МОЛЕКУЛЯРНАЯ БИОЛОГИЯ, 2011, том 45, № 4, с. 724–737

БИОИНФОРМАТИКА

УДК 577.21

ИЗУЧЕНИЕ СВЯЗЫВАНИЯ ДНК ФАКТОРАМИ ТРАНСКРИПЦИИ СЕМЕЙСТВА LacI МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

© 2011 г. Г. Г. Федонин[1]*, А. Б. Рахманинова[2], Ю. Д. Коростелёв[2], О. Н. Лайкова[1], М. С. Гельфанд[1]

[1]*Институт проблем передачи информации им. А.А. Харкевича Российской академии наук, Москва, 127994*

[2]*Факультет биоинженерии и биоинформатики Московского государственного университета им. М.В. Ломоносова, Москва, 119991*

Поступила в редакцию 23.11.2010 г.
Принята к печати 28.01.2011 г.

С помощью алгоритмов машинного обучения и методов отбора признаков изучали 1372 фактора транскрипции семейства LacI и 4484 их сайта связывания с ДНК. Использовали наивный байесовский классификатор и логистическую регрессию для предсказания сайтов связывания по последовательностям факторов транскрипции, а также для классификации пар "фактор–сайт" как связывающихся и не связывающихся. Точность прогноза и классификации оценивали, используя скользящий контроль по десяти блокам. Лучший прогноз плотностей распределения нуклеотидов в выбранных позициях сайтов получается при использовании небольшого числа ключевых позиций белковой последовательности. Эти позиции стабильно отбираются прямым последовательным отбором признаков на основе взаимной информации пар позиций в белке и в сайте.

Ключевые слова: **факторы транскрипции, наивный байесовский классификатор, логистическая регрессия, взаимная информация, прокариоты, семейство LacI.**

MACHINE LEARNING STUDY OF DNA BINDING BY TRANSCRIPTION FACTORS FROM THE LacI FAMILY, by *G. G. Fedonin[1]*, A. B. Rakhmaninova[2], U. D. Korostelev[2], O. N. Laikova[1], M. S. Gelfand[1]*

([1]Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, 127994, Russia, *e-mail: gennady.fedonin@gmail.com; [2]Department of Bioengineering and Bioinformatics, Moscow State University, Moscow, 119991 Russia).

We studied 1372 LacI-family transcription factors and their 4484 DNA binding sites using machine learning algorithms and feature selection techniques. The Naive Bayes classifier and Logistic Regression were used to predict binding sites given transcription factor sequences and to classify factor-site pairs on binding and non-binding ones. Prediction accuracy was estimated using 10-fold cross-validation. Experiments showed that the best prediction of nucleotide densities at selected site positions is obtained using only a few key protein sequence positions. These positions are stably selected by the forward feature selection based on the mutual information of factor-site position pairs.

***Keywords*: transcription factors, Naive Bayes classifier, Logistic Regression, Mutual Information, prokaryotes, Lac1 family.**

Многие биологические процессы включают в себя специфическое взаимодействие между ДНК-связывающими белками и сайтами ДНК. Механизм распознавания, специфичного по структуре и последовательности, остается малоизученным, несмотря на некоторые успехи, достигнутые при экспериментальных исследованиях мутаций и при компьютерном анализе известных рентгеновских структур комплексов белок–ДНК. Одной из причин этого отставания может быть недостаток данных.

При анализе экспериментально определенных структур комплексов белок–ДНК можно уловить (выделить) ряд закономерностей: предпочтительно образуются пары аланин–тимин (за счет взаимодействия метильных групп), а также пары водородных связей между аргинином и гуанином и аспарагином и аденином [1]. Показано, что область контакта белок–ДНК обогащена полярными аминокислотами [2], пурины более избирательны в отношении аминокислот, чем пиримидины [2], а ароматические аминокислоты могут иметь различные предпочтения [3].

Однако исключений оказалось не меньше, чем правил, и никакого универсального кода не обнаружено [4]. Со структурной точки зрения взаимодействие зависит от числа фиксированных контактов,

* Эл. почта: gennady.fedonin@gmail.com

специфичных для каждого семейства [1, 5, 6]. С целью определить участки белка, взаимодействующие с ДНК в различных семействах, применяли методы распознавания образов [7–9], а также методы, основанные на анализе детерминант специфичности [10–12].

Исследовать код взаимодействия белок–ДНК в больших семействах ДНК-связывающих белков можно не только в опыте, но и при сравнительном геномном анализе регулирующих взаимодействий. Обширный источник таких данных – бактериальные факторы транскрипции, как, например, семейство LacI, рассмотренное здесь. Имея данные о сайтах связывания этих белков, можно исследовать корреляции между аминокислотными последовательностями и соответствующими сайтами ДНК, а затем использовать известные структуры для сравнения, чтобы подтвердить, действительно ли наблюдаемые позиции контактируют в комплексах белок–ДНК.

Предыдущие исследования [13] показали, что корреляции не ограничены парами позиций в выравниваниях белков и ДНК: во многих случаях предпочтение белком конкретного нуклеотида в конкретной позиции, по-видимому, зависит от того, имеются ли некие специфические остатки в нескольких других позициях белка. Это ставит задачу выбора оптимальной сложности модели. В статье мы пытаемся решить эту задачу, используя прогнозирующую силу алгоритмов распознавания образов как средство определения оптимального числа параметров модели.

При этом множественное выравнивание аминокислотных последовательностей (АП) регуляторов при сопоставлении каждой последовательности с набором сайтов связывания представляется в виде множества пар “АП-сайт”. Первая задача – предсказание сайта по АП. Один фактор транскрипции может связываться с разными сайтами ДНК. Поэтому корректно предсказание для всевозможных нуклеотидных последовательностей с длиной 20 вероятности того, что эти последовательности являются сайтами связывания данного фактора транскрипции. На практике оценить такое распределение по обучающей выборке невозможно без упрощающих предположений. В настоящей работе предполагается, что позиции в сайте условно независимы. Тогда задача сводится к предсказанию вероятностей появления нуклеотидов каждого из четырех типов в каждой позиции сайта, т.е. распределение нуклеотидов в каждой позиции сайта предсказывается отдельно.

Вторая задача – классификация пар “АП-сайт” на две группы. При рассмотрении каждой АП в качестве объектов первого класса (положительного) используются пары, образованные АП с ее сайтами связывания, в качестве объектов второго (отрицательного) – пары, образованные данной АП с сайтами других АП того же организма.

ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

Подготовка данных. В работе использовали выборку бактериальных транскрипционных регуляторов семейства LacI и их сайтов связывания как показанных экспериментально, так и предсказанных методами сравнительной геномики. Выборка доступна в базе данных RegPrecise [14].

С помощью инструментов базы данных SMART [15] в аминокислотной последовательности каждого регулятора из выборки определяли границы ДНК-связывающего домена (HTH_LACI), а затем полученную последовательность домена выравнивали против эталонного выравнивания семейства HTH_LACI [16]. Всего получено 1372 последовательности, три необычно короткие последовательности (43, 51, 61 а.о.) не рассматривались. Для уточнения границ петель применяли минимальное ручное редактирование.

Нуклеотидные последовательности сайтов связывания в выборке в подавляющем своем большинстве (94%) образуют (представляют собой) четные палиндромы длиной 20 п.н. с консервативной парой CG в центре. Остальные – нечетные палиндромы длиной 21 п.н. В данной работе рассмотрены только первые из них. Конечная выборка содержит 4484 нуклеотидные последовательности длиной в 20 п.н. со средним попарным сходством 45%.

Выравнивание аминокислотных последовательностей содержит 87 позиций. Среди них в шестнадцати позициях более 30% содержат пробелы и потому эти позиции не использовали при обучении алгоритмов классификации. Оставшиеся позиции пронумерованы подряд, начиная с 1.

Разбиение на обучающую и тестовую выборку. Для оценки качества алгоритмов предсказания и классификации исходную выборку случайным образом разбивали на десять частей, каждую из которых поочередно использовали для тестирования алгоритма, обученного на оставшихся девяти частях. Так как многие белки в выборке близкородственны (имеют очень сходные аминокислотные последовательности), разумно требовать, чтобы в тестовой выборке не было АП, сильно похожей на какую-нибудь АП обучающей выборки. Чтобы это обеспечить, все АП объединяли на основе попарной близости в кластеры, которые никогда не разделялись при разбиении в выборки. Для этого вычисляли близость (процент совпадающих аминокислотных остатков во всех позициях) всех пар. Далее строили полный граф, вершинами которого являются АП, а ребрам приписаны веса, равные степени близости пары АП. Далее удаляли ребра с весом меньше порогового значения. Максимальные связные компоненты считали кластерами близких АП, все алго-

ритмы обучали на обучающей выборке и вычисляли функционал ошибки на тестовой выборке. Результат усредняли. Процедуру повторяли десять раз для лучшего качества усреднения.

Оценка качества алгоритмов. Качество алгоритмов предсказания оценивали по логарифму правдоподобия нуклеотидов в исследуемой позиции в тестовой выборке. Для всех сайтов каждого фактора транскрипции вычисляли логарифм условной вероятности появления наблюдаемого нуклеотида в данной позиции. Эти значения для всех факторов с весами, пропорциональными филогенетическим весам, суммировали:

$$\lg L = \frac{\sum_i w_i \sum_j \lg P(n_{ij}|AAS_i)}{\sum_i w_i},$$

где AAS_i – i-ая АП; n_{ij} – нуклеотид, наблюдаемый в j-ом сайте связывания белка, имеющего i-ую АП; w_i – вес i-ой АП.

Полученные таким образом для каждого разбиения значения правдоподобия усредняли. Деление на суммарный вес всех АП уравнивает вклад каждого тестового блока в общее значение правдоподобия: блоки могут содержать разное число АП и иметь различный вес.

Качество алгоритмов классификации оценивали как усредненную по всем разбиениям долю ошибочных ответов на тестовой выборке. При усреднении также использовали веса.

Использованные алгоритмы. Взвешивание аминокислотных последовательностей и сайтов связывания. Кластеры близких АП имеют разные размеры: некоторые группы сходных последовательностей слишком многочисленны, что приводит к заметному искажению статистических характеристик выборки. Для уменьшения этого эффекта последовательности взвешивали при помощи алгоритма Герштейна–Сонхаммера–Чотьи (Gerstein, Sonnhammer, Chothia) [17]: белки, имеющие много близких родственников в выборке, получали меньший вес, чем белки, непохожие на остальные. Для получения весов пар, вес каждой АП делили поровну на все сайты, соответствующие данной АП.

Отрицательные пары взвешивали тем же методом. В результате, суммарные веса отрицательных и положительных пар у каждого белка были равны.

Полученные таким образом веса использовали для вычисления частот аминокислот и нуклеотидов при построении байесовского классификатора, при вычислении взаимной информации, при обучении алгоритма логистической регрессии, а также для оценки качества алгоритмов.

Наивный байесовский классификатор. Байесовский классификатор [18] оценивает вероятности наблюдения в заданной позиции сайта каждого из нуклеотидов по формуле Байеса.

$$P(n_i|AAS) = \frac{P(n_i)P(AAS|n_i)}{\sum_j P(n_j)P(AAS|n_j)},$$

где n_i – i-й нуклеотид; AAS – аминокислотная последовательность; $P(n)$ – априорная вероятность появления нуклеотида n.

Наивный байесовский классификатор предполагает, что позиции в АП условно независимы в совокупности, т.е.

$$P(AAS|n) = \prod_i P(a_i|n),$$

где a_i – аминокислотный остаток в позиции i.

Вероятности $P(a_i|n)$ оценивали по соответствующим частотам выборки. При этом использовали филогенетические веса АП и технику псевдоотсчетов. Априорные вероятности оценивали по формуле:

$$P(n) = \frac{W(n) + k\frac{\sum_m W(m)P(m \to n)}{\sqrt{W}}}{W + k\sqrt{W}} = \frac{W(n) + 0.25k\frac{\sum_m W(m)}{\sqrt{W}}}{W + k\sqrt{W}},$$

где $W(n) = \sum_{i=1}^{N} w_i f(n)$ – сумма весов АП, умноженных на частоты нуклеотида n; $W = \sum_{i=1}^{N} w_i$ – суммарный вес всех АП; k – коэффициент, регулирующий вклад псевдоотсчетов [15].

Условная вероятность появления а.о. a_i в позиции j:

$$P(a_i|n) = \frac{W_j(a_i|n) + k\sqrt{W(n)}p}{W(n) + k\sqrt{W(n)}},$$

где a_i – тип аминокислотного остатка; n – тип нуклеотида; $W_j(a_i|n)$ – сумма весов АП, содержащих a_i в позиции j, умноженных на частоты нуклеотида n в выбранной позиции сайтов этих АП; $p = 1/m$; m – число возможных аминокислотных остатков.

Эксперименты показали, что прямое применение такого подхода дает плохие результаты. Это связано с консервативностью большинства позиций. Вероятности редко встречающихся остатков оказываются статистически не значимыми. Чтобы избежать этого эффекта, остатки, встречающиеся в вы-

борке в одной позиции реже некоторого порогового значения, объединяли в одну группу и считали одним типом остатка. При этом $p = 1/m_j$, где m_j – количество групп остатков в позиции j после группировки.

В задаче классификации пар "АП-сайт", с учетом равновероятности классов, формула Байеса принимает вид:

$$P(c|AAS, NS) = \frac{\frac{1}{2}P(AAS, NS|c)}{\sum_j \frac{1}{2}P(AAS, NS|c_j)},$$

где AAS – аминокислотная последовательность; NS – последовательность нуклеотидов сайта; c – класс.

Далее предполагаем, что пары позиций в АП и в сайте условно независимы в совокупности:

$$P(AAS, NS|c) = \prod_{i,j} P(a_i, n_j|c),$$

где a_i – аминокислотный остаток в позиции i; n_j – нуклеотид в позиции j.

Условные вероятности пар позиций вычисляли с использованием псевдоотсчетов и группировки редких а.о.:

$$P(a_i, n_j|c) = \frac{W_{ij}(a_i, n_j|c) + k\sqrt{W}p}{W + k\sqrt{W}},$$

где a_i – тип аминокислотного остатка; n_j – тип нуклеотида; $W_{ij}(a_i, n_j|c)$ – сумма весов пар класса c, содержащих а.о. a_i в позиции i АП и нуклеотид n_j в позиции j сайта; $p = 1/m$; m – число возможных аминокислотных остатков; k – коэффициент, регулирующий вклад псевдоотсчетов.

Логистическая регрессия. Логистическая регрессия [19] – популярный метод машинного обучения для решения задач классификации на два класса. Предполагается, что обучающие объекты задаются K-мерными векторами числовых признаков $\bar{x} = (x_1, ..., x_K)$ с бинарной меткой класса $y \in \{-1, 1\}$. Алгоритм строит линейное решающее правило. Каждому числовому признаку присваивается вес. При классификации вычисляется линейная комбинация признаков классифицируемого объекта, и ответ определяется знаком этой линейной комбинации.

Решающее правило имеет вид:

$$f(x_1, ..., x_K) = \mathrm{sing}\left(\sum_{i=1}^{K} \alpha_i x_i\right),$$

или в векторной форме:

$$f(\bar{x}) = \mathrm{sign}(\langle \bar{\alpha}, \bar{x} \rangle),$$

где α_i – вес i-го признака; x_i – значение i-го признака.

Обучение классификатора состоит в подборе весов путем максимизации функционала качества на обучающей выборке. Функционал качества имеет вид:

$$L(\bar{\alpha}) = \sum_{i=1}^{l} w_i \ln(\sigma(y_i \langle \bar{\alpha}_i, \bar{x}_i \rangle)) \to \max_{\bar{\alpha}},$$

где индекс i пробегает объекты обучающей выборки; $y_i \in \{-1, 1\}$ – класс i-го обучающего объекта; $\sigma(z) = 1/(1 + \exp(-z))$ – логистическая (сигмоидная) функция; w_i – веса объектов.

Максимизация данного функционала эквивалентна максимизации правдоподобия обучающей выборки в предположении, что функции правдоподобия классов принадлежат экспоненциальному классу распределения с равными значениями параметра разброса. При таких предположениях и наличии среди признаков константы, байесовское решающее правило линейно.

Для повышения качества классификатора использовали метод регуляризации. Для этого в функционал добавляется слагаемое, штрафующее большие по модулю значения параметров:

$$L'(\bar{\alpha}) = L(\bar{\alpha}) - \beta \sum_{i=1}^{K} \alpha_i^2,$$

где i пробегает все номера признаков; β – коэффициент регуляризации.

Вероятности наблюдения каждого из классов у объекта с данным вектором признаков вычисляются по формуле:

$$P(y) = \frac{1}{1 + \exp(-y\langle \bar{\alpha}, \bar{x} \rangle)},$$

где $y \in \{-1, 1\}$ – класс; $\bar{x}$ – вектор признаков; $\bar{\alpha}$ – вектор весов.

Логистическая регрессия предполагает использование числовых признаков. В нашем случае признаки номинальные. Использовали стандартный подход бинаризации таких признаков. Для задачи прогнозирования – с каждым i-м остатком сопоставляли его признак-индикатор: $f_i(a) = 1$, когда $a = a_k$, и $f_i(a) = 0$ в остальных случаях. Для задачи классификации признак-индикатор сопоставляли с каждой парой "а.о. – нуклеотид": $f_i(a, n) = 1$, когда $a = a_i$ и $n = n_j$, и $f_i(a, n) = 0$ в остальных случаях.

Сведение прогнозирования к бинарной задаче распознавания. Для предсказания распределения нуклеотидов в выбранной позиции сайта для каждого нуклеотида строили отдельный классификатор. В качестве положительных примеров использовали пары белок–сайт с данным нуклеотидом в данной позиции сайта, а в качестве отрицательных – все

остальные пары. При прогнозировании вероятность каждого типа нуклеотида определяли по формуле:

$$P(n_i|AAS) = \frac{P_i(+|AAS)}{\sum_{j=1}^{4} P_j(+|AAS)},$$

где AAS – АП, для которой делается прогноз; $P_i(+|AAS)$ – вероятность положительного класса, вычисленная i-м классификатором.

При разбиении на отрицательный и положительный класс белков с несколькими сайтами, различающимися в исследуемой позиции, возникает проблема. Такой белок для нескольких классификаторов является и положительным, и отрицательным примером. Наличие в выборке идентичных обучающих примеров с разными метками классов сильно ухудшают качество прогнозирования. В такой ситуации можно либо игнорировать такие конфликтные белки, составляющие существенную часть выборки, либо включать их только в положительную выборку. Эксперименты показали преимущество второго подхода и его применяли во всех экспериментах.

В этой части работы использовали следующие веса: для отрицательных объектов – вес соответствующей АП, для положительных – тот же вес, умноженный на частоту данного нуклеотида в выбранной позиции сайтов данной АП.

Метод *k* ближайших соседей. Метод k ближайших соседей, применяемый обычно для решения задач классификации, основан на использовании меры близости на множестве объектов. Обучение алгоритма состоит в запоминании обучающей выборки. В процессе классификации в простейшем случае, при $k = 1$, новому объекту приписывается класс ближайшего объекта обучающей выборки. Часто можно повысить качество классификации, определяя класс нового объекта голосованием k ближайших к нему объектов обучающей выборки. Простое голосование можно заменить взвешиванием ответов объектов пропорционально их близости к классифицируемому объекту.

В настоящей работе в качестве меры близости для прогнозирования распределения нуклеотидов в исследуемой позиции сайта использовали долю совпавших а.о. в отобранных позициях АП. В качестве прогноза вероятности появления нуклеотида использовали эффективную частоту появления этого нуклеотида в сайтах k ближайших соседей:

$$P(n_i|AAS) = \frac{\sum_{j=1}^{k} s_j w_j f_j(n_i)}{\sum_{j=1}^{k} s_j w_j},$$

где n_i – i-й нуклеотид; AAS – аминокислотная последовательность; s_j – близость j-ой ближайшей АП к АП, для которой строится прогноз; w_j – вес j-ой АП; $f_j(n_i)$ – частота i-го нуклеотида в исследуемой позиции сайтов j-ой АП.

Оценка частот при $1 \le k \le 10$ статистически не значима. Сравнительно неплохое качество прогноза может быть достигнуто при $k \ge 50$. Использовали значение $k = 200$, дающее лучшее значение на тестовой выборке.

В задаче классификации пар “АП-сайт” в качестве меры близости пар использовали произведение значений близости АП и близости сайтов, вычисляемой как доля совпавших нуклеотидов в отобранных позициях сайтов. Опыты показали, что лучшее качество достигается при использовании только одного ближайшего соседа.

Отбор признаков по взаимной информации. Взаимная информация (Mutual information, MI [20]) пары позиций в АП и в сайте – мера скоррелированности этих позиций, позволяющая быстро оценить предсказательную силу данной позиции в АП для предсказания нуклеотида в заданной позиции в сайте. Достоинство MI – быстрота ее вычисления, что позволяет использовать ее для отбора признаков. Этот метод позволяет избавиться от неинформативных признаков, но не учитывает зависимость между признаками.

Для регуляризации статистически ненадежных оценок частот редких а.о. и нуклеотидов (в заданной позиции) использовали псевдоотсчеты, добавляющие небольшие значения редким событиям.

Эффективная частота появления остатка a в позиции i вычисляли по формуле:

$$f_i(a) = \frac{W_i^A(a) + k\sum_{b=1}^{20} W_i^A(b) P(b \to a)/\sqrt{W^A}}{W^A + k\sqrt{W^A}},$$

где $W_i^A(a)$ – суммарный вес АП с остатком a в позиции i; W^A – суммарный вес всех АП в выборке. Вероятности переходов $P(b \to a)$ брали из BLOSUM60 [21] .

Эффективная частота появления нуклеотида n в позиции j:

$$f_j(n) = \frac{W_j^S(n) + k\sum_{m=1}^{4} W_j^S(m) P(m \to n)/\sqrt{W^S}}{W^S + k\sqrt{W^S}} =$$

$$= \frac{W_j^S(n) + 0.25k\sum_{m=1}^{4} W_j^S(m)/\sqrt{W^S}}{W^S + k\sqrt{W^S}},$$

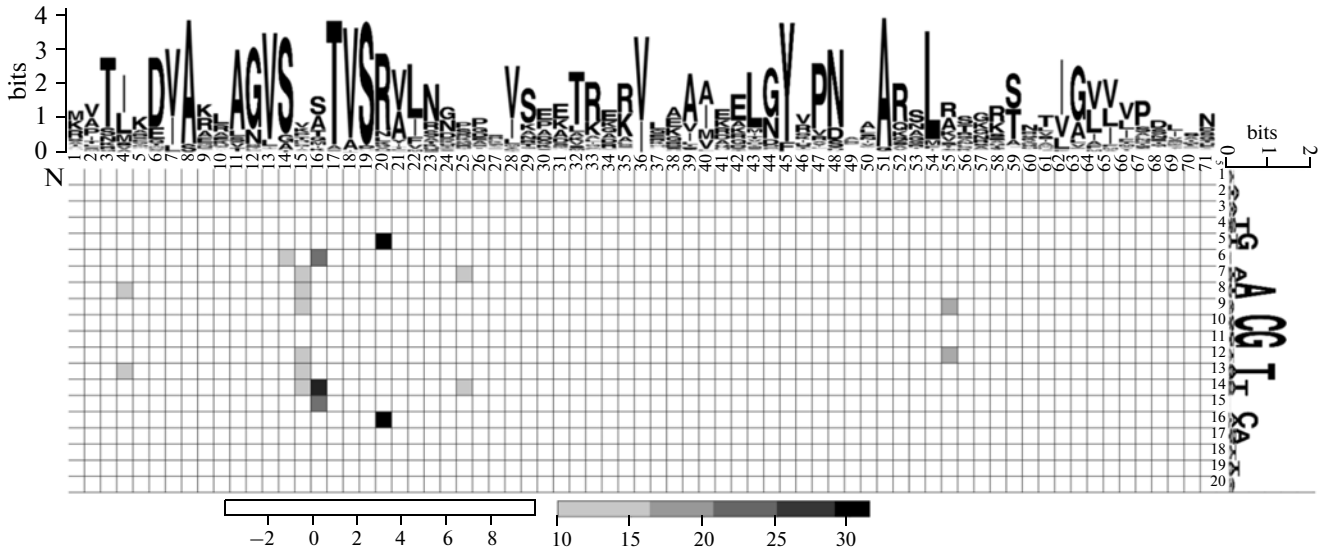

Рис. 1. Взаимная информация пар позиций “АП–сайт” [13]. Более темные цвета соответствуют более значимым корреляциям.

где W_j^S – суммарный вес сайтов с нуклеотидом n в позиции j; W^S – суммарный вес сайтов в выборке.

Наблюдаемая эффективная частота пары позиций (одна в белке, одна в сайте):

$$f_{ij}^o(a, n) = \frac{W_{ij}^P(a, n) + k\sqrt{W^P} f_{ij}^e(a, n)}{W^P + k\sqrt{W^P}},$$

где $W_{ij}^P(a, n)$ – суммарный вес пар “АП–сайт”, имеющих остаток а в позиции i АП и нуклеотид n в позиции j сайта; W^P – суммарный вес всех пар АП в выборке; $f_{ij}^e(a, n)$ – ожидаемая эффективная частота пары (a, n):

$$f_{ij}^e(a, n) = f_i(a) f_j(n),$$

где $f_i(a)$ и $f_j(n)$ – эффективные частоты остатка а в позиции i и нуклеотида n в позиции j соответственно.

Взаимную информацию вычисляли по формуле:

$$I_{ij} = \sum_{a.} \sum_{n} f_{ij}^o(a, n) \lg \frac{f_{ij}^o(a, n)}{f_{ij}^e(a, n)}.$$

“**Жадный**” **отбор признаков.** Другой метод отбора – перебор подмножеств признаков с обучением алгоритма на части обучающей выборки и оценкой ошибки на оставшейся части. Выбирается множество, дающее наименьшую ошибку.

На практике перебрать все подмножества нельзя, поэтому здесь использовали так называемый “жадный” алгоритм, добавляющий последовательно к текущему наилучшему множеству признаков каждый из оставшихся и выбирающий тот, добавление которого дает наилучший классификатор. Этот признак добавляли к наилучшему множеству, и процесс повторялся.

Жадная стратегия учитывает связь между признаками, однако может давать неоптимальные множества. Тем не менее, этот алгоритм самый быстрый метод, после отбора по MI.

Жадный алгоритм может быть усовершенствован несколькими способами, в частности, путем последовательного добавления и удаления признаков, либо путем сохранения нескольких лидеров. Опыты с использованием байесовского классификатора (обучающегося быстрее всего) показали, что усовершенствованные алгоритмы отбирают те же множества признаков, что и простой жадный алгоритм.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Выбор позиций для прогнозирования

Разные позиции сайтов могут быть предсказаны с разной точностью. В этой работе исследовали позиции сайтов, для которых были найдены значимо скоррелированные с ними позиции в АП [13]. В качестве меры скоррелированности использовали взаимную информацию. Как видно из тепловой карты (рис. 1), значимые корреляции наблюдаются у позиций сайта 5, 6, 7, 8, 9 и симметричных им. Далее будут рассмотрены эти пять позиций.

Отбор значимых позиций

Использовали два способа. По взаимной информации для каждой из пяти позиций сайта было отобрано двадцать наиболее информативных позиций, т.е. имеющих в паре с исследуемой позицией сайта

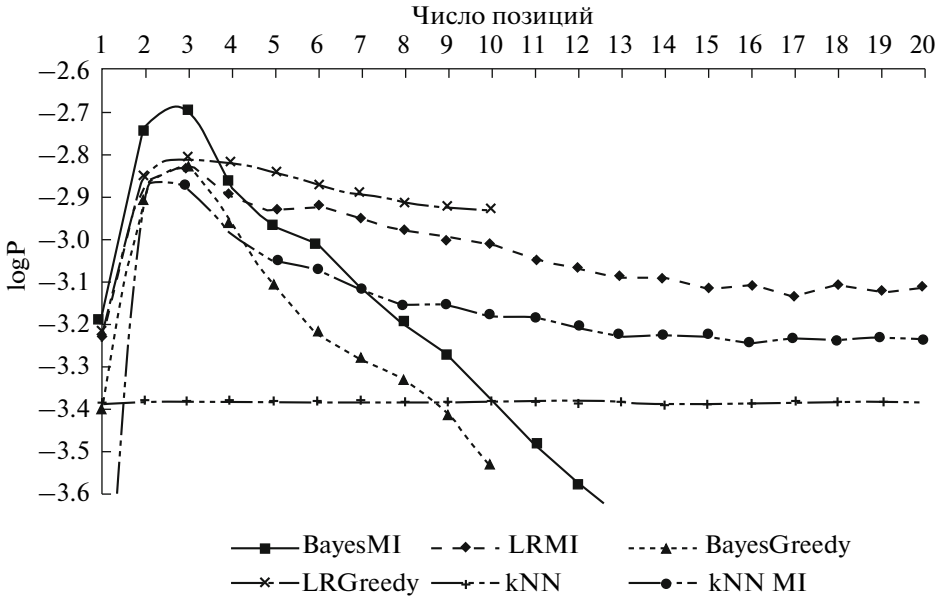

Рис. 2. Зависимость точности прогноза от числа отобранных позиций для позиции 9 выравнивания сайтов. Представлены результаты использования байесовского классификатора (BayesMI) и логистической регрессии (LRMI) с отбором признаков по взаимной информации (MI) и по отбору жадным алгоритмом (BayesGreedy и LRGreedy соответственно), а также метода *k* ближайших соседей с отбором признаков по MI (kNN MI) и метода *k* ближайших соседей без отбора признаков (kNN).

наибольшие значения MI. Позиции отбирали последовательно, начиная с самой информативной. На каждой итерации по этому множеству обучали классификаторы (байесовский, логистическая регрессия и метод *k* ближайших соседей с $k = 200$) и оценивали качество прогноза. Жадный отбор был организован тем же способом, но только с использованием десяти позиций АП для каждой позиции сайта. В обоих случаях процесс повторяли для разных разбиений выборки при скользящем контроле и результаты усредняли.

По результатам тестов построили графики зависимости между точностью прогноза и числом отобранных признаков, а также таблицы, показывающие, какие позиции использованные алгоритмы отбирали на каждой итерации отбора. При разных разбиениях выборки могут быть отобраны различные наборы позиций. Поэтому мы приводим только частоту появления данной позиции во множестве отобранных на некотором шаге алгоритма позиций, т.е. частоту появления позиции в отобранных наборах длины от 1 до 20. Для наглядности все позиции упорядочены по суммарной частоте (сумме частот в наборах всевозможных длин) и приводятся только наиболее частые из них. Отбор позиций по MI не зависит от применяемого метода прогнозирования: оба использованных классификатора обучались с использованием одних и тех же наборов позиций. Жадный отбор предполагает использование классификатора в процессе отбора: каждому классификатору при этом соответствуют свои наборы признаков.

Отбор позиций в АП для позиции 9 выравнивания сайтов

Значения точности, полученные на тестовой выборке для позиции 9 различными алгоритмами и стратегиями отбора позиций, представлены на рис. 2. Максимум достигается на трех позициях при использовании всех методов, причем качество прогноза существенно выше, чем у алгоритма *k* ближайших соседей, использующего все позиции (kNN).

В табл. 1 указаны наиболее часто выбираемые позиции. Отбор по MI и жадный отбор наивным байесовским алгоритмом стабильно отбирают три позиции – 55, 15 и 5. Жадная логистическая регрессия стабильно отбирает те же три позиции, а также часто выбирает позицию 27.

Максимальное качество прогноза достигается при использовании трех позиций. Следовательно, позиции 55, 15 и 5 аминокислотного выравнивания значимо связаны с позицией 9 выравнивания сайтов.

Отбор позиций в АП для позиции 8 выравнивания сайтов

Значения точности, полученные на тестовой выборке для позиции 8 различными алгоритмами и стратегиями отбора позиций, представлены на рис. 3. График точности байесовского классификатора с отбором по MI имеет четкие максимумы при использовании двух и пяти позиций. Так же, но с

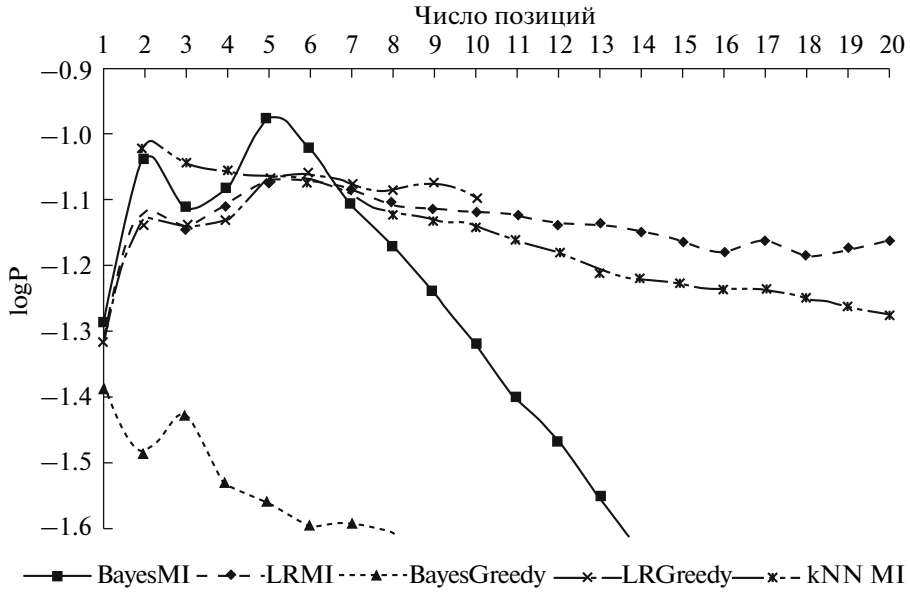

Рис. 3. Зависимость точности прогноза от числа отобранных позиций для позиции 8 выравнивания сайтов.

менее выраженными максимумами, выглядят графики логистической регрессии. Жадный байесовский алгоритм имеет два максимума при использовании одной и трех позиций. Метод *k* ближайших соседей с отбором по MI – только один максимум при двух позициях. При использовании всех позиций алгоритм ближайшего соседа дает точность меньше –4 (на рисунке не отображен), т.е. значительно уступает всем алгоритмам.

В табл. 2 указаны наиболее часто отбираемые позиции. Отбор по взаимной информации и жадная логистическая регрессия абсолютно стабильно отбирают две позиции – 4 и 15. Дальнейший отбор обоими методами менее стабильно отбирает одни и те же позиции в разном порядке. Жадный отбор байесовским алгоритмом стабильно отбирает позицию 4, но на втором шаге отбирает позицию 5, что приводит к снижению качества прогноза.

Несмотря на наличие нескольких максимумов на графиках точности, стабильно отбираются только две позиции. Следовательно, позиции 4 и 15 выравнивания аминокислотных последовательностей значимо связаны с позицией 8 выравнивания сайтов.

Отбор позиций в АП для позиции 7 выравнивания сайтов

Значения точности, полученные на тестовой выборке для позиции 7 различными алгоритмами и стратегиями отбора позиций, представлены на рис. 4. Хорошо различимый максимум достигается на трех

Таблица 1. Частоты шести наиболее частых позиций в наборах различной длины, отобранных по методам MI, жадного наивного байесовского классификатора (NB) и жадной логистической регрессии (LR) для предсказания позиции 9 выравнивания сайтов (в %)

	MI						Bayes						LR					
	55	15	5	68	56	16	55	15	5	1	70	26	55	15	5	27	49	56
1	0.98	0.02	0	0	0	0	0.91	0.09	0	0	0	0	0.96	0.04	0	0	0	0
2	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0
3	1	1	0.9	0	0	0	1	1	0.96	0	0	0	1	1	0.9	0.09	0	0
4	1	1	0.9	0.2	0.35	0.39	1	1	0.99	0.36	0.05	0.06	1	1	0.96	0.82	0.05	0.04
5	1	1	0.95	0.5	0.64	0.57	1	1	0.99	0.52	0.23	0.23	1	1	0.98	0.94	0.38	0.37
6	1	1	0.97	0.79	0.8	0.8	1	1	0.99	0.68	0.42	0.4	1	1	0.99	0.96	0.64	0.54

Примечание. Номера столбцов соответствуют номерам позиций, начиная с самой частой. Номера столбцов – число позиций в отобранном наборе.

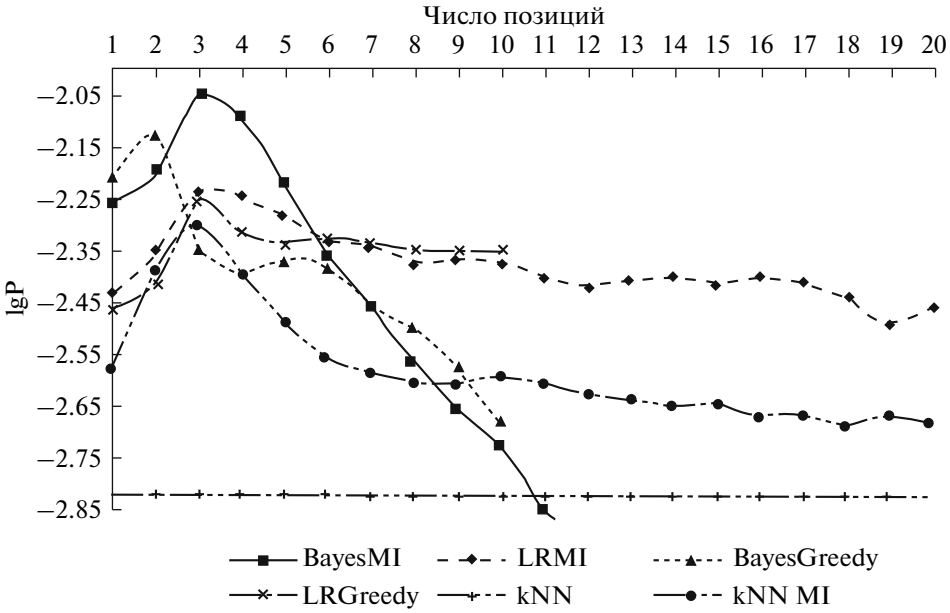

Рис. 4. Зависимость точности прогноза от числа отобранных позиций для позиции 7 выравнивания сайтов.

позициях всеми методами, кроме жадного наивного байесовского классификатора, имеющего максимум при двух позициях. Качество прогноза у всех алгоритмов существенно выше, чем у алгоритма *k* ближайших соседей, использующего все позиции (kNN).

Самые часто отбираемые позиции представлены в табл. 3, с принятыми выше обозначениями. Отбор

Таблица 2. Частоты пяти наиболее частых позиций в наборах различной длины, отобранных по методам MI, жадного наивного байесовского классификатора (NB) и жадной логистической регрессии (LR) для предсказания позиции 8 выравнивания сайтов (в %)

	MI					Bayes					LR				
	4	15	5	27	25	4	5	61	15	19	4	15	19	27	5
1	0.89	0.11	0	0	0	0.92	0	0	0.07	0.01	0.89	0.11	0	0	0
2	0.99	1	0	0	0	0.93	0.67	0.05	0.18	0.1	0.98	1	0.02	0	0
3	0.99	1	0.62	0.31	0.05	0.93	0.7	0.65	0.26	0.26	1	1	0.64	0.16	0.17
4	0.99	1	0.85	0.94	0.11	0.93	0.7	0.65	0.52	0.3	1	1	0.83	0.56	0.46
5	1	1	0.99	0.97	0.46	0.96	0.72	0.65	0.53	0.37	1	1	0.91	0.9	0.89

Таблица 3. Частоты шести наиболее частых позиций в наборах различной длины, отобранных по методам MI, жадного наивного байесовского классификатора (NB) и жадной логистической регрессии (LR) для предсказания позиции 7 выравнивания сайтов (в %)

	MI						Bayes						LR					
	16	25	15	68	5	46	16	15	49	68	50	19	16	15	25	49	68	50
1	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
2	1	1	0.04	0	0	0	1	0.97	0	0.02	0	0	1	0.69	0.3	0	0	0
3	1	1	1	0	0	0	1	0.97	0.71	0.11	0.09	0.02	1	0.99	0.99	0	0	0
4	1	1	1	0.84	0.05	0.03	1	0.98	0.89	0.59	0.33	0.07	1	1	1	0.56	0.2	0.05
5	1	1	1	0.94	0.25	0.18	1	0.98	0.92	0.94	0.75	0.12	1	1	1	0.78	0.57	0.16
6	1	1	1	0.97	0.38	0.46	1	0.99	0.93	1	0.86	0.64	1	1	1	0.89	0.84	0.42

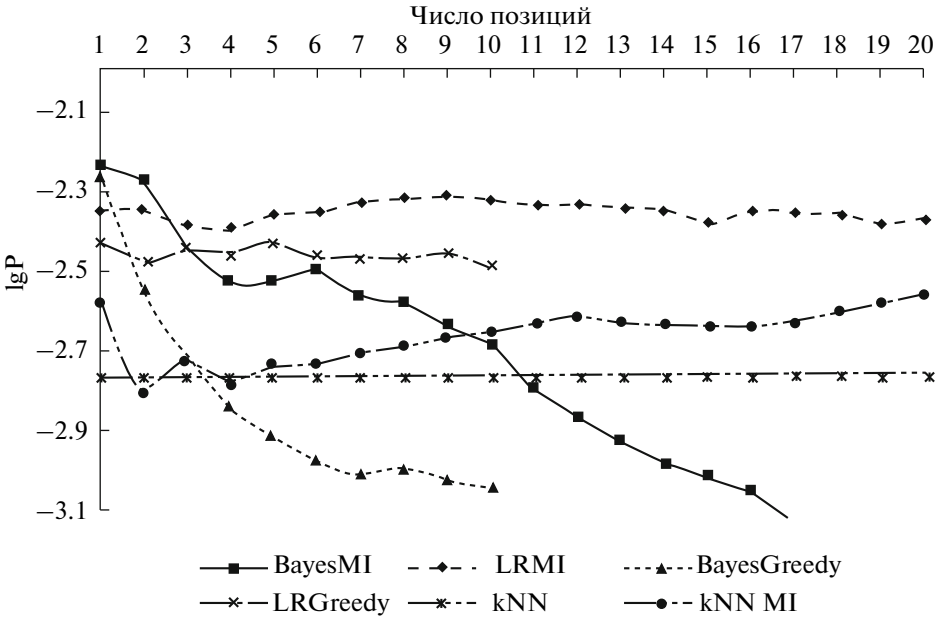

Рис. 5. Зависимость точности прогноза от числа отобранных позиций для позиции 6 выравнивания сайтов.

на основе взаимной информации стабильно отбирает три позиции – 16, 25 и 15, и иногда позицию 68. Жадная логистическая регрессия стабильно отбирает те же позиции, тогда как жадный отбор на основе байесовского классификатора ошибается на третьем шаге, стабильно выбирая позицию 49, что, как видно на графике точности прогноза, приводит к существенному снижению качества прогноза.

Максимальная точность прогноза достигается при использовании трех позиций. Следовательно, позиции 16, 25 и 15 выравнивания аминокислотных последовательностей значимо связаны с позицией 7 выравнивания сайтов.

Отбор позиций в АП для позиции 6 выравнивания сайтов

Зависимость точности от длины набора позиций показана на рис. 5. Наивный байесовский классификатор с отбором на основе взаимной информации имеет два максимума на одной и трех позициях, тогда как жадная стратегия имеет максимум на одной и семи позициях. Кривая логистической регрессии медленно растет, имея много локальных максимумов с максимальными значениями около шести и одиннадцати позиций для жадного и основанного на MI отборов соответственно. Алгоритм *k* ближайших соседей с отбором по MI имеет четкий максимум на трех позициях.

В табл. 4 указаны наиболее часто отбираемые позиции. Отбор по взаимной информации имеет одну абсолютно стабильную позицию – 16, а также две дополнительные стабильные позиции – 25 и 15, которые взаимозаменяемы на втором шаге отбора. Жадные стратегии отбирают две позиции: абсолютно стабильно 16 и достаточно стабильно 15. Дальнейший отбор нестабилен.

В случае предсказания позиции 6 выравнивания сайтов связывания разные алгоритмы ведут себя по-разному: график качества предсказания при ис-

Таблица 4. Частоты пяти наиболее частых позиций в наборах различной длины, отобранных по методам MI, жадного наивного байесовского классификатора (NB) и жадной логистической регрессии (LR) для предсказания позиции 6 выравнивания сайтов (в %)

	MI					Bayes					LR				
	16	25	15	68	26	16	15	20	27	49	16	15	27	25	49
1	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0
2	1	0.6	0.4	0	0	1	0.9	0	0.1	0	1	0.85	0.05	0.08	0
3	1	0.96	0.91	0	0.08	1	0.94	0.61	0.28	0.06	1	0.93	0.65	0.19	0.04
4	1	0.98	0.95	0.45	0.29	1	0.94	0.82	0.64	0.21	1	0.95	0.78	0.35	0.22
5	1	1	0.97	0.66	0.58	1	0.97	0.89	0.82	0.68	1	0.98	0.86	0.55	0.47

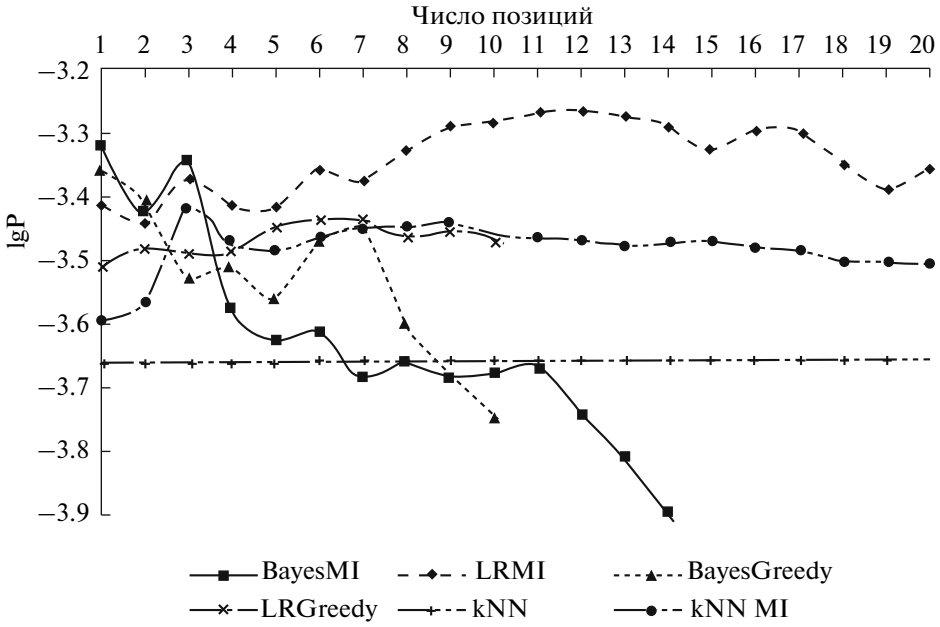

Рис. 6. Зависимость точности прогноза от числа отобранных позиций для позиции 5 выравнивания сайтов.

пользовании наивного байесовского классификатора имеет два максимума, а логистическая регрессия имеет размытый максимум в районе 12-ой позиций и, тем самым, видимо, переобучается. Тем не менее, все методы стабильно отбирают позицию 16 выравнивания аминокислотных последовательностей, которая существенно связана с позицией 6.

Отбор позиций в АП для позиции 5 выравнивания сайтов

Точность прогноза разными алгоритмами и стратегиями отбора, оцененная на тестовой выборке для позиции 5, показана на рис. 6. При использовании метода *k* ближайших соседей и наивного байесовского классификатора, а также при отборе по MI и жадным алгоритмом максимум достигается при использовании только одной позиции. У графиков логистической регрессии нет выраженного максимума.

Наиболее часто отбираемые позиции показаны в табл. 5. Позиция 20 абсолютно стабильна, позиция 25 стабильна при отборе по MI. Дальнейший отбор нестабилен.

Максимальное качество прогноза достигается при использовании только одной позиции. Добавление второй позиции существенно снижает качество прогноза. Следовательно, только позиция 20 выравнивания аминокислотных последовательностей существенно связана с позицией 5 выравнивания сайтов.

Отбор пар позиций в задаче классификации

Как и в описанной выше задаче прогнозирования, отбор пар позиций для классификации пар “АП-сайт” проводили последовательно по взаимной информации и жадными алгоритмами с использованием наивного байесовского классификатора и логистической регрессии.

Таблица 5. Частоты пяти наиболее частых позиций в наборах различной длины, отобранных по методам MI, жадного наивного байесовского классификатора (NB) и жадной логистической регрессии (LR) для предсказания позиции 5 выравнивания сайтов (в %)

	MI					Bayes					LR				
	20	25	27	68	16	20	27	15	69	50	20	25	16	50	27
1	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0
2	1	0.95	0.03	0.02	0	1	0.55	0.02	0.2	0	1	0.54	0.33	0	0.13
3	1	0.96	0.35	0.41	0.21	1	0.61	0.28	0.58	0.18	1	0.87	0.69	0.16	0.25
4	1	0.99	0.62	0.62	0.53	1	0.62	0.48	0.6	0.28	1	0.94	0.73	0.6	0.44
5	1	0.99	0.85	0.83	0.77	1	0.64	0.67	0.61	0.49	1	1	0.75	0.85	0.62

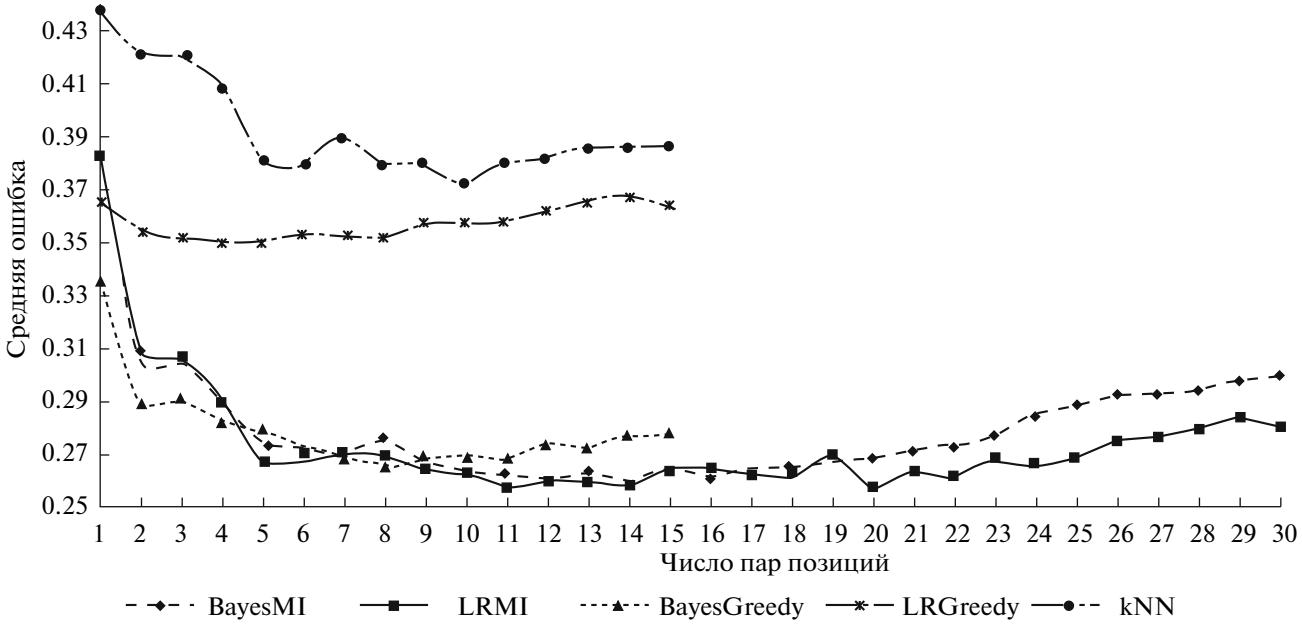

Рис. 7. Зависимость ошибки классификации от числа отобранных пар позиций. Представлены байесовский классификатор (Bayes MI), логистическая регрессия (LR MI) с отбором признаков по взаимной информации (MI) и они же с отбором жадным алгоритмом (Bayes Greedy и LR Greedy соответственно), метод ближайшего соседа (kNN) с отбором признаков по MI.

График зависимости средней ошибки всех алгоритмов классификации и методов отбора признаков от количества использованных пар позиций представлен на рис. 7. Кривые не имеют выраженных минимумов. Ошибка классификации при отборе по взаимной информации и с использованием жадного байесовского классификатора резко падает при добавлении второй пары позиций, и падение продолжается, пока уровень ошибки не стабилизируется в районе пяти–восьми пар позиций. Далее монотонность нарушается, и следующий минимум достигается уже на десяти–одиннадцати парах. Ошибка жадной логистической регрессии медленно падает до четырех пар позиций. Отметим также, что точность классификации алгоритма ближайшего соседа, использующего все позиции АП и сайта (на рисунке не изображен), составляет около 50%, что соответствует случайному угадыванию.

В табл. 6 представлены частоты наиболее часто отбираемых различными методами пар позиций. Номера столбцов соответствуют парам “номер позиций в АП – номер позиции в сайте”, начиная с самой частой. Номера строк – число пар позиций в отобранном наборе. Отбор по взаимной информации, в целом, более стабилен, чем жадный: абсолютно стабильно отбираются первые две пары позиций (55-9 и 20-5), после чего достаточно стабильно отбираются еще три (16-7, 16-6, 15-9). Жадный

Таблица 6. Частоты наиболее частых пар позиций в наборах различной длины, отобранных по методам MI, жадного наивного байесовского классификатора (NB) и жадной логистической регрессии (LR) (в %)

	MI							Bayes							LR						
	55-9	20-5	16-7	16-6	15-9	15-7	27-4	55-9	20-5	27-4	16-7	4-8	15-9	27-3	20-5	27-4	27-3	26-4	16-4	25-3	26-3
1	**0.58**	0.42	0	0	0	0	0	**1**	0	0	0	0	0	0	**0.92**	0.08	0	0	0	0	0
2	1	**1**	0	0	0	0	0	1	**1**	0	0	0	0	0	1	**1**	0	0	0	0	0
3	1	1	**0.92**	0.08	0	0	0	1	1	**0.36**	0.24	0	0	0	1	1	**0.46**	0.08	0.18	0.02	0.01
4	1	1	1	**0.99**	0.01	0	0	1	1	0.61	**0.6**	0.09	0.06	0	1	1	0.76	**0.12**	0.23	0.04	0.03
5	1	1	1	1	**0.9**	0.08	0.02	1	1	0.68	0.66	**0.38**	0.18	0.01	1	1	0.88	0.24	**0.28**	0.1	0.06
6	1	1	1	1	0.95	**0.73**	0.21	1	1	0.83	0.67	0.6	**0.23**	0.02	1	1	0.93	0.38	0.3	**0.2**	0.09
7	1	1	1	1	1	0.93	**0.62**	1	1	0.92	0.71	0.72	0.37	**0.1**	1	1	0.99	0.54	0.35	0.24	**0.19**

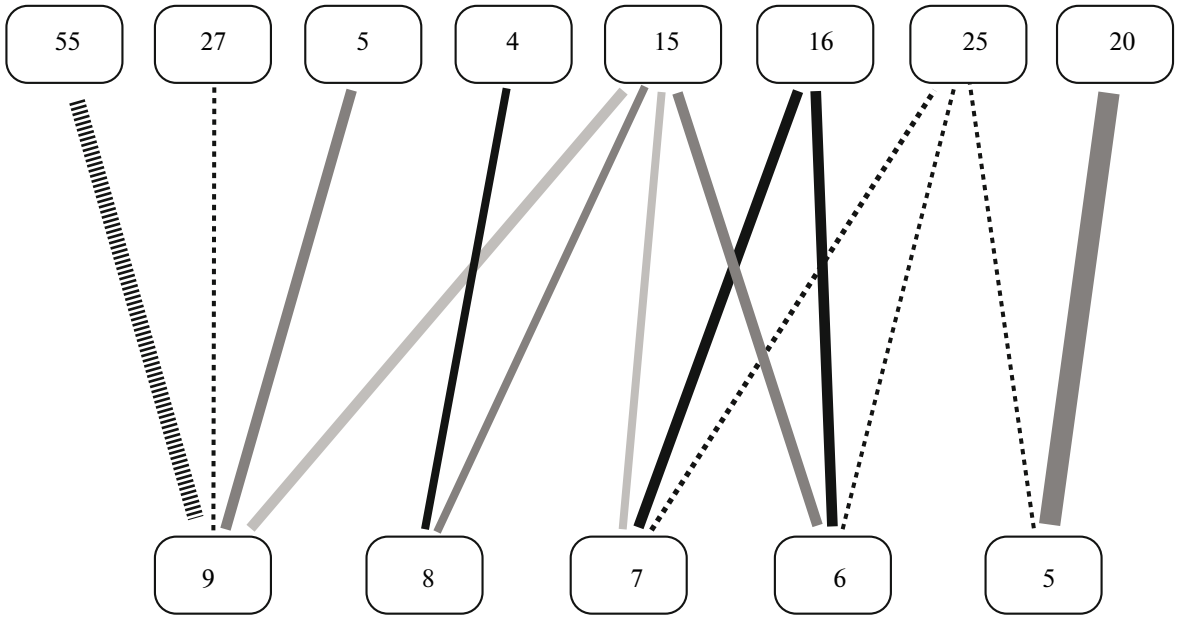

Рис. 8. Пары позиций, стабильно отбираемые разными методами. Вверху – позиции в аминокислотной последовательности (АП), внизу – позиции сайта. Связанные пары позиций соединены линиями. Толщина линий пропорциональна числу методов, в которых данная пара позиций отбиралась стабильно. Оттенок и штриховка линии показывают, во скольких известных структурах комплексов белок•ДНК данная пара образует контакт (по данным сервиса WHATIF [23]): черный – во всех трех структурах имеются специфические контакты для данной пары, темно-серый – в двух структурах, светло-серый — в одной, пунктир – ни в одной.

отбор с использованием наивного байесовского классификатора стабильно отбирает те же первые две пары (55-9 и 20-5), после чего отбор становится нестабильным. Логистическая регрессия стабильно отбирает две пары (20-5 и 27-4), после чего отбор нестабилен. Среди всех методов логистическая регрессия дает наибольшую ошибку классификации, что есть следствие выбора неправильных пар позиций (по-видимому, среди них только 20-5 значима).

Лучшие пары позиций для классификации пар "АП-сайт" – пары с наибольшим значением взаимной информации. Для лучшей классификации достаточно использования пяти–семи пар позиций. Можно предположить, что наиболее значимы для специфического связывания факторов транскрипции с сайтами ДНК пары позиций 55-9 и 20-5.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Эксперименты показали, что, если известно даже небольшое число ключевых позиций белковой последовательности, то этого достаточно для предсказания распределения нуклеотидов в выбранной позиции сайта. Эти позиции образуют значимо скоррелированные пары с соответствующими позициями выравнивания сайтов, имеют высокие значения взаимной информации и стабильно отбираются различными методами отбора. Этим наборам позиций соответствуют максимумы на графиках точности прогноза. Дальнейшее увеличение числа признаков ведет к переобучению.

Хотя качество прогноза сильно различается при разных разбиениях выборки на обучающую и тестовую, общие результаты (положение локальных максимумов, отобранные позиции, относительное качество прогноза разными алгоритмами) устойчивы к возмущениям исходных данных.

Стабильность отбора и образование очевидного максимума на графиках точности прогноза можно считать доказательством связи между отобранными позициями в аминокислотных последовательностях и позициями в сайтах связывания. Позиции в белковых последовательностях не взаимозаменяемы: лучшие предсказания получены при их одновременном использовании.

Те же пары позиций позволяют предсказать, связывается ли данный белок с данным сайтом ДНК: наиболее значимыми для специфического связывания факторов транскрипции с сайтами ДНК являются пары позиций 55-9 и 20-5 выравнивания аминокислотных последовательностей и выравнивания сайтов соответственно.

На рис. 8 изображены позиции в АП, стабильно отбирающиеся в процессе прогнозирования хотя бы одной из исследованных позиций сайта. Девять из четырнадцати представленных на рис. 8 пар (64.3%) соответствуют парам позиций, остатки в которых образуют специфические контакты между азотистым основанием и боковой группой а.о. хотя бы в одной из трех рассмотренных структур (коды PDB [22]: 1qpz, 1efa, 1rzr). И эти девять пар состав-

ляют ровно половину от объединенного множества всех специфических пар в структурах, что свидетельствует о том, что модель не случайна. Более того, важная функциональная роль взаимодействий 20-5, 16-6, 16-7, 15-7 подтверждена экспериментально [24]. Предсказываются также три пары с позицией 25 в АП. Это особая позиция в конце распознающей спирали предсказывалась ранее и как СДП-позиция [16], и как скоррелированная позиция [13], причем она всегда оказывалась в топ-списке. По-видимому, функциональная роль аминокислотных остатков в этой позиции недооценена. Из оставшихся трех пар в двух случаях (5-9, 27-9) аминокислотный остаток в указанной позиции образует специфические контакты с ДНК, но с нуклеотидами в других позициях.

Сопоставление с данными о специфических контактах в трех известных структурах комплексов белок-сайт для семейства LacI показывает, что большая часть наблюдавшихся пар образует специфические контакты хотя бы в одной структуре. При этом в различных структурах контакты могут быть образованы разными парами. Это дает основания полагать, что остальные пары также могут образовывать контакты в других, не исследованных пока, членах семейства.

Работа получила финансовую поддержку Российского фонда фундаментальных исследований (09-04-92745 и 10-04-00431), и Государственного контракта 2.740.11.0101.

СПИСОК ЛИТЕРАТУРЫ

1. Suzuki M., Brenner S.E., Gerstein M., Yagi N. 1995. DNA recognition code of transcription factors. *Protein Eng.* **8**, 319–328.
2. Jones S., Shanahan H.P., Berman H.M., Thornton J.M. 2003. Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. *Nucl. Acids Res.* **31,** 7189–7198.
3. Baker C.M., Grant G.H. 2007. Role of aromatic amino acids in protein-nucleic acid recognition. *Biopolymers.* **85,** 456–470.
4. Sarai A., Kono H. 2005. Protein-DNA recognition patterns and predictions. *Annu. Rev. Biophys. Biomol. Struct.* **34,** 379–398.
5. Sandelin A., Wasserman W.W. 2004. Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J. Mol. Biol.* **338,** 207–215.
6. Mahony S., Auron P.E., Benos P.V. 2007. Inferring protein-DNA dependencies using motif alignments and mutual information. *Bioinformatics.* **23**, i297–i304.
7. Ahmad S., Sarai A. 2005. PSSM-based prediction of DNA binding sites in proteins. *BMC Bioinformatics.* **6,** 33–34.
8. Ofran Y., Mysore V., Rost B. 2007. Prediction of DNA-binding residues from sequence. *Bioinformatics.* **23,** i347–i353.
9. Yan C., Terribilini M., Wu F., et al. 2006. Predicting DNA-binding sites of proteins from amino acid sequence. *BMC Bioinformatics.* **7,** 262–262.
10. Mirny L.A., Gelfand M.S. 2002. Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. *J. Mol. Biol.* **321,** 7–20.
11. Kalinina O.V., Mironov A.A., Gelfand M.S., Rakhmaninova A.B. 2004. Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. *Protein Sci.* **13,** 443–456.
12. Donald J.E., Shakhnovich E.I. 2005. Predicting specificity-determining residues in two large eukaryotic transcription factor families. *Nucl. Acids Res.* **33,** 4455–4465.
13. Korostelev Y., Laikova O.N., Rakhmaninova A.B., Gelfand M.S. *First RECOMB Satellite Conference on Bioinformatics Education.* 2009. San Diego, USA. Abstract book. P. 13.
14. Novichkov P.S., Laikova O.N., Novichkova E.S., Gelfand M.S., Arkin A.P., Dubchak I., Rodionov D.A. 2010. RegPrecise: a database of curated genomic inferences of transcriptional regulatory interactions in prokaryotes. *Nucl. Acids Res.* **38,** D111–D118.
15. Schultz J., Milpetz F., Bork P., Ponting C.P. 1998. SMART, a simple modular architecture research tool: Identification of signaling domains. *Proc. Natl. Acad. Sci. USA.* **95,** 5857–5864.
16. Kalinina O.V., Novichkov P.S., Mironov A.A., Gelfand M.S., Rakhmaninova A.B. 2004. SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. *Nucl. Acids Res.* **32,** W424–W428.
17. Gerstein M., Sonnhammer E.L., Chothia C. 1994. Volume changes in protein evolution. *J. Mol. Biol.* **236,** 1067–1078.
18. Domingos P., Pazzani M. 1997. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning.* **29,** 103–137.
19. Hosmer D., Lemeshow S. 2000. *Applied Logistic Regression, 2nd ed.* N.Y. Chichester: Wiley.
20. Peng H.C., Long F., Ding C. 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* **27**, 1226–1238.
21. Henikoff S., Henikoff J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA.* **89**, 10915–10919.
22. Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. 2000. The protein data bank. *Nucl. Acids Res.* **28,** 235–242.
23. Rodriguez R., Chinea G., Lopez N., Pons T., Vriend G. 1998. Homology modeling, model and software evaluation: three related resources. *Comput. Appl. Biosci.* **14,** 523–528.
24. Sartorius J., Lehming N., Kisters B., von Wilcken-Bergmann B., Müller-Hill B. 1989. Lac repressor mutants with double or triple exchanges in the recognition helix bind specifically to lac operator variants with multiple exchanges. *EMBO J.* **8**, 1265–1270.