coli*, имеющих длину > 700 н.п. и разделённых ≤ 50 н.п. межгенным пространством. Присутствие функциональных промоторов в таких генах наименее вероятно. Второй набор содержал 400 случайных последовательностей, имеющих равный с исследуемым геномом AT/GC-состав. Использование каждого из этих наборов имеет свои преимущества и недостатки. Достоинством первой компиляции является её биологическая аутентичность, но для её создания необходимо иметь уже аннотированный геном и не всегда удаётся собрать достаточное для статистического анализа число фрагментов. Случайные последовательности можно получить в любом количестве, но среди них с определённой вероятностью окажутся и промоторы, распределение которых в геноме определяется эволюционным отбором.

В данной работе предлагается новый способ определения пороговых уровней. Он заключается в поиске участков, наименее похожих на промоторную ДНК. Для этого в режиме скользящего окна определяли среднее значение скора (F) и StD на фрагменте в 1000 н.п. (средний размер гена). Затем геном разбивали на сегменты равной длины и в каждом из них искали позицию с минимальным F ($F_{\min}$), как показано на рис. 1. Среднее значение $F_{\min}$ по всему геному ($\overline{F}$) считали фоновым уровнем, а среднее значение соответствующих им StD — характеристикой вариабельности скоров в непромоторных участках.

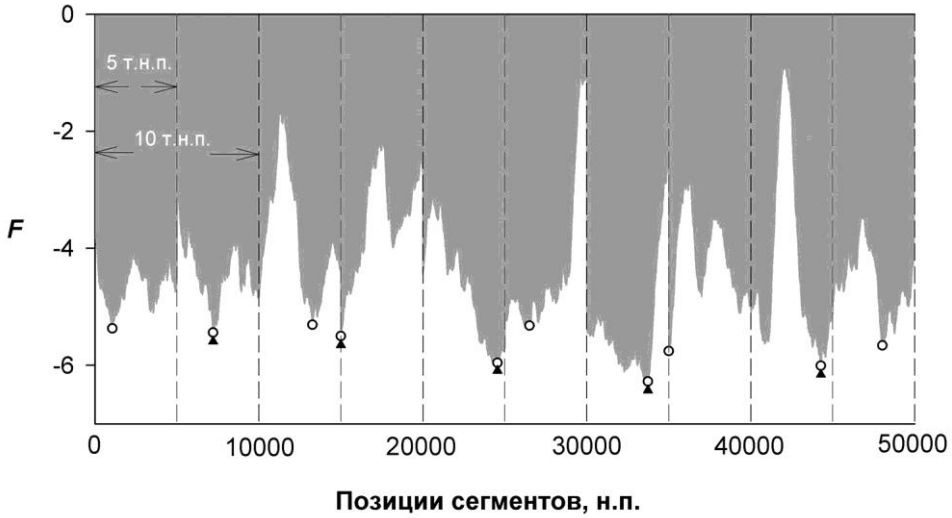

Рисунок 1. Поиск непромоторных участков для первых 50000 н.п. хромосомы *E.coli* MG1655 (NC_000913 в [4]). Серые столбики соответствуют значениям *F* для фрагментов длиной в 1000 н.п. Локальные минимумы (F_{min}), выявляемые при разбиении последовательности на сегменты длиной 5000 и 10000 н.п. отмечены окружностями и треугольниками, соответственно.

Значения StD не проявляли существенной зависимости от плотности сегментации, а величина F_{min} закономерно была тем меньше, чем больше размер сегмента (см. рис. 1 и табл. 3). Для того чтобы выбрать оптимальный размер сегмента, величины $\overline{F}$, полученные для разной плотности разбиения, сравнивали с фоновым значением, рассчитанным для контрольной компиляции CS1. Наиболее близкими эти значения оказались при разбиении генома *E.coli* на сегменты длиной 5000 н.п. Именно такие значения $\overline{F}$ были использованы для расчета трех пороговых уровней (L) при сканировании генома *C.glutamicum* разными версиями PlatProm:

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

Таблица 3. Зависимость значений $\overline{F}$ и StD от плотности сегментации

Размер сегмента (н.п.)	$\overline{F}$	StD	Второй пороговый уровень (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. Адаптация PWM PlatProm к контексту элементов −35 и −10 *C.glutamicum*

Исходя из предположения о консервативности пространственной структуры бактериального транскрипционного комплекса, для адаптации PlatProm к узнаванию промоторов *C.glutamicum* были определены частотные веса только для PWM. Изначально эти матрицы были построены с использованием обучающего набора, содержащего 308 известных промоторов *E.coli* (см. выше), а предсказательный потенциал программы оценивали с помощью тестовой компиляции (290

последовательностей), не содержащей промоторов обучающего набора [1, 2, 52]. Эта независимость обучающего и тестового наборов принципиально важна для оценки предсказательного потенциала, но ограниченное число промоторов, экспериментально картированных в геноме *C.glutamicum* (всего 160 последовательностей, причем 3 из них идентичны), не позволяет создать две независимые компиляции. Поэтому для тестирования качества специализированной программы использовали стратегию сменных мишеней. При этом каждый из известных промоторов поочерёдно был тестовым, а 157 остальных использовали для построения PWM. Но сквозное сканирование генома для определения фонового значения было осуществлено специализированной версией программы (PlatPromC), созданной на основе всех 158 промоторов.

2.5. Унификация PlatProm

Построение унифицированной программы является многостадийной задачей, решение которой может потребовать учёта дополнительных факторов и, возможно, дальнейшего упрощения расчётной системы PlatProm. В данной работе в качестве первой версии PlatPromU применялась программа, использующая только каскадные матрицы PlatProm.

2.6. Критерии оценки предсказательного потенциала компьютерных алгоритмов

«Чувствительность» (“*sensitivity*”) алгоритмов определяли как процент идентифицированных промоторов на разных уровнях достоверности. Значимыми считали скоры, превышающие фоновое значение на 3, 4 и 5 StD ($p < 0.0014$, $p < 0.00004$ и 0.000001, соответственно). Промотор считали узнанным, если предсказанная точка инициации транскрипции находилась в диапазоне ± 5 н.п. от экспериментально картированного старта. Так как экспериментальные методы допускают некоторую погрешность при определении 5’-концов РНК, промотор считали узнанным точно, если позиция предсказанной точки старта совпадала с экспериментальной, или отличалась от нее не более чем на 2 н.п.

3. РЕЗУЛЬТАТЫ

На рис. 2, в качестве примера, показаны результаты сканирования генома *C.glutamicum* вблизи гена *phoR*, кодирующего сенсорную киназу-фосфотрансферазу фосфатного регулона. Позиция стартовой точки транскрипции у этого гена находится на расстоянии 44 н.п. левее инициирующего кодона ATG [10].

Видно, что все три алгоритма (PlatProm, PlatPromU и PlatPromC) обнаруживают промотор-подобный участок перед геном *phoR*, но исходная версия программы (PlatProm) в качестве наиболее вероятного старта предлагает позицию, отстоящую от инициирующего кодона на 75 н.п., а настоящий старт не находит (средний график на рис. 2). Специализированная программа PlatPromC предсказывает его точно (красный столбик на верхнем графике), хотя тоже обнаруживает промотор со стартом в позиции −74. Сигнал в этой позиции превышает фоновый уровень на 4.99 StD, что соответствует $p<0.000001$. Это значит, что в геноме *C.glutamicum* случайным образом могут оказаться всего 7 промотор-подобных сигналов с такой амплитудой. Вероятность того, что он фальшивый, следовательно, очень мала. Кроме того, унифицированная программа (нижний график на рис. 2) предсказывает этот дополнительный старт (и настоящую точку инициации транскрипции) с очень высокой надёжностью. Скорее всего, это означает, что экспрессия гена *phoR* контролируется двумя тандемными промоторами.

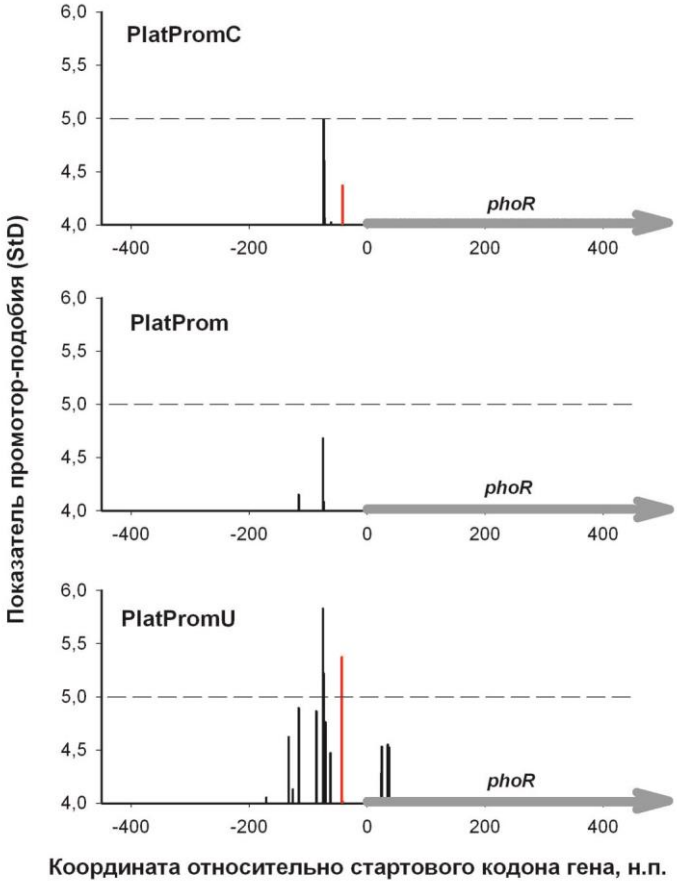

Рисунок 2. Старты транскрипции, предсказанные тремя алгоритмами для гена *phoR* (серая стрелка) *C.glutamicum*. Красным цветом обозначена настоящая точка старта. Ось X соответствует второму уровню достоверности ($\overline{F}$ + 4 StD), пунктиром отмечен третий уровень ($\overline{F}$ + 5 StD).

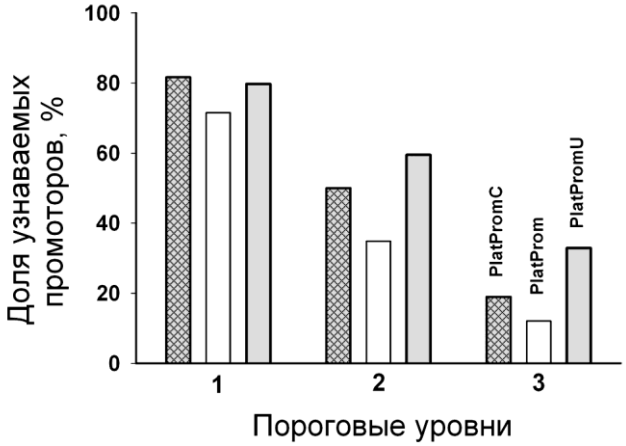

Рисунок 3. Сопоставление способности PlatPromC, PlatProm и PlatPromU (указано на графике) распознавать промоторы *C.glutamicum* на разных уровнях достоверности. Значимыми считали сигналы, превышающие фон ($\overline{F}$) на 3, 4 и 5 StD для пороговых уровней достоверности 1, 2 и 3, соответственно.

На рис. 3 показан суммарный результат сравнительного анализа. На первом уровне ($p < 0.0014$) наиболее эффективной оказалась адаптированная программа PlatPromC (заштрихованные столбики), с помощью которой удалось распознать 81.6% промоторов, т.е. столько же, сколько на этом уровне достоверности находит PlatProm в тестовой компиляции, составленной из промоторов *E.coli*, (81%, данные не показаны). Это значит, что каскадные матрицы, которые в PlatPromC остались «настроенными» на

структурно-конформационные свойства промоторов *E.coli*, равноэффективны и для поиска промоторов *C.glutamicum*. Их использование в сочетании с PWM, отражающими контекст консервативных модулей в промоторах *E.coli*, существенно снижало предсказательный потенциал программы (белые столбики). Это вполне соответствует общепринятому представлению о необходимости специфической адаптации алгоритмов поиска промоторов. «*Чувствительность*» унифицированной программы, работающей только с каскадными матрицами (серые столбики), на первом уровне практически не отличалась от «*чувствительности*» адаптированного алгоритма (79.7%), а на более высоких уровнях достоверности даже превышала её. Т.е. учитываемые каскадными матрицами структурообразующие модули промоторной ДНК действительно могут быть использованы в качестве эффективных индикаторов.

ОБСУЖДЕНИЕ

В работе сделана первая попытка создания унифицированной компьютерной программы, способной обнаруживать места инициации транскрипции в геномах с плохо изученными или совсем неизученными регуляторными элементами. Для этого была использована редуцированная программа PlatProm, в системе обсчёта которой были отключены PWM, оценивающие соответствие нуклеотидных последовательностей контексту консервативных элементов −10 и −35 σ^D промоторов *E.coli*. Для оценки эффективности этой программы были выбраны промоторы *C.glutamicum*. Этот грамположительный микроорганизм относится к классу *Actinobacteria*, в то время как грамотрицательная *E.coli* является гаммапротеобактерией. Т.е. предсказательный потенциал унифицированной версии оценивался в жёстких условиях гетерологичной генетической системы. Полученные данные не оставляют сомнений в том, что структурные свойства промоторной ДНК применимы для идентификации мест инициации транскрипции. Однако большинство из используемых в настоящее время структурообразующих модулей обогащены A/T-парами. Поэтому пока не понятно, насколько применимым окажется унифицированный алгоритм для поиска промоторов в геномах с высоким и, наоборот, с низким GC-составом.

Анализ полученных результатов свидетельствует о том, что потенциальные сигналы транскрипции, выявляемые унифицированной программой, проявляют большую тенденцию к кластеризации, чем аналогичные сигналы, обнаруживаемые специализированными алгоритмами (рис. 2). Само явление кластеризации сигналов транскрипции вблизи настоящих промоторов давно известно [1, 58, 59]. Было высказано предположение, что перекрывающиеся промотор-подобные участки используются транскрипционным аппаратом клетки для повышения локальной концентрации РНК-полимеразы вблизи транскрибируемых генетических локусов [58, 59]. Наряду с этим, в геноме *E.coli* нами были обнаружены аномально длинные (≥ 300 н.п.) «промоторные островки», эффективно взаимодействующие с РНК-полимеразой, но обладающие парадоксально низкой транскрипционной активностью [1]. Не исключено, что эти новые структурные элементы играют какую-то особую биологическую роль, не обязательно связанную с синтезом РНК. Детальный сравнительный скрининг «промоторных островков» унифицированным и специфическими алгоритмами в разных геномах может оказаться полезным для того, чтобы принять или отвергнуть предположение об участии этих новых элементов генома в структурном ремоделировании хромосомной ДНК.

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (Грант №10-04-01218).

СПИСОК ЛИТЕРАТУРЫ

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. Озолинь О.Н., Пуртов Ю.А, Брок-Волчанский А.С., Деев А.А., Лукьянов В.И. Особенности ДНК-белковых взаимодействий в транскрипционных комплексах *Escherichia coli. Молекулярная биология.* 2004. Т. 38. С. 786–797.
4. URL: ftp://ftp.ncbi.nih.gov/genomes/Bacteria/ (дата обращения: 24.01.2011).
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. Часов В.В., Деев А.А., Масулис И.С., Озолинь О.Н. A/T-треки в структуре промоторов *Escherichia coli*: характер распределения и функциональное значение. *Молекулярная биология.* 2002. Т. 36. С. 682–688.
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.

Материал поступил в редакцию 21.01.2011, опубликован 03.02.2011.