УДК 577.214:577.218:004.651

ВЛИЯНИЕ ФЛАНКИРУЮЩИХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ НА ТОЧНОСТЬ РАСПОЗНАВАНИЯ САЙТОВ СВЯЗЫВАНИЯ ТРАНСКРИПЦИОННЫХ ФАКТОРОВ

© 2014 г. **Т.М. Хлебодарова[1], Д.Ю. Ощепков[1], В.Г. Левицкий[1, 2], О.А. Подколодная[1], Е.В. Игнатьева[1], Е.А. Ананько[1], И.Л. Степаненко[1], Н.А. Колчанов[1, 2]**

[1] Федеральное государственное бюджетное учреждение науки Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия, e-mail: tamara@bionet.nsc.ru;
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

Поступила в редакцию 25 сентября 2014 г. Принята к публикации 1 октября 2014 г.

Развитие *in vitro* технологий привело к появлению новых экспериментальных данных о связывании белков с ДНК, которые накапливаются в базах данных и используются при исследовании механизмов регуляции экспрессии генов и разработке компьютерных методов распознавания сайтов связывания в геномах про- и эукариот. Однако пока не ясно, насколько *in vitro* селектированные последовательности отражают истинную структуру природных сайтов связывания транскрипционных факторов (ТФ). С использованием Кульбака – Лейблера критерия расстояний проведено сравнение сходства частотных матриц сайтов связывания ТФ, построенных на основе выборок искусственно селектированных последовательностей и природных сайтов. Показано, что для 80 % ТФ (из числа исследованных) наблюдается высокое сходство коровых последовательностей природных и искусственных сайтов. Для 20 % ТФ их *in vitro* селектированные последовательности имеют в коровой структуре сайта более широкий спектр допустимых значимых нуклеотидов, не встречающихся среди природных сайтов. Методом весовых матриц проведена оценка оптимальной длины последовательностей ДНК, включающих природные сайты связывания, при которой удается достичь максимальной точности их распознавания. Обнаружено, что примерно для 80 % ТФ (из исследованных) оптимальная для распознавания длина сайта связывания значительно превышает длину коровой последовательности и длину *in vitro* селектированных сайтов. Выявленные особенности *in vitro* селектированных сайтов связывания ТФ накладывают определенные ограничения на их использование при разработке компьютерных методов распознавания потенциальных сайтов в геномных последовательностях.

Ключевые слова: транскрипционные факторы, сайты связывания, частотные и весовые матрицы, *in vitro* селектированные последовательности.

ВВЕДЕНИЕ

Ключевым звеном тонкой регуляции экспрессии генов является структура регуляторных последовательностей промоторов генов, определяющая спектр возможных воздействий со стороны регуляторов транскрипции. Поэтому не удивительно, что поток информации по изучению механизмов регуляции транскрипции и структуры промоторов не ослабевает. В связи с этим в последнее десятилетие всё большее внимание уделяется созданию и развитию баз данных по регуляции транскрипции как у эу-, так и у прокариот (Wingender *et al.*, 2001; Lescot *et al.*, 2002; Praz *et al.*, 2002; Matys *et al.*, 2003; 2006; Kolchanov *et al.*, 2002; 2008; Munch *et al.*, 2003; Zhao *et al.*, 2005; Liu *et al.*, 2008; Grote *et al.*, 2009 и др.).

Развитие новых технологий создает большие возможности для накопления и анализа

информации о структуре регуляторных районов генов. Так, с развитием *in vitro* технологий, в частности SELEX (Systematic Evolution of Ligands by EXponential enrichment), SAAB (Selected And Amplified Binding site imprint assay), REPSA (Restriction Endonuclease Protection Selection and Amplification), CASTing (Cyclical Amplification and Selection of Targets) и других более поздних модификаций методов, например SELEX SAGE, SELEX-seq и др., используемых для селекции сайтов связывания транскрипционных факторов (Blackwell, Weintraub, 1990; Pollock, Treisman, 1990; Wright *et al.*, 1991; Hardenbol *et al.*, 1997; Roulet *et al.*, 2002; обзоры: Djordjevic, 2007; Wang *et al.*, 2011), появилось много информации о структуре сайтов связывания для различных ТФ как про-, так и эукариот.

Подобного рода данные необходимы для изучения механизмов функционирования ТФ, построения методов распознавания сайтов связывания ТФ и районов, регулирующих транскрипцию генов, а также для функциональной аннотации геномов. Созданы специализированные базы данных, предназначенные для накопления и систематизации информации об искусственно селектированных сайтах связывания ТФ (Ponomarenko *et al.*, 2000; Sandelin *et al.*, 2004; Bryne *et al.*, 2008; Newburger, Bulyk, 2009; Portales-Casamar *et al.*, 2010; Chen *et al.*, 2011 и др.). Накопление подобной информации идет также и в базе TRANSFAC, одной из наиболее известных баз по регуляции транскрипции (Matys *et al.*, 2003, 2006).

Однако вопрос о том, насколько *in vitro* селектированные последовательности отражают истинную структуру природных сайтов связывания ТФ и каковы возможности их использования для создания компьютерных методов поиска природных сайтов в геномах различных видов организмов, остается открытым. Информация об этом противоречива и неоднозначна (Robison *et al.*, 1998; Shultzaberger, Schneider, 1999; Roulet *et al.*, 2000; Ehret *et al.*, 2001).

Чтобы ответить на этот вопрос, необходим сравнительный анализ большого количества данных, полученных из разных источников. Имея значительный объем информации о структуре природных сайтов связывания ТФ в базе TRRD (Kolchanov *et al.*, 2002, 2008), мы пошли по пути объединения этих данных с данными, полученными с помощью *in vitro* технологий, и создали базу данных ArtSite (Khlebodarova *et al.*, 2006).

В этой базе накапливаются частотные матрицы, описывающие структуру как природных, так и *in vitro* селектированных сайтов связывания ТФ эу- и прокариот. (txt-файл базы может быть получен по запросу у авторов.) Матрицы получены на основе выравнивания последовательностей этих сайтов относительно наиболее консервативных нуклеотидов. В настоящее время база данных ArtSite содержит более 630 матриц, которые описывают структуру сайтов связывания более чем 300 ТФ. Из них более 100 матриц построено на основе природных, функциональных сайтов, которые описывают структуру сайтов связывания для 134 транскрипционных факторов.

Такое большое количество данных позволяет сопоставить результаты распознавания сайтов связывания ТФ, полученных методами, построенными на основе выборок природных сайтов, и искусственно селектированных последовательностей. Ранее мы сравнили структуры коровых последовательностей природных и искусственных сайтов связывания для пяти ТФ, имеющих разные ДНК-связывающие домены и, соответственно, разные типы связывания с ДНК (USF, SP1, YY1, RXR/RAR и E2F1/DP1) (Khlebodarova *et al.*, 2006). Мы получили очень высокий уровень сходства матриц, взятых из разных источников.

Эти данные позволили нам предположить, что, по крайней мере, для исследованных факторов существует возможность использования выборок *in vitro* селектированных последовательностей для распознавания потенциальных природных сайтов ТФ в геномах различных организмов. Для проверки этого предположения мы решили провести более полное сравнение матриц, построенных с использованием последовательностей сайтов, выявленных на основе селекции *in vitro*, и природных сайтов. Ресурс базы ArtSite позволил провести подобное сравнение для 35 ТФ. Кроме того, мы поставили задачу оценить оптимальную длину последовательностей ДНК природных сайтов связывания ТФ, при которой удается достичь максимальной точности их распознавания.

МАТЕРИАЛ И МЕТОДЫ

Для анализа сходства природных и *in vitro* селектированных сайтов связывания транскрипционных факторов были использованы последовательности сайтов, аннотированные в базе ArtSite, и частотные матрицы, созданные на их основе (Khlebodarova *et al.*, 2006). Использованы выборки сайтов связывания только тех транскрипционных факторов, для которых присутствовали данные и по природным, и по *in vitro* селектированным сайтам.

Для измерения сходства частотных матриц сайтов связывания ТФ был использован критерий, основанный на расстоянии Кульбака – Лейблера (Kullback – Leibler), описанный Aerts с соавт. (Aerts *et al.*, 2003). Расстояние Кульбака – Лейблера – широко известный статистический метод сравнения и оценки различия распределений. В применении к частотным матрицам он позволяет сравнивать выборки сайтов связывания транскрипционных факторов и оценивать степень их различия. Согласно критерию, значения расстояний менее и равные 0,2 определяют высокий уровень сходства матриц, от 0,2 до 0,3 – средний, более 0,3 – слабый.

Для оценки оптимальной длины сайтов связывания транскрипционных факторов, при которой достигается максимальная точность распознавания их потенциальных сайтов в геномных последовательностях, использован метод оптимизации весовых матриц (Levitsky *et al.*, 2007).При распознавании методом весовых матриц ошибка первого рода (недопредсказание сайтов из выборки обучения) была зафиксирована на уровне 50 %, а ошибка второго рода (перепредсказание случайных последовательностей, полученных из выборки обучения путем перемешивания) минимизировалась.

Точность распознавания оценена стандартным методом независимого скользящего контроля (jackknife test). Максимальная длина последовательности природного сайта, для которой проведена оценка точности распознавания, была равна 50 нуклеотидам. Нами отобрано 29 выборок сайтов связывания транскрипционных факторов, для которых оказалось возможным произвести расчеты точности распознавания для длин матриц вплоть до 50 п. н. (табл. 1).

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Анализ сходства матриц, полученных на основе природных и *in vitro* селектированных сайтов

Анализ сходства матриц, полученных на основе природных и *in vitro* селектированных сайтов связывания для 28 выборок сайтов связывания ТФ (ССТФ) эукариот с использованием критерия расстояний Кульбака – Лейблера приведен в табл. 2. Согласно полученным оценкам, расстояние менее 0,2, свидетельствующее о высоком уровне сходства природных и искусственных матриц ССТФ, показано для 80 % матриц из числа исследованных. Для шести ТФ (~20 %), а именно: C/EBPα, C/EBPβ, PEA3, PDX1, MYOD и SREBP1 (SRE тип сайтов) – это расстояние превышало 0,2, и только для одного ТФ, EGR1, оно было больше 0,3, что указывало на средний и слабый уровень сходства природных и искусственных матриц соответствующих ТФ (табл. 2). Для PEA3 средний уровень сходства (0,25) искусственных и природных матриц можно было бы объяснить видовыми различиями ТФ, так как в первом случае это были сайты связывания ТФ PEA3 *Brachydanio rerio* (рыбы), а во втором – человека и мыши.

Что касается остальных ТФ, то во всех случаях матрицы построены на основе сайтов связывания ТФ млекопитающих, причем различия в структуре сайтов C/EBPβ, EGR1 и MYOD были выражены сильнее, чем для PEA3 (табл. 2). Эти различия нельзя объяснить особенностями какого-либо ДНК-связывающего домена соответствующих ТФ, так как в рассматриваемых случаях типы доменов были разные: C/EBP – bZIP, EGR1 – Zinc finger, MYOD – bHLH, PDX1 – Homeo домен, PEA3 – Ets домен. Более того, матрицы сайтов связывания ТФ, имеющих в своей структуре те же типы ДНК-связывающих доменов, что и перечисленные выше ТФ, например CREB (bZIP), GATA (zinc finger), MYOG (bHLH) и ELK1 (Ets домен), различались незначительно (табл. 2).

Одним из возможных объяснений выявленных различий может быть то, что для некоторых ТФ формирование тонких механизмов регуляции их генов-мишеней привело к отбору сайтов с узким диапазоном аффинности, что, несомненно, отразилось на структуре сайтов.

Таблица 1

Количество и длина природных и *in vitro* селектированных последовательностей сайтов связывания транскрипционных факторов в матрицах, использованных для сравнения

Транскрипционный фактор	Кол-во последовательностей сайтов ТФ, использованных для построения матрицы		Длина последовательностей сайтов ТФ, использованных для сравнения, п. н.	
	природные сайты	*in vitro* селекция	природные сайты (max†/min‡)	*in vitro* селекция
AP2	43	185	50/9	20
AHR/ARNT	16	24	50/7	13
CEBPA	48	81	50/12	16
CEBPB	48	99	50/12	16
c-MYB	16	28	50/9	14
MYC/MAX	22	26	50/9	26
EGR1	23	55	50/9	21
ELK1	13	18	50/12	26
ETS1	59	15	50/10	10
GATA1	45	25	50/8	10
GATA2	27	53	50/7	20
GATA3	12	67	50/7	20
HMG1Y	14	15	50/10	20
HSF1	31	41	50/12	27
HSF2	13	33	50/14	27
MEF2	10	104	50/10	40
MYOD	13	24	50/7	10
MYOG	19	44	50/10	14
PDX1	10	30	50/8	10
PEA3	10	36	50/9	26
PPAR/RXR	39	72	50/14	27
USF	52	31	50/10	20
SOX5	19	23	50/7	26
SOX9	9	73	50/9	26
PU.1	23	57	50/13	16
SREBP1*	38	30	50/8	16
SREBP1**	15	7	50/10	16
SRF	13	46	50/10	26
YY1	22	55	50/10	15

Примечание. * – SRE тип сайта; ** – E-box тип сайта; † – максимальная длина последовательности сайта, для которой проведена оценка точности распознавания; ‡ – соответствует длине, при которой различие Кульбака – Лейблера матриц минимально.

Таблица 2

Оценка сходства и точности распознавания природных и *in vitro* селектированных сайтов связывания транскрипционных факторов эукариот

Транскрипционный фактор	Длина коровой последовательности сайтов, п. н. ‡	Оптимальная для распознавания длина сайта, п. н.	Расстояние Кульбака – Лейблера	Увеличение точности распознавания при опт. длине сайтов †
AP2	9	49	0,12	5,06
AHR/ARNT	7	9	0,18	2,14
C/EBPα	12	14	**0,23**	2,17
C/EBPβ	12	13	**0,27**	**1,16**
c-MYB	9	14	0,15	4,30
MYC/MAX	9	13	0,20	7,10
EGR1	9	49	**0,32**	33,0
ELK1	12	44	0,16	18,8
ETS1	10	49	0,17	18,2
GATA1	8	46	0,08	17,6
GATA2	7	49	0,19	5,06
GATA3	7	**11**	0,10	2,57
HMG1Y	10	19	0,17	**1,24**
HSF1	12	47	0,19	25,5
HSF2	14	45	0,16	23,3
MYOD	7	16	**0,26**	8,82
MYOG	10	32	0,18	3,08
PDX1	8	10	**0,21**	2,16
PEA3	9	36	**0,25**	16,3
PPAR	14	44	0,17	8,84
USF	10	31	0,13	3,26
SOX5	7	48	0,16	**162**
SOX9	9	30	0,18	18,3
PU1	13	28	0,14	27,4
SREBP1*	8	16	**0,22**	9,19
SREBP1**	10	17	0,09	2,79
SRF	10	38	0,18	32,6
YY1	10	25	0,18	2,44

Примечание. * – SRE тип сайта; ** – E-box тип сайта; ‡ – соответствует длине, при которой различие Кульбака – Лейблера матриц минимально; † – соответствует значению отношения точности распознавания сайта при оптимальной длине к точности его распознавания при учете только кора.

При искусственной селекции сайты отбираются в зависимости от условий эксперимента, которые могут быть настроены на отбор как высоко-, так и низкоаффинных сайтов, что может не соответствовать сформированной в результате эволюции структуре сайта. В этом смысле показательна ситуация с ТФ SREBP1, для которого исследователи выявили два типа сайтов, значительно различающихся по структуре. Именно те сайты, через которые осуществляется специфическая регуляция транскрипции генов в зависимости от уровня холестерина в клетке (SRE тип), сильнее отличаются от искусственных сайтов (уровень сходства 0,22), нежели те, которые участвуют в широком спектре регуляторных событий (E-box тип). Для этого типа сайтов показан высокий уровень сходства (0,09) с *in vitro* селектированными последовательностями (табл. 2). В целом, эти данные свидетельствуют о достаточно высоком уровне сходства природных и искусственных сайтов связывания ТФ эукариот.

Хотелось бы отдельно рассмотреть те случаи, когда для селекции последовательностей *in*

vitro использовали ТФ, различные по видовому происхождению. В базе имеются такие данные для пяти факторов: ETS1, c-Myb, GATA1, GATA2 и GATA3. Для сравнения использовали матрицы природных сайтов, которые построены на основе сайтов связывания человека, мыши и цыпленка. Как видно из табл. 3, для трех типов матриц (c-Myb, GATA2 и GATA3) уровень сходства высок и не зависит от видового происхождения фактора, использованного для селекции. Что же касается ETS1 и GATA1, то, согласно нашим оценкам, корректность результата поиска будет зависеть от видового происхождения ТФ, использованного для селекции сайтов. Особенно это значимо для GATA1. Анализ матриц показал, что для обоих ТФ в популяции природных сайтов, независимо от их видового происхождения, присутствует практически один тип кора, gata для GATA1 (рис. 1, *а*) и ggaa для ETS1. В селекционном эксперименте выявлены два значимых типа кора, и частота выявления второго кора, gatt для GATA1 (рис. 1, *б*, *в*) и ggat для ETS1, зависит от видового происхождения фактора. До 40 % последовательностей содержат второй тип кора при использовании для селекции сайтов GATA1 мыши (рис. 1, *в*) или ETS1 цыпленка.

Как видно из табл. 3, именно появление второго значимого нуклеотида в коровой последовательности сайта связывания сильно отражается на сходстве природных и *in vitro* селектированных матриц (рис. 1, пара сравниваемых матриц *а*/*б*, расстояние 0,0793; пара *а*/*в*, расстояние 0,3511), поскольку разделение сайтов в этой матрице только по этому признаку существенно увеличивает уровень сходства для gata-сайтов (рис. 1, пара *а*/*г*, расстояние 0,1618) и снижает таковой для gatt-сайтов (рис. 1, пара *а*/*д*, расстояние 0,4750). Это означает, что использование последних для распознавания потенциальных сайтов в геноме приведет к обнаружению большого числа не характерных для *in vivo* сайтов. Насколько широко распространено это явление, мы пока сказать не можем. Однако его наличие ставит под сомнение корректность создания широкомасштабных баз потенциальных сайтов

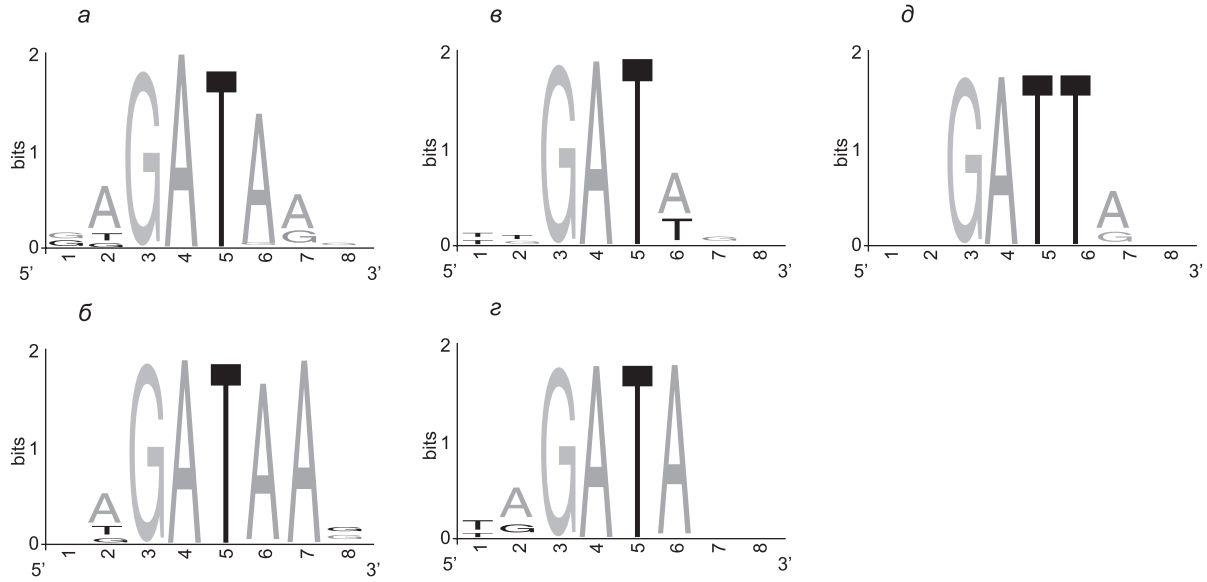

Рис. 1. Графическое изображение структуры сайта GATA1, построенное с помощью программы WebLogo (http://weblogo.berkeley.edu/):

а – матрица природных сайтов ТФ GATA1, полученная на основе последовательностей из генов человека (25), мыши (10), крысы (6) и цыпленка (6); *б* и *в* – матрицы искусственных сайтов, полученные в результате селекции с GATA1 цыпленка и мыши соответственно; *г* и *д* – матрицы искусственных сайтов с gata и gatt кором соответственно, полученные в результате селекции с GATA1 мыши. Цифры в скобках указывают количество проанализированных последовательностей природных сайтов из геномов соответствующих организмов. Ось ординат – степень консервативности нуклеотида в последовательности сайта (показана высотой буквы), ось абсцисс – позиция нуклеотида в матрице от 5′ конца сайта.

Таблица 3

Расстояние Кульбака – Лейблера между матрицами, построенными на основе природных и *in vitro* селектированных сайтов, полученных в результате селекции с ТФ различного видового происхождения

Тип матрицы	Видовое происхождение ТФ	Расстояние Кульбака – Лейблера	Длина сайтов в матрице, п. н.	Число сайтов связывания ТФ в сравниваемых парах матриц	
				in vitro селекция	природные
GATA1	Цыпленок	0,0793	8	25	47
GATA1	Мышь	**0,3511**	7	69	47
GATA1 (gata-core)	Мышь	0,1618	8	42	47
GATA1 (gatt-core)	Мышь	**0,4750**	8	24	47
GATA2	Цыпленок	0,0988	7	49	14
GATA2	Человек	0,1864	7	53	14
GATA3	Цыпленок	0,0996	7	67	12
GATA3	Человек	0,0980	7	63	12
ETS1	Цыпленок	**0,2251**	10	59	40
ETS1	Мышь	0,1731	10	15	40
c-Myb	Мышь	0,1510	9	28	16
c-Myb	Цыпленок	0,1335	11	49	16

связывания ТФ (Marinescu *et al.*, 2005) без предварительного изучения структуры их сайтов связывания.

Оценка оптимальной длины сайтов связывания транскрипционных факторов, при которой достигается максимальная точность распознавания их потенциальных сайтов

Так как длина фланговых последовательностей сайтов связывания ТФ влияет на точность их распознавания (Levitsky *et al.*, 2007), мы решили оценить оптимальную для распознавания длину сайтов связывания тех ТФ, для которых существуют выборки искусственных сайтов. Для этого мы использовали метод весовых матриц (Levitsky *et al.*, 2007) и выборки природных сайтов с увеличенными фланговыми последовательностями. Нами были построены методы динуклеотидных весовых матриц, критерием отбора длины которых стала оценка точности предсказания с помощью jack-knife теста. Не для всех 35 выборок сайтов связывания ТФ удалось набрать достаточное количество необходимых последовательностей, поэтому в данном исследовании использовано только 28 выборок. Для этих типов сайтов (табл. 2) приведены результаты анализа сходства матриц (расстояние Кульбака – Лейблера), построенных на основе коровых последовательностей природных и *in vitro* селектированных сайтов, а также изменения точности распознавания природных сайтов (по отношению ошибок второго рода) при увеличении длины их последовательности до оптимальной, т. е. такой, когда ошибка второго рода (перепредсказание) была минимальной.

Как следует из данных (табл. 2), оптимальная для распознавания длина сайтов связывания для всех исследованных ТФ превышала длину коровой последовательности сайта и только у шести ТФ (AHR, C/EBPα, C/EBPβ, MYC/MAX, GATA3 и PDX1) это превышение было незначительным (на 1–4 нуклеотида). При этом точность распознавания, изменение которой оценивали по отношению ошибок перепредсказания, увеличилась в два раза и более для большинства исследованных ТФ. А в случае ТФ SOX5 точность возросла на 2 порядка.

Исключение составили только ТФ C/EBPβ и HMG1Y. Однако и для этих ТФ увеличение длины флангов их сайтов связывания при построении матрицы улучшило точности их распознавания и снизило уровень перепред-

сказания почти на 20 % (табл. 2). Что касается столь высокого значения увеличения точности распознавания для ТФ SOX5, то анализ выборки сайтов связывания этого ТФ показал, что, в отличие от выборок других ССТФ, она содержит большое количество высокогомологичных сайтов, что, как показано нами, может приводить к завышению оценки точности распознавания. Удаление таких сайтов из выборки существенно (в три раза) снизило точность распознавания сайта.

Полученные данные позволяют предположить, что фланговые последовательности сайтов связывания практически всех исследованных ТФ эукариот содержат дополнительную информацию, необходимую для более точного распознавания их потенциальных сайтов. Например, согласно данным, представленным в табл. 2, длина коровой последовательности сайтов связывания ТФ SRF, построенной на основе выборки искусственно селектированных последовательностей, равна 10 пн. На рис. 2 она соответствует десятибуквенному мотиву CCA(T/A)ATAAGG, представленному в позициях 17–26. Оказалось, что оптимальная для распознавания длина сайта, полученная на основе анализа выборки геномных последовательностей сайтов связывания этого фактора равна 38 п. н.

Как видно на рис. 2, эта последовательность за пределами консенсуса как в 5′-, так и 3′-фланкирующих участках обогащена короткими консервативными кластерами нуклеотидов (позиции 2–4, 9, 11, 12, 15, 29, 32–35, 39). Отметим, что, хотя во всех случаях уровень консерватизма нуклеотидов оказался более низким, чем в коровой части сайта, точность распознавания сайтов связывания ТФ SRF на основе расширенной матрицы длиной 38 нуклеотидов оказалась в 32 раза выше, чем при распознавании на основе матрицы, соответствующей коровой последовательности длиной 10 нуклеотидов (см. табл. 2).

Консервативные участки в выравнивании отображены в виде стопки букв, причем высота каждой буквы пропорциональна частоте ее встречаемости в данной позиции, а общая высота стопки пропорциональна консерватизму последовательности в данной позиции, выраженной в битах (ось ординат). Номера по оси абсцисс соответствуют номеру позиции в выравнивании. Коровая часть сайта соответствует позициям 17–26.

Отметим, что рост точности распознавания при оптимизации матриц происходит благодаря двум факторам: (1) привлечению динуклеотидной, а не мононуклеотидной статистики для построения матриц и (2) собственно наращива-

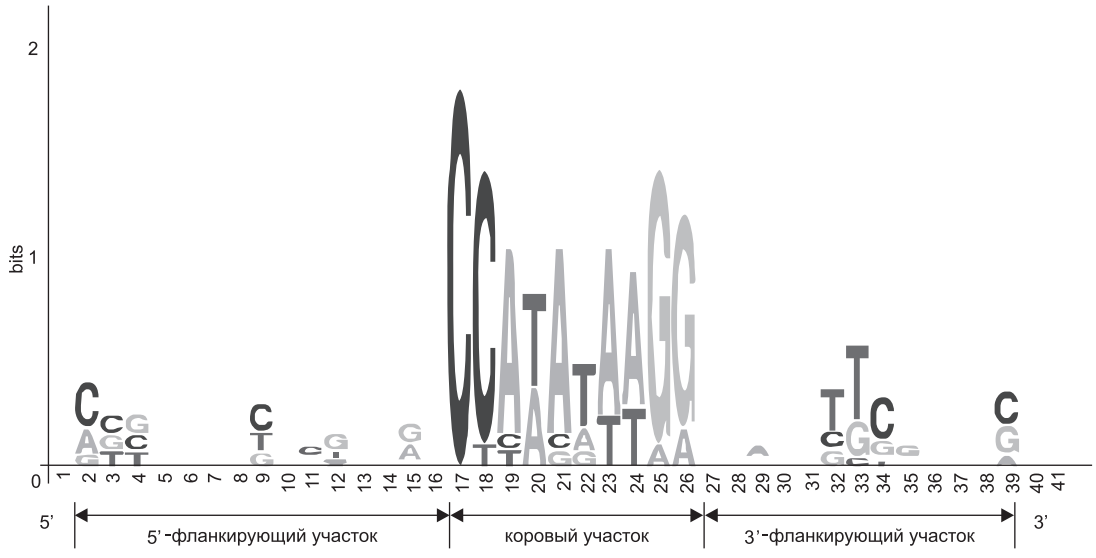

Рис. 2. Представление консервативных нуклеотидов в пределах коровой части сайта связывания фактора SRF и его фланкирующих участках, построенное с помощью программы WebLogo (http://weblogo.berkeley.edu/).

нию длины матрицы за счет менее консервативных флангов. Преимущество динуклеотидных матриц по сравнению с мононуклеотидными подтверждено как давними работами, так и современными исследованиями (Zhang, Marr, 1993; Gershenson *et al.*, 2005; Siddharthan, 2010; Nahdi, Ioshikhes, 2012). Нами ранее было показано, что привлечение фланкирующих последовательностей существенно повышает точность методов распознавания (Levitsky *et al.*, 2007; 9 типов ССТФ). Недавнее применение использованной в настоящей работе технологии учета консервативного контекста в последовательностях, фланкирующих кор сайта, к последовательностям, извлеченным из эксперимента по массовому секвенированию сайтов связывания ТФ, показывает, что длина динуклеотидной матрицы, обеспечивающей наивысшую точность, составляет около 30 нт (Kulakovskiy *et al.*, 2013; ССТФ FoxA). Хотя фланкирующие последовательности коровых районов сайтов существенно менее консервативны по сравнению с кором (рис. 2), их привлечение позволяет правильно распознавать неверные предсказания, выявляемые с помощью мононуклеотидных матриц короткой длины (10 нуклеотидов и менее).

Таким образом, проведенное сравнение структуры сайтов связывания ТФ, полученных из разных источников, позволило сделать заключение: несмотря на то что примерно для 80 % ТФ (из исследованных) показано высокое сходство коровых последовательностей природных и искусственных сайтов, использование выборок *in vitro* селектированных сайтов связывания ТФ для разработки компьютерных методов распознавания в геномных последовательностях их потенциальных сайтов не вполне корректно. Ограниченная длина селектированных сайтов (табл. 2) резко снижает точность их распознавания и приводит к выявлению значительного числа неверно предсказанных сайтов. Более того, для 20 % исследованных ТФ их *in vitro* селектированные СС имеют более широкий спектр допустимых значимых нуклеотидов в коровой структуре сайта (см. рис. 1, табл. 2). Их использование при распознавании также может привести к увеличению числа неверно предсказанных сайтов (росту ошибки перепредсказания) за счет последовательностей, вообще не встречающихся среди природных сайтов. Причина последнего, вероятно, кроется в том, что для природных сайтов связывания ряда ТФ характерна более высокая частота встречаемости среднеаффинных сайтов, нежели среди выявленных *in vitro* (Khlebodarova *et al.*, 2006), что представляет результат отбора природных сайтов скорее по функциональному критерию, чем по критерию аффинности к белку.

БЛАГОДАРНОСТИ

Работа выполнена при поддержке гранта РНФ № 14-24-00123.

Авторы выражают благодарность Н.Л. Подколодному за ценные замечания при обсуждении результатов работы.

ЛИТЕРАТУРА

Aerts S., Van Loo P., Thijs G. *et al.* Computational detection of cis -regulatory modules // Bioinformatics. 2003. V. 19. Suppl 2. P. ii5–14.

Blackwell T.K., Weintraub H. Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection // Science. 1990. V. 250. P. 1104–1110.

Bryne J.C., Valen E., Tang M.H. *et al.* JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update // Nucl. Acids Res. 2008. V. 36. P. D102–D106.

Chen L., Wu G., Ji H. hmChIP: a database and web server for exploring publicly available human and mouse ChIP-seq and ChIP-chip data // Bioinformatics. 2011. V. 27. P. 1447–1448.

Djordjevic M. SELEX experiments: new prospects, applications and data analysis in inferring regulatory pathways // Biomol. Eng. 2007. V. 24. P. 179–189.

Ehret G.B., Reichenbach P., Schindler U. *et al.* DNA binding specificity of different STAT proteins. Comparison of *in vitro* specificity with natural target sites // J. Biol. Chem. 2001. V. 276. P. 6675–6688.

Gershenzon N.I., Stormo G.D., Ioshikhes I.P. Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites // Nucleic Acids Res. 2005. V. 33. P. 2290–2301.

Grote A., Klein J., Retter I. *et al.* PRODORIC (release 2009): a database and tool platform for the analysis of gene regulation in prokaryotes // Nucl. Acids Res. 2009. V. 37. P. D61–D65.

Hardenbol P., Wang J., Van Dyke M. Identification of preferred hTBP DNA binding sites by the combinatorial method REPSA // Nucl. Acids Res. 1997. V. 25. P. 3339–3344.

Khlebodarova T.M., Podkolodnaya O.A., Oshchepkov D.Y. *et al.* ARTSITE DATABASE: Structures of natural and *in vitro* selected transcription factor binding sites //

Bioinformatics of Genome Regulation and Structure II. Ed. By N. Kolchanov and R. Hofestaedt, Springer Science+Business Media, Inc., 2006. P. 55–65.

Kolchanov N.A., Ignatieva E.V., Ananko E.A. *et al.* Transcription Regulatory Regions Database (TRRD): its status in 2002 // Nucl. Acids Res. 2002. V. 30. P. 312–317.

Kolchanov N.A., Ignatieva E.V., Podkolodnaya O.A. *et al.* TRRD: Technology for extraction, storage, and use of knowledge about the structural-functional organization of the transcriptional regulatory regions in the eukaryotic genes // Intelligent Data Analysis, 2008. V. 12. No. 5. P. 443–461.

Kulakovskiy I., Levitsky V., Oshchepkov D. *et al.* From binding motifs in ChIP-Seq data to improved models of transcription factor binding sites // J. Bioinform. Comput. Biol. 2013. V. 11. P. 1340004.

Lescot M., Dehais P., Thijs G. *et al.* PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences // Nucl. Acids Res. 2002. V. 30. P. 325–327.

Levitsky V.G., Ignatieva E.V., Ananko E.A. *et al.* Effective transcription factor binding site prediction using a combination of optimization, a genetic algorithm and discriminant analysis to capture distant interactions // BMC Bioinformatics. 2007. V. 8. P. 481.

Liu X., Yu X., Zack D.J. *et al.* TiGER: a database for tissue-specific gene expression and regulation // BMC Bioinformatics. 2008. V. 9. P. 271. doi: 10.1186/1471-2105-9-271.

Matys V., Fricke E., Geffers R. *et al.* TRANSFAC: transcriptional regulation, from patterns to profiles // Nucl. Acids Res. 2003. V. 31. P. 374–378.

Matys V., Kel-Margoulis O.V., Fricke E. *et al.* TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes // Nucl. Acids Res. 2006. V. 34. P. D108–D110.

Munch R., Hiller K., Barg H. *et al.* PRODORIC: prokaryotic database of gene regulation. Nuc. Acids Res. 2003. V. 31. P. 266–269.

Nandi S., Ioshikhes I. Optimizing the GATA-3 position weight matrix to improve the identification of novel binding sites // BMC Genomics. 2012. V. 13. P. 416.

Newburger D.E., Bulyk M.L. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions // Nucl. Acids Res. 2009. V. 37. P. D77–D82.

Pollock R., Treisman R. A sensitive method for the determination of protein-DNA binding specificities // Nucl. Acids Res. 1990. V. 18. P. 6197–6204.

Ponomarenko J.V., Orlova G.V., Ponomarenko M.P. *et al.* SELEX_DB: a database on *in vitro* selected oligomers adapted for recognizing natural sites and for analyzing both SNPs and site-directed mutagenesis data // Nucl. Acids Res. 2000. V. 28. P. 205–208.

Portales-Casamar E., Thongjuea S., Kwon AT. *et al.* JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles // Nucl. Acids Res. 2010. V. 38. P. D105–D110.

Praz V., Perier R., Bonnard C., Bucher P. The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data // Nucl. Acids Res. 2002. V. 30. P. 322–324.

Robison K., McGuire A.M., Church G.M. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome // J. Mol. Biol. 1998. V. 284. P. 241–254.

Roulet E., Bucher P., Schneider R. *et al.* Experimental analysis and computer prediction of CTF/NFI transcription factor DNA binding sites // J. Mol. Biol. 2000. V. 297. P. 833–848.

Roulet E., Busso S., Camargo A.A. *et al.* High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites // Nat. Biotechnol. 2002. V. 20. P. 831–835.

Sandelin A., Alkema W. Engstrom P., Wasserman W.W., Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles // Nucl. Acids Res. 2004. V. 32. P. D91–94.

Shultzaberger R.K., Schneider T.D. Using sequence logos and information analysis of Lrp DNA binding sites to investigate discrepancies between natural selection and SELEX // Nucl. Acids Res. 1999. V. 27. P. 882–887.

Siddharthan R. Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix // PLoS One. 2010. V. 5. P. e9722.

Wang J., Lu J., Gu G., Liu Y. *In vitro* DNA-binding profile of transcription factors: methods and new insights // J. Endocrinol. 2011. V. 210. P. 15–27.

Wingender E., Chen X., Fricke E. *et al.* The TRANSFAC system on gene expression regulation // Nucl. Acids Res. 2001. V. 29. P. 281–283.

Wright W.E., Binder M., Funk W. Cyclic amplification and selection of targets (CASTing) for the myogenin consensus binding site // Mol. Cell. Biol. 1991. V. 11. P. 4104–4110.

Zhang M.Q., Marr T.G. A weight array method for splicing signal analysis // Comput. Appl. Biosci. 1993. V. 9. P. 499–509.

Zhao F., Xuan Z., Liu L., Zhang M.Q. TRED: a Transcriptional Regulatory Element Database and a platform for in silico gene regulation studies // Nucl. Acids Res. 2005. V. 33. P. D103–D107.

EFFECT OF FLANKING SEQUENCES ON THE ACCURACY OF THE RECOGNITION OF TRANSCRIPTION FACTOR BINDING SITES

T.M. Khlebodarova[1], D.Yu. Oshchepkov[1], V.G. Levitsky[1, 2], O.A. Podkolodnaya[1], E.V. Ignatieva[1], E.A. Ananko[1], I.L. Stepanenko[1], N.A. Kolchanov[1, 2]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia,
e-mail: tamara@bionet.nsc.ru;
[2] Novosibirsk National Research State University, Novosibirsk, Russia

Summary

The development of *in vitro* methods produced new experimental information on protein binding to DNA, which is accumulated in databases and used in studies of mechanisms regulating gene expression and in the development of computer-assisted methods of binding site recognition in pro- and eukaryotic genomes. However, it is still questionable to what extent sequences selected in vitro reflect the actual structures of natural transcription factor (TF) binding sites. The Kullback – Leibler divergence was applied to the comparison of frequency matrices of TF binding sites constructed on samples of artificially selected sequences and natural sites. Core sequences of natural and artificial sites showed high similarity for 80 % of all TFs studied. For 20 % of TFs, binding site sequences selected in vitro had a broader range of permissible significant nucleotides not found in natural sites. The optimum lengths of DNA sequences including natural binding sites, at which they are recognized most accurately, were estimated by the weight matrix method. For approximately 80 % of the TFs studied, the optimum binding site length notably exceeded the lengths of the core sequences, as well as the lengths of *in vitro* selected sites. The detected features of *in vitro* selected TF binding sites impose constraints on their use in the development of computer-assisted methods of the recognition of candidate sites in genomic sequences.

Key words: transcription factors, binding sites, frequency and weight matrices, *in vitro* selected sequences.