КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2015 Т. 7 № 2 С. 347–359

Ки&М

АНАЛИЗ И МОДЕЛИРОВАНИЕ СЛОЖНЫХ ЖИВЫХ СИСТЕМ

УДК: 519.8

Определение промоторных и непромоторных последовательностей *E.coli* по профилям их электростатического потенциала

Е. А. Темлякова[a], А. А. Сорокин[b]

Институт биофизики клетки РАН,
Россия, 142290, г. Пущино, ул. Институтская, д. 3

E-mail: [a] evgenia.teml@gmail.com, [b] lptolik@gmail.com

Получено 16 сентября 2013 г.,
после доработки 26 января 2015 г.

В рамках данной работы была продемонстрирована возможность использования характеристик профилей электростатического потенциала вдоль последовательностей ДНК для определения их функционального класса. Построены модели, позволяющие разделять промоторные и непромоторные последовательности (случайные бернуллиевские, кодирующие и псевдопромоторы) с точностью порядка 83–85 %. Определены наиболее значимые участки для такого разделения, по-видимому играющие важную роль при ДНК-полимеразном узнавании.

Ключевые слова: электростатические свойства ДНК, поиск промоторов, PLS–DA, VIP-анализ

Detection of promoter and non-promoter *E.coli* sequences by analysis of their electrostatic profiles

E. A. Temlyakova, A. A. Sorokin

Institute of Cell Biophysics, 3, Institutskaya st., Pushchino, 142290, Russia

Abstract. — The article is devoted to the idea of using physical properties of DNA instead of sequence along for the aspect of accurate search and annotation of various prokaryotic genomic regions. Particulary, the possibility to use electrostatic potential distribution around DNA sequence as a classifier for identification of a few functional DNA regions was demonstrated. A number of classification models was built providing discrimination of promoters and non-promoter regions (random sequences, coding regions and promoter-like sequences) with accuracy value about 83–85 %. The most valueable regions for the discrimination were determined and expected to play a certain role in the process of DNA-recognition by RNA-polymerase.

Keywords: electrostatics of DNA, promoter location, PLS–DA, VIP-analysis

Citation: *Computer Research and Modeling*, 2015, vol. 7, no. 2, pp. 347–359 (Russian).

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 14-04-31793-мол_а, а также совместно РФФИ и Московской областью в рамках научного проекта № 14-44-03679-р_центр_а.

Введение

Известно, что молекула ДНК — это двухцепочечный полимер, цепи которого образованы комбинацией четырех типов мономеров-нуклеотидов: аденина (А), тимина (Т), гуанина (Г) и цитозина (Ц), формирующих правую двойную спираль за счет водородного связывания комплементарных пар. Существуют методы предсказания по нуклеотидной последовательности таких физических характеристик молекулы ДНК, как геометрия двойной спирали [Travers, 1989; Loziński, Wierzchowski, 1996], гибкость [Drew, Travers, 1984], термодинамическая стабильность [Jost, Everaers, 2009; Yeramian, 2000] и устойчивость двойной спирали к суперскручиванию [Wang, Benham, 2006; Benham, 1996; Benham et al., 1997]. Однако для широкого круга практических и теоретических задач оказывается достаточным рассмотрение только последовательности ДНК в виде строки букв АТГЦ.

При исследовании генетических текстов в основном используется лексический (текстовый) анализ последовательностей. Он интуитивно понятен и обычно не требует длительных и громоздких вычислений. При помощи текстового анализа генетических последовательностей были решены такие задачи, как поиск повторов, устойчивых мотивов в определенных участках генома, сайтов рестрикции и старт-кодонов, а также оценка гомологии последовательностей. Это дало немало новых знаний об устройстве генетического аппарата различных организмов и их филогенетическом положении.

Существует ряд задач, применительно к которым текстовый анализ последовательностей ДНК не работает или дает неудовлетворительные результаты. Как правило, это задачи, направленные на поиск последовательностей, нуклеотидный состав которых сильно варьирует и не имеет устойчивых паттернов. В большинстве своем эти последовательности не являются кодирующими, осуществляя структурные или регуляторные функции в геноме. К таким последовательностям относятся промоторы — важнейшие функциональные участки ДНК, расположенные перед точкой старта транскрипции, узнаваемые бактериальной РНК-полимеразой и обеспечивающие инициацию транскрипции. Экспериментально было показано, что замена нуклеотидов в определенных позициях промоторных последовательностей влияет на силу связывания промоторов с РНК-полимеразой таким образом, что аффинность может изменяться в пределах 2-х порядков [Schaumburg, 2003; Kiryu et al., 2005]. Это подтверждает основную гипотезу, согласно которой различия нуклеотидного состава промоторов необходимы клетке для контроля экспрессии отдельных генов на уровне промоторно-полимеразного взаимодействия [Barker et al., 2001a; Barker et al., 2001b].

Методы, использующие нуклеотидную последовательность для предсказания положения промоторов на хромосоме, обладают слишком низкой селективностью, и их использование приводит к заметному количеству ложных срабатываний [Киселев, Озолинь, 2011; Tutukina et al., 2007; Panyukov et al., 2013; Mendoza-Vargas et al., 2009]. Вероятнее всего, это связано с тем, что текстовый анализ применим только для описания взаимодействия нуклеиновых кислот между собой. В тех случаях, когда текстовый анализ давал хорошие результаты, например при предсказании открытых рамок считывания или кодирования аминокислотных последовательностей, узнавание и связывание осуществляются за счет комплементарного спаривания нуклеиновых оснований в рибосомальной РНК, транспортной РНК и т. д., а белки осуществляют лишь структурные функции. Сложности текстового анализа возникают при поиске участков ДНК, для которых белок становится распознающим элементом, например сайты посадки транскрипционных факторов и ферментов системы модификации-рестрикции, места посадки нуклеосом. В этом случае необходимо искать характерные особенности физико-химических свойств ДНК, отличающих искомые участки от всех остальных. В ряде случаев такие характеристики удается найти экспериментально, так, например, белок ocr фага T7 ингибирует систему бактериальной рестрикции за счет имитации структуры и электростатических свойств последовательности длиной 24

п.о., обычно распознаваемой бактериальной рестриктазой [Stephanou et al., 2009; Tsonis, Dwivedi, 2008].

Различные группы исследователей рассматривали в качестве возможных промоторных детерминант такие физические свойства и геометрические характеристики молекулы ДНК, как электростатический потенциал [Polozov et al., 2005; Камзолова и др., 2009; Sorokin et al., 2006; Kamzolova et al., 2005; Polozov et al., 1999], устойчивость к суперскрученности [Wang, Benham, 2006; Benham, 1996; Benham et al., 1997], термодинамическую стабильность [Jost, Everaers, 2009; Yeramian, 2000], изгибность цепи [Rohs et al., 2009], способность формировать изломы [Jensen et al., 1999], а также доступность и расположение доноров и акцепторов водородных связей в большой борозде промоторной ДНК [Weindl et al., 2009]. Все эти характеристики играют важную роль на различных этапах взаимодействия ДНК и РНК-полимеразы: при первичном узнавании, формировании комплекса «белок–ДНК», синтеза первичного РНК-транскрипта и формировании элонгирующего комплекса. При этом из перечисленных выше свойств только электростатический потенциал может «узнаваться» РНК-полимеразой на расстоянии, начиная со стадии диффузии и до формирования молекулярных контактов. Более того, пример «молекулярной мимикрии» [Stephanou et al., 2009; Tsonis, Dwivedi, 2008] доказывает, что электростатические взаимодействия играют важную роль в белково-нуклеиновом узнавании.

Материалы и методы

Электростатический потенциал молекулы ДНК

Молекула ДНК — один из наиболее сильно заряженных биополимеров. Основной вклад в общий заряд молекулы вносят нескомпенсированные фосфатные группы сахаро-фосфатного остова. Молекулярное окружение частично компенсирует этот заряд противоионами. Как правило, для вычисления распределения электростатического потенциала вокруг молекулы ДНК используют решения уравнения Пуассона-Больцмана [Polozov et al., 2005; Koehl, 2006; Jayaram et al., 1989; Misra et al., 1998], которые дают возможность описать электростатические взаимодействия с хорошей точностью:

$$-\nabla\left(\varepsilon(\vec{R})\nabla V(\vec{R})\right) + \sum_i c_i \mathrm{e} z_i e^{-\frac{\mathrm{e} z_i V(\vec{R})}{kT}} = \frac{4\pi}{kT}\rho(\vec{R}), \quad (1)$$

где $V(\vec{R})$ — электростатический потенциал, $\varepsilon(\vec{R})$ — диэлектрическая проницаемость, $\rho(\vec{R})$ — распределение заряда молекулы ДНК, c_i, z_i — концентрация и заряд i-го иона раствора, e — заряд электрона. Несмотря на последние достижения в вычислительной технике и новые алгоритмы, уравнение Пуассона–Больцмана удается использовать для исследования только относительно коротких фрагментов ДНК (100–200 п.о.), при этом высокая плотность заряда ДНК не позволяет использовать лианеризованную форму уравнения, что еще больше усложняет вычисления.

Ранее нами был разработан подход [Sorokin et al., 2006; Kamzolova et al., 2005; Polozov et al., 1999], позволяющий получать достоверную оценку характеристик электростатических взаимодействий для больших последовательностей и даже целых геномов. Он может быть использован для оценки поведения электростатического потенциала прокариотических геномов и любых других двухцепочечных молекул ДНК, представленных в линейном виде и не упакованных в надмолекулярные структуры. Полное описание метода можно найти в [Polozov et al., 1999], ниже приведены основные принципы нашего подхода к вычислению профиля электростатического потенциала.

Для расчета электростатического потенциала строилась полноатомная модель молекулы ДНК. При ее построении учитывались геометрические характеристики исследуемой последовательности — расстояние между соседними парами оснований (rise) и угол поворота между

соседними парами оснований (twist). Заряды помещались в центр атомов; величины зарядов были определены в работе [Журкин и др., 1980] как сумма σ- и π-зарядов атомов.

Значение электростатического потенциала $V(\vec{R})$ вычислялось по формуле Кулона

$$V(\vec{R}) = \sum_i \frac{Q_i}{\varepsilon(\vec{R})R_i} \tag{2}$$

на поверхности цилиндра радиуса 15 Å, ось которого совпадает с осью двойной спирали ДНК. Полученное распределение усреднялось по азимутальному углу и в дальнейшем использовалось только одномерное распределение электростатического потенциала вдоль исследуемой последовательности ДНК с шагом в 1 Å [Камзолова и др., 2007].

Для того чтобы учесть влияние противоионов раствора вокруг молекулы ДНК, эффективный заряд на фосфатных группах был снижен вдвое. Помимо этого, диэлектрическая постоянная была определена как функция расстояния $\varepsilon(\vec{r}) = |\vec{r}|$. Указанные модификации позволили приблизить результаты вычислений к тем, что наблюдаются при использовании формулы Пуассона-Больцмана [Polozov et al., 1999].

Представленный метод позволяет расчитывать распределения вдоль целых бактериальных геномов с минимальными временными затратами и позволяет получить представление об основных характеристиках распределения электростатического потенциала вдоль молекулы ДНК. Использование нами зарядовой схемы из работы [Журкин и др., 1980] обусловлено главным образом необходимостью обеспечения согласованности результатов исследований разных лет. Как показал анализ результатов построения одномерных профилей электростатического потенциала с использованием нелинейного уравнения Пуассона–Больцма (1), разница в одномерных профилях между зарядовыми схемами AMBER и Charmm была только в положении средней линии потенциала (Charmm давал слегка повышенный потенциал по сравнению с AMBER), при этом разница между формой профилей не превышала 0.5 %. Более того, разница между профилем, полученным решением уравнения Пуассона–Больцмана (1) и вычисленным по формуле (2), в среднем не превосходит 1 %, а максимальное отклонение меньше 10 % [Сорокин].

Промоторные и непромоторные последовательности

В нашем исследовании мы использовали последовательности полной хромосомы *E.coli*, взятой из GenBank (NC 000913). Информация о положении точек старта транскрипции была взята из RegulonDB 7.5 [Gama-Castro et al., 2011]. При этом использовались только σ^{70}-промоторы, имеющие экспериментальное подтверждение, — всего 699 последовательностей.

В качестве непромоторных последовательностей рассматривались 3 класса последовательностей:

- **rand**: случайные нескоррелированные бернуллиевские последовательности;
- **cod**: кодирующие последовательности, взятые из центральной части кодирующих участков генов *E.coli*;
- **isl**: псевдопромоторы из промоторных «островков» [Tutukina et al., 2007].

Все исследуемые последовательности рассматривались в интервале $[-540, +179]$ Å, где 0 соответствовал положению центра пары оснований точки старта транскрипции. В случае построения выборки кодирующих последовательностей 0 выбирался случайным образом так, чтобы ближайшая точка старта транскрипции находилась на расстоянии не менее 900 Å (приблизительно 250 п.о.). В итоге нами рассматривались 720 значений электростатического потенциала вдоль последовательностей ДНК, в интервале, приблизительно соответствующем $[-150, +50]$ п.о.

Логотипы мотивов и весовых матриц

Логотипы мотивов и весовых матриц были построены при помощи онлайн-приложения WebLogo версии 2.8.2. [Crooks, 2004].

Дискриминантный анализ

Дискриминантный анализ проводился в вычислительной среде R/Bioconductor с использованием пакетов pls [Mevik et al., 2011] и reldna [Sorokin et al.]. В качестве матрицы дескрипторов выступали значения электростатического потенциала в интервале координат $[-540, +179]$ Å, точка старта транскрипции при этом помещалась в 0. Матрица ответов содержала в себе информацию о принадлежности последовательностей к одному из классов (промоторным и непромоторным последовательностям). Столбцы матриц были предварительно центрированы вокруг среднего значения и масштабированы на величину стандартного отклонения.

Результаты и их обсуждение

Структура промотора, промоторные «островки» и ложно-положительные сигналы

Промоторами называют последовательности ДНК длиной 60–70 п. о., расположенные перед точкой старта транскрипции. В ряде экспериментов было показано, что при полимеразном узнавании промоторов наибольшую роль играют два гексануклеотида, расположенные на расстоянии в 10 и 35 п.о. перед точкой старта транскрипции, так называемые −10- и −35-области. Для них были вычислены консенсусные последовательности — TATAAT и TTGACA соответственно (см. рис. 1). Однако в реальных промоторах практически не встречается полного совпадения с этими каноническими последовательностями [Киселев, Озолинь, 2011; Mendoza-Vargas et al., 2009]. Помимо гексануклеотидов −10- и −35-областей были обнаружены AT-богатые участки, расположенные в up-stream-области (далее −40-позиции), с ними взаимодействует одна из субъединиц РНК-полимеразы [Hirvonen et al., 2001; Borukhov, Nudler, 2008], что обеспечивает дополнительные контакты между белком и промоторной последовательностью.

По-видимому, наиболее важной для полимеразного узнавания является −10-область, она может быть обнаружена у большинства промоторов. При этом, в отсутствие других элементов, этот гексануклеотид может быть представлен в виде нонануклеотида вида GTNTATAAT (так называемая «расширенная» −10-область) [Borukhov, Nudler, 2008]. Динуклеотид GT также связывается с РНК-полимеразой, создавая благоприятные условия для распознания промотора и формирования ДНК-белкового комплекса [Hook-Barnard, Hinton, 2009; Yuzenkova et al., 2011].

Алгоритмы поиска промоторов, опирающиеся на текстовый анализ последовательностей, часто используют весовые матрицы, содержащие информацию о частотах встречаемости нуклеотидов в той или иной позиции. Помимо этого, могут учитываться наличие AT-богатых последовательностей в up-stream-области и длина спэйсера между гексануклеотидами −10- и −35-области. Используя эти характеристики в качестве критериев отбора, можно добиться неплохих показателей специфичности метода. Многие алгоритмы поиска промоторов демонстрируют действительно высокую специфичность, но при этом обладают довольно низкой селективностью. Это значит, что при достаточном числе обнаруженных промоторных последовательностей определяется большое число ошибочных положительных предсказаний. Интересно, что некоторые из таких участков хоть и не являются промоторами, но могут распознаваться РНК-полимеразой и связываться с ней. Ввиду возникновения полимеразного узнавания такие предсказания некорректно называть ложно-положительными и стоит рассматривать отдельно. Именно к такому типу последовательностей относятся промоторные «островки» [Tutukina et al., 2007].

Промоторные «островки» — это особые участки прокариотических геномов, содержащие большое количество близкорасположенных промотороподобных участков. В работах [Tutukina

Рис. 1. Текстовые элементы, встречающиеся в промоторных последовательностях

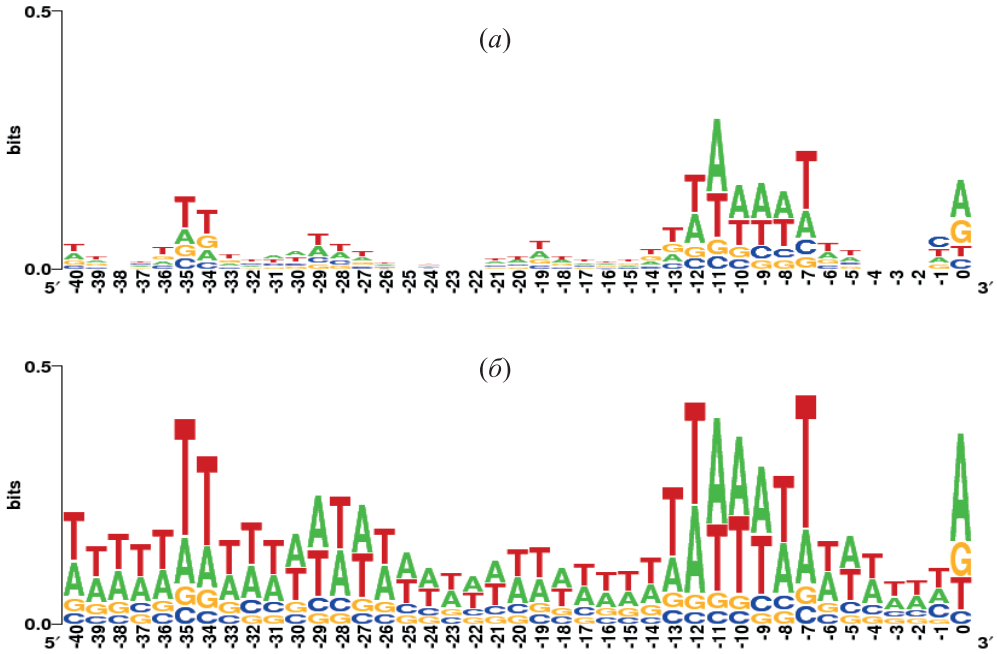

Рис. 2. Закономерности нуклеотидного состава (*а*) промоторов и (*б*) псевдопромоторов, расположенных в промоторных «островках». По горизонтальной оси отложено расстояние до точки старта транскрипции, по вертикальной — информационное содержание. Для каждого нуклеотида последовательности показаны общее информационное содержание и вклад нуклеотидов конкретного типа; так, на рисунке (*а*) наибольший вклад дает нуклеотид в положении −11, при этом буквы A и T в этом положении равновероятны

et al., 2007; Shavkunov et al., 2009] в геноме *E.coli* было выявлено 72 промоторных «островка», для которых при помощи анализа нуклеотидной последовательности (Platprom) были определены более 4000 псевдоточек старта транскрипции. Экспериментальное исследование этих участков методом ChIP-chip показало, что с некоторыми из псевдопромоторов связывается бактериальная РНК-полимераза. Фермент может задерживаться на данных участках, однако синтеза РНК-транскриптов при этом не наблюдается. Функции и значение промоторных «островков» для генома прокариот пока неизвестны.

Нуклеотидный состав промоторных «островков» характеризуется высоким содержанием АТ-пар (49–92 % с медианой 74 %, в то время как для 699 экспериментальных промоторов она составляет 31–89 %, с медианой 61 %). На рисунке 2 представлены мотивы для 699 промоторов и такого же количества случайным образом выбранных псевдоточек старта транскрипции. Видно, что области −10- и −35-псевдопромоторов имеют ярко выраженную структуру на фоне общего обогащения АТ-парами, в то время как у экспериментально подтвержденных промоторов степень выраженности указанных мотивов гораздо ниже. Очевидно, что методы, направленные на поиск закономерностей в −10- и −35-областях, классифицируют псевдопромоторы из промоторных «островков» как промоторные последовательности с высокими показателями достоверности.

Дискриминантный анализ

Для построения моделей мы использовали дискриминантный анализ при помощи проекций на латентные структуры (projections on latent structures discriminant analysis, PLS–DA) [Le Cao et al., 2011]. Привлекательность данного классификационного метода обусловлена следующими возможностями:

- работа с большим количеством зашумленных и корреллирующих переменных;
- сокращение размерности исходных данных при минимальной потере значимой информации;
- исследование вклада каждой из переменных в процесс разделения классов, определение скрытых связей между переменными.

Для построения классификационной модели необходимо составить бинарную матрицу Y, содержащую данные о принадлежности объектов к одному из классов (матрица ответов), и матрицу X, содержащую все признаки, по которым предполагается разделять классы объектов (матрица дескрипторов). Посредством PLS–DA между матрицей ответов Y и матрицей дескрипторов X строится линейная (полиномиальная) зависимость вида

$$Y = BX + E,$$

где E — матрица остатков.

Методы PLS имеют простую геометрическую интерпретацию. Если представить матрицы X и Y как облако точек в двух гиперпространствах, то создание моделей при помощи PLS при этом будет представлять собой одновременную проекцию пространств X и Y на гиперпространство с меньшей размерностью A. Новые координаты этих точек при этом будут составлять значения матриц T и U — матриц счетов. Проекции строятся таким образом, чтобы хорошо описывать пространства X и Y и добиться максимальной корреляции между ними. Таким образом, главной задачей PLS-методов является поиск максимальной ковариации между координатами точек в гиперпространстве или определение максимальной ковариации между матрицами счетов T и U. Корреляция между матрицами U и X выражена через матрицу взвешенных нагрузок W, которая используется для вычисления значений счетов T.

Исходные матрицы получают следующее представление:

$$X = 1 \times \bar{X} + TP + E, \tag{3}$$

$$Y = 1 \times \bar{Y} + UQ + F, \tag{4}$$

$$U = T + H, \tag{5}$$

где E, F и H — матрицы остатков.

Предсказания для новых объектов вычисляют следующим образом:

$$Y = X_{\text{new}}B, \tag{6}$$

$$B = W(P^T W)^{-1}Q^T, \tag{7}$$

PLS–DA также предоставляет возможность анализа вклада каждой j-ой переменной в процесс построения новых переменных a при помощи метода, называемого VIP-анализом (variable importance on projections). VIP-значения переменных можно определить по следующей формуле:

$$VIP_{Ak} = \sqrt{\sum_{a=1}^{A}(w_{ak}^2(SSY_{a-1} - SSY_a)) \times \frac{I}{SSY_0 - SSY_A}}, \tag{8}$$

где SSY_a — выборочная дисперсия элементов проекции матрицы Y на латентную переменную a. Чем выше VIP-значение, тем большую роль в процессе построения модели играет данная переменная.

В нашем исследовании при построении классификационных моделей в качестве матрицы дескрипторов X выступали значения электростатического потенциала с шагом, равным 1 Å, в интервале координат $[-540, +179]$ Å, точка старта транскрипции при этом помещалась в 0. Таким образом, все нуклеотидные последовательности были выровнены относительно точки старта транскрипции и были описаны 720-ю значениями электростатического потенциала вдоль них. Матрица ответов Y содержала в себе информацию о принадлежности последовательностей к одному из классов (промоторным и непромоторным последовательностям).

Предсказание промоторов при помощи PLS–DA

В результате дискриминантного анализа PLS–DA были построены 3 класса моделей, обученных разделять промоторы и непромоторные последовательности (случайные бернуллиевские и кодирующие последовательности, псевдопромоторы из промоторных «островков»). При этом были реализованы 5 разных соотношений тренировочных и тестовых частей: 10 к 90, 20 к 80, 30 к 70, 40 к 60 и 50 к 50. В качестве данных для разделения последовательностей использовались только значения электростатического потенциала с шагом в 1 Å, выровненные относительно точки старта транскрипции. Предсказательная способность моделей оценивалась по точности (Acc), селективности (Sens) и специфичности (Spec), которые вычисляли по формулам

$$\text{Sens} = \frac{TP}{TP + FN}, \quad \text{Spes} = \frac{TN}{FP + TN}, \quad \text{Acc} = \frac{TP + TN}{FP + TP + FN + TN}. \tag{9}$$

Результаты предсказаний для наших моделей представлены в таблице 1. Точность предсказания для всех трех моделей колеблется в пределах 83–85 %. При уменьшении доли тренировочных выборок точность практически не изменяется, что, возможно, указывает на значительные различия профилей электростатического потенциала в наборах промоторных и непромоторных последовательностей.

По сравнению с методами, использующими для построения предсказаний текстовый анализ последовательностей, показатели селективности наших моделей выше показателей специфичности. Таким образом, используя электростатический потенциал как характеристику поиска промоторных последовательностей, мы наблюдаем меньшее число ложно-положительных предсказаний.

Таблица 1. Результаты предсказаний дискриминантного анализа PLS–DA. Приведены значения точности (Acc), селективности (Sens) и специфичности (Spec) для каждой из моделей: rand — случайные бернуллиевские последовательности, cod — кодирующие последовательности, isl — псевдопромоторы из промоторных «островков» и пяти разных соотношений тренировочных и тестовых частей

Модель		90/10	80/20	70/30	60/40	50/50
rand	Acc	84.1 ± 4.4	83.9 ± 1.6	83.8 ± 1.2	83.4 ± 1.3	83.3 ± 1.1
	Sens	81.4 ± 7.6	81.2 ± 4.3	80.6 ± 4.3	80.5 ± 3.8	80.5 ± 3.3
	Spec	86.8 ± 5.4	86.6 ± 4.4	87.0 ± 4.3	86.3 ± 4.4	86.0 ± 4.1
cod	Acc	84.6 ± 4.3	85.1 ± 2.4	85.1 ± 1.9	85.0 ± 2.0	84.7 ± 1.4
	Sens	82.8 ± 7.4	83.0 ± 3.7	83.0 ± 2.8	82.7 ± 2.9	82.6 ± 1.9
	Spec	86.5 ± 3.2	87.3 ± 2.7	87.3 ± 2.2	87.4 ± 1.9	86.9 ± 1.9
isl	Acc	83.9 ± 5.8	83.8 ± 2.5	83.3 ± 1.9	83.1 ± 1.2	82.6 ± 0.9
	Sens	81.0 ± 9.5	81.0 ± 6.8	81.1 ± 7.1	81.2 ± 6.3	80.7 ± 6.1
	Spec	86.8 ± 9.9	86.6 ± 6.7	85.6 ± 6.8	85.1 ± 6.8	84.5 ± 6.5

Другие методы предсказаний

Для сравнения результатов нашего алгоритма с существующими методами использовались три алгоритма текстового анализа последовательностей:

- BPROM — алгоритм, использующий линейный дискриминантный анализ для предсказания положения точек старта транскрипции [BPROM; Solovyev, Salamov, 2011]; считается одним из наиболее популярным в применении;
- NNPP — предсказания строятся при помощи нейронных сетей [Reese, 2001];
- PlatProm — отечественный алгоритм, рассчитывающий показатель «промотороподобия» для каждого нуклеотида последовательности [Brok-Volchanski et al., 2006].

Были вычислены точность (Acc), селективность (Sens) и специфичность (Spec) этих методов для тестовой выборки, содержащей промоторы и кодирующие последовательности из генов *E.coli*. Результаты вычислений приведены в таблице 2.

Таблица 2. Результаты предсказаний для алгоритмов с текстовым анализом последовательностей. Приведены значения точности (Acc), селективности (Sens) и специфичности (Spec) для четырех различных алгоритмов предсказания промоторов и пяти разных соотношений тренировочных и тестовых частей. Наилучшие значения для каждого соотношения выделены черным

Модель		90/10	80/20	70/30	60/40	50/50
BPROM	Acc	57.3 ± 1.6	57.2 ± 1.0	57.2 ± 0.7	57.2 ± 0.5	57.2 ± 0.4
	Sens	27.7 ± 4.3	27.6 ± 2.8	27.6 ± 1.8	27.6 ± 1.6	27.6 ± 1.4
	Spec	96.8 ± 2.8	96.7 ± 1.5	96.7 ± 1.2	96.7 ± 1.0	96.7 ± 0.8
NNPP	Acc	54.4 ± 1.0	54.4 ± 0.6	54.4 ± 0.4	54.4 ± 0.3	54.4 ± 0.2
	Sens	19.4 ± 3.8	19.4 ± 2.6	19.3 ± 1.8	19.3 ± 1.3	19.3 ± 0.9
	Spec	96.0 ± 2.7	96.1 ± 1.9	96.1 ± 1.4	96.1 ± 1.0	96.1 ± 0.7
PlatProm	Acc	67.7 ± 3.3	67.5 ± 1.6	67.5 ± 1.0	67.5 ± 0.8	67.5 ± 0.6
	Sens	52.8 ± 6.7	52.6 ± 3.3	52.5 ± 1.8	52.5 ± 1.7	52.5 ± 1.2
	Spec	**98.6 ± 2.0**	**98.5 ± 1.2**	**98.5 ± 0.8**	**98.5 ± 0.6**	**98.5 ± 0.5**
PLS–DA	Acc	**84.6 ± 4.3**	**85.1 ± 2.4**	**85.1 ± 1.9**	**85.0 ± 2.0**	**84.7 ± 1.4**
	Sens	**82.8 ± 7.4**	**83.0 ± 3.7**	**83.0 ± 2.8**	**82.7 ± 2.9**	**82.6 ± 1.9**
	Spec	86.5 ± 3.2	87.3 ± 2.7	87.3 ± 2.2	87.4 ± 1.9	86.9 ± 1.9

Из таблицы 2 видно, что при сравнимых значениях специфичности наш метод дает на 30–40 % большую селективность. При анализе полной хромосомы это означает в два раза меньшее количество ложных срабатываний.

Также были исследованы закономерности нуклеотидного состава для положительных предсказаний методов BPROM, NNPP, PlatProm, нашего алгоритма (для модели cod) и последовательностей реальных промоторов *E.coli*. Результаты представлены на Рисунке 3. Для алгоритмов, основанных на текстовом анализе (рис. 3, *а–в* основной вклад в построение предсказаний вносит −10-область. Алгоритмы NNPP и PlatProm, вероятно, также учитывают вклад −35-области. Структура положительных предсказаний метода PLS–DA наиболее близка к структуре реальных промоторов; малая величина информационного содержания на рисунке 3, *г* свидетельствует о большой вариабельности промоторных последовательностей и отсутствии четких консенсусов, наблюдаемых на рисунках 3, *а–в*. Необходимо также отметить, что у всех текстовых методов степень выраженности мотивов в −10- и −35-областях заметно превышает степень выраженности этих мотивов в реальных промоторах, именно поэтому для достижения приемлемой специфичности разработчики данных алгоритмов были вынуждены пожертвовать селективностью. В случае метода PLS–DA текстовые мотивы, наоборот, выражены слабее, чем у реальных промоторов.

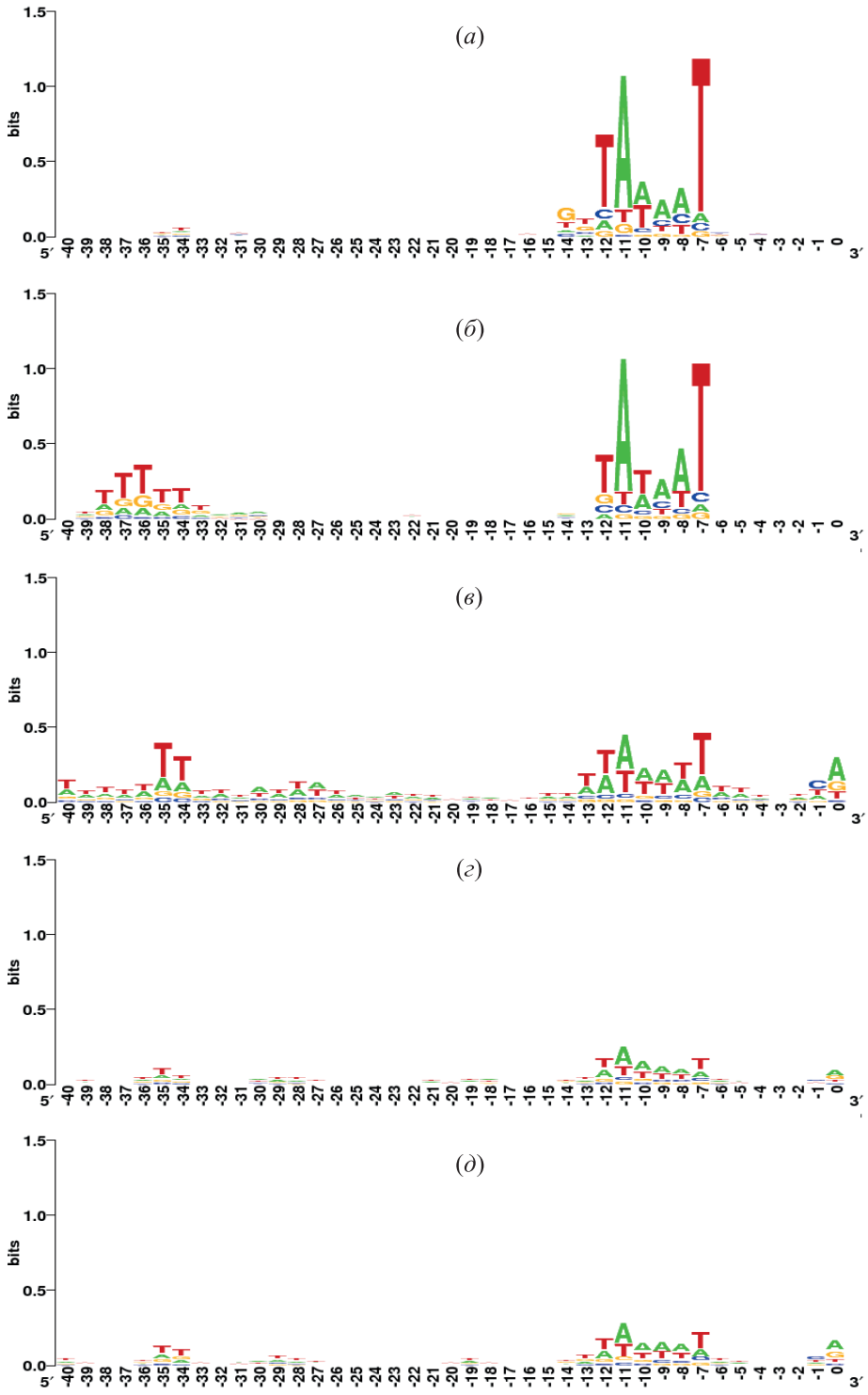

Рис. 3. Закономерности нуклеотидного состава положительных предсказаний алгоритмов: (*а*) BPROM, (*б*) NNPP, (*в*) PlatProm, (*г*) PLS–DA алгоритма из данной работы и (*д*) реальных промоторов. По горизонтальной оси отложено расстояние до точки старта транскрипции, по вертикальной — информационное содержание. Для каждого нуклеотида последовательности показано общее информационное содержание и вклад нуклеотидов конкретного типа

Это свидетельствует о том, что комбинация метода PLS–DA с каким-либо из текстовых методов может дать лучшие результаты за счет взаимной компенсации недостатков.

Выводы

Нами предложен алгоритм предсказания промотров, работающий без привлечения текстового анализа нуклеотидной последовательности, использующий только профиль распределения электростатического потенциала в промоторных областях. Показано, что, основываясь только на данных о распределении электростатического потенциала вдоль последовательностей ДНК, можно разделять различные классы последовательностей с точностью порядка 80–85 %. Эти результаты позволяют сделать вывод, что электростатический потенциал может быть использован в качестве одной из характеристик для доэкспериментального аннотирования генетических текстов.

Список литературы

Журкин В. Б., Полтев В. И., Флорентьев В. Л. Атом-атомные потенциальные функции для конформационных расчетов нуклеиновых кислот // Мол. биол. — 1980. — Сентябрь. — Т. 14, № 5. — С. 1116–1130.

Камзолова С. Г., Осипов А. А., Бескаравайный П. М. и др. Регуляция активности промоторной ДНК через электростатические взаимодействия с РНК-полимеразой // Биофизика. — 2007. — Т. 52, № 2. — С. 228–236.

Камзолова С. Г., Сорокин А. А., Осипов А. А., Бескаравайный П. М. Электростатическая карта генома бактериофага Т7. 1. Сравнительный анализ электростатических свойств σ70-специфических промоторов Т7 ДНК, взаимодействующих с РНК-полимеразой *E.coli* // Биофизика. — 2009. — Т. 54, № 6. — С. 975–983.

Киселев С. С., Озолинь О. Н. Структурообразующие модули как индикаторы промоторной ДНК в бактериальных геномах // Математическая биология и биоинформатика. — 2011. — Февраль. — С. 39–52.

Сорокин А. А. Сравнение одномерных профилей электростатического потенциала ДНК, полученных методом решения уравнения Пуассона–Больцмана, с результатами кулоновского подхода. — Неопубликованные данные.

Barker M. M., Gaal T., Gourse R. L. Mechanism of regulation of transcription initiation by ppGpp. II. Models for positive control based on properties of RNAP mutants and competition for RNAP // J. Mol. Biol. — 2001a. — January. — Vol. 305, no. 4. — P. 689–702.

Barker M. M., Gaal T., Josaitis C. A., Gourse R. L. Mechanism of regulation of transcription initiation by ppGpp. I. Effects of ppGpp on transcription initiation in vivo and in vitro // J. Mol. Biol. — 2001b. — January. — Vol. 305, no. 4. — P. 673–688.

Benham C. J. Duplex destabilization in superhelical DNA is predicted to occur at specific transcriptional regulatory regions // J. Mol. Biol. — 1996. — January. — Vol. 255, no. 3. — P. 425–434.

Benham C. J., Kohwi-Shigematsu T., Bode J. Stress-induced duplex DNA destabilization in scaffold/matrix attachment regions // J. Mol. Biol. — 1997. — November. — Vol. 274, no. 2. — P. 181–196.

Borukhov S., Nudler E. RNA polymerase: the vehicle of transcription // Trends Microbiol. — 2008. — February. — Vol. 16, no. 3. — P. 126–134.

BPROM. — Prediction of bacterial promoters. — http://linux1.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb.

Brok-Volchanski A., Masulis I., Shavkunov K. et al. Predicting sRNA genes in the genome of E. coli by the promoter-search algorithm PlatProm // Bioinformatics of Genome Regulation and Structure II. — 2006. — P. 11–20.

Crooks G. E. WebLogo: A Sequence Logo Generator // Genome Res. — 2004. — May. — Vol. 14, no. 6. — P. 1188–1190.

Drew H. R., Travers A. A. DNA structural variations in the E. coli tyrT promoter // Cell. — 1984. — June. — Vol. 37, no. 2. — P. 491–502.

Gama-Castro S., Salgado H., Peralta-Gil M. et al. RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units) // Nucleic Acids Res. — 2011. — January. — Vol. 39, no. Database issue. — P. D98–105.

Hirvonen C. A., Ross W., Wozniak C. E. et al. Contributions of UP elements and the transcription factor FIS to expression from the seven rrn P1 promoters in Escherichia coli // J. Bacteriol. — 2001. — November. — Vol. 183, no. 21. — P. 6305–6314.

Hook-Barnard I. G., Hinton D. M. The promoter spacer influences transcription initiation via sigma70 region 1.1 of Escherichia coli RNA polymerase // Proceedings of the National Academy of Sciences. — 2009. — January. — Vol. 106, no. 3. — P. 737–742.

Jayaram B., Sharp K. A., Honig B. The electrostatic potential of B-DNA // Biopolymers. — 1989. — May. — Vol. 28, no. 5. — P. 975–993.

Jensen L. J., Friis C., Ussery D. W. Three views of microbial genomes // Res Microbiol. — 1999. — Vol. 150, no. 9-10. — P. 773–777.

Jost D., Everaers R. Genome wide application of DNA melting analysis // J. Phys. Condens. Matter. — 2009. — January. — Vol. 21, no. 3. — P. 034108.

Kamzolova S. G., Sorokin A. A., Dzhelyadin T. D. et al. Electrostatic potentials of E.coli genome DNA // J. Biomol. Struct. Dyn. — 2005. — November. — Vol. 23, no. 3. — P. 341–345.

Kiryu H., Oshima T., Asai K. Extracting relations between promoter sequences and their strengths from microarray data // Bioinformatics. — 2005. — March. — Vol. 21, no. 7. — P. 1062–1068.

Koehl P. Electrostatics calculations: latest methodological advances // Current opinion in structural biology. — 2006. — Vol. 16, no. 2. — P. 142–151.

Le Cao K.-A., Boitard S., Besse P. Sparse PLS Discriminant Analysis: biologically relevant feature selection and graphical displays for multiclass problems // BMC Bioinformatics. — 2011. — June. — Vol. 12, no. 1. — P. 253.

Loziński T., Wierzchowski K. L. Effect of reversed orientation and length of An.Tn DNA bending sequences in the −35 and spacer domains of a consensus-like Escherichia coli promoter on its strength in vivo and gross structure of the open complex in vitro // Acta biochimica Polonica. — 1996. — Vol. 43, no. 1. — P. 265–279.

Mendoza-Vargas A., Olvera L., Olvera M. et al. Genome-Wide Identification of Transcription Start Sites, Promoters and Transcription Factor Binding Sites in E. coli // PLoS ONE. — 2009. — October. — Vol. 4, no. 10. — P. e7526.

Mevik B.-H., Wehrens R., Liland K. H. — pls: Partial Least Squares and Principal Component regression, 2011. — R package version 2.3-0.

Misra V. K., Hecht J. L., Yang A. S., Honig B. Electrostatic Contributions to the Binding Free Energy of the [lambda] cI Repressor to DNA // Biophys J. — 1998. — Vol. 75, no. 5. — P. 2262–2273.

Panyukov V. V., Kiselev S. S., Shavkunov K. S. et al. Mixed promoter islands as genomic regions with specific structural and functional properties // Matematicheskaya Biologiya i Bioinformatika [Mathematical Biology and Bioinformatics]. — 2013. — Vol. 8, no. 2. — P. 432–448.

Polozov R. V., Dzhelyadin T. R., Sorokin A. A. et al. Electrostatic potentials of DNA. Comparative analysis of promoter and nonpromoter nucleotide sequences // J. Biomol. Struct. Dyn. — 1999. — May. — Vol. 16, no. 6. — P. 1135–1143.

Polozov R. V., Sivozhelezov V. S., Ivanov V. V., Melnikov Y. B. On a classification of E. coli promoters according to their electrostatic potentials // Physics of Particles and Nuclei Letters. — 2005. — July. — Vol. 2, no. 4. — P. 241–246.

Reese M. G. Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome // Comput. Chem. — 2001. — December. — Vol. 26, no. 1. — P. 51–56.

Rohs R., West S. M., Liu P., Honig B. Nuance in the double-helix and its role in protein–DNA recognition // Current opinion in structural biology. — 2009. — April. — Vol. 19, no. 2. — P. 171–177.

Schaumburg C. S. Mutational analysis of the Chlamydia trachomatis dnaK promoter defines the optimal −35 promoter element // Nucleic Acids Res. — 2003. — January. — Vol. 31, no. 2. — P. 551–555.

Shavkunov K. S., Masulis I. S., Tutukina M. N. et al. Gains and unexpected lessons from genome-scale promoter mapping // Nucleic Acids Res. — 2009. — August. — Vol. 37, no. 15. — P. 4919–4931.

Solovyev V., Salamov A. Automatic annotation of microbial genomes and metagenomic sequences // Metagenomics and its applications in agriculture, biomedicine and environmental studies. — 2011. — P. 61–78.

Sorokin A. A., Osypov A. A., Dzhelyadin T. R. et al. Electrostatic properties of promoter recognized by E. coli RNA polymerase Esigma70 // Journal of Bioinformatics and Computational Biology. — 2006. — April. — Vol. 4, no. 2. — P. 455–467.

Sorokin A. A., Dzhelyadin T. R., Temlyakova E. A. — RelDNA: DNA electrostatics in R. — http://reldna.github.io.

Stephanou A. S., Roberts G. A., Tock M. R. et al. A mutational analysis of DNA mimicry by ocr, the gene 0.3 antirestriction protein of bacteriophage T7 // Biochem Biophys Res Commun. — 2009. — January. — Vol. 378, no. 1. — P. 129–132.

Travers A. A. DNA conformation and protein binding // Annual review of biochemistry. — 1989.

Tsonis P. A., Dwivedi B. Molecular mimicry: structural camouflage of proteins and nucleic acids // Biochim Biophys Acta. — 2008. — February. — Vol. 1783, no. 2. — P. 177–187.

Tutukina M. N., Shavkunov K. S., Masulis I. S., Ozoline O. N. Intragenic promotor-like sites in the genome of Escherichia coli discovery and functional implication // Journal of Bioinformatics and Computational Biology. — 2007. — April. — Vol. 5, no. 2B. — P. 549–560.

Wang H., Benham C. J. Promoter prediction and annotation of microbial genomes based on DNA sequence and structural responses to superhelical stress // BMC Bioinformatics. — 2006. — Vol. 7. — P. 248.

Weindl J., Dawy Z., Hanus P. et al. Modeling promoter search by E. coli RNA polymerase: One-dimensional diffusion in a sequence-dependent energy landscape // J. Theor. Biol. — 2009. — August. — Vol. 259, no. 3. — P. 628–634.

Yeramian E. Genes and the physics of the DNA double-helix // Gene. — 2000. — September. — Vol. 255, no. 2. — P. 139–150.

Yuzenkova Y., Tadigotla V. R., Severinov K., Zenkin N. A new basal promoter element recognized by RNA polymerase core enzyme // The EMBO Journal. — 2011. — July. — Vol. 30, no. 18. — P. 3766–3775.