МОЛЕКУЛЯРНАЯ БИОЛОГИЯ, 2007, том 41, № 1, с. 151–162

МАТЕМАТИЧЕСКАЯ И СИСТЕМНАЯ БИОЛОГИЯ

УДК 577.112:004.02

Ускоренная публикация

ВЫЧИСЛИТЕЛЬНЫЙ МЕТОД ДЛЯ ПРЕДСКАЗАНИЯ ФУНКЦИОНАЛЬНЫХ САЙТОВ БЕЛКА С ИСПОЛЬЗОВАНИЕМ ДЕТЕРМИНАНТ СПЕЦИФИЧНОСТИ

© 2007 г. О. В. Калинина[1, 2], Р. Б. Расселл[2], А. Б. Рахманинова[1, 3], М. С. Гельфанд[1, 3]*

[1]*Факультет биоинженерии и биоинформатики Московского государственного университета, Москва, 119992*

[2]*European Molecular Biology Laboratory, 69117 Heidelberg, Germany*

[3]*Институт проблем передачи информации Российской академии наук, Москва, 127994*

Поступила в редакцию и принята к печати 05.09.2006 г.
Представлена к публикации А.В. Финкельштейном

В настоящее время объем расшифрованных аминокислотных последовательностей белков во много раз превышает экспериментальные возможности по их функциональной аннотации. Поэтому все большую роль начинает играть аннотация *in silico* – методами биоинформатики. Такая аннотация с необходимостью носит характер предсказания, но может служить важной отправной точкой для дальнейших лабораторных исследований. В настоящей работе представлен новый метод предсказания функционально важных сайтов белка “SDPsite”, основанный на предсказании детерминант специфичности. На основании выравнивания белкового семейства и филогенетического дерева алгоритм предсказывает консервативные позиции и детерминанты специфичности, картирует их на структуру белка и ищет области кластеризации предсказанных позиций. Сравнение полученных предсказаний с экспериментальными данными и опубликованными результатами других методов предсказания функционального сайта показывает, что результаты SDPsite хорошо согласуются с экспериментом и превосходят результаты большинства ранее известных методов. SDPsite свободно доступен через Интернет по адресу http://bioinf.fbb.msu.ru/SDPsite.

Ключевые слова: структурная геномика, функциональный сайт, предсказание, сравнительный анализ последовательностей, детерминанты специфичности, определяющие специфичность позиции.

COMPUTATIONAL METHOD FOR PREDICTION OF PROTEIN FUNCTIONAL SITES USING SPECIFICITY DETERMINANTS, by *O. V. Kalinina[1, 2], R. B. Russell[2], A. B. Rakhmaninova[1, 3], M. S. Gelfand[1, 3]** ([1]Department of Bioengineering and Bioinformatics, Moscow State University, Moscow, 119992 Russia, *e-mail: gelfand@iitp.ru; [2]European Molecular Biology Laboratory, 69117, Heidelberg, Germany; [3]Institute for Information Transmission Problems Russian Academy of Sciences, Moscow, 127994 Russia). The current available data on protein sequences largely exceeds the experimental capabilities to annotate their function. So annotation *in silico*, i.e. using computational methods becomes increasingly important. This annotation is inevitably a prediction, but it can be an important starting point for further experimental studies. Here we present a method for prediction of protein functional sites, SDPsite, based on the identification of protein specificity determinants. Taking as an input a protein sequence alignment and a phylogenetic tree, the algorithm predicts conserved positions and specificity determinants, maps them onto the protein’s 3D structure, and searches for clusters of the predicted positions. Comparison of the obtained predictions with experimental data and data on performance of several other methods for prediction of functional sites reveals that SDPsite agrees well with the experiment and outperforms most of the previously available methods. SDPsite is publicly available under http://bioinf.fbb.msu.ru/SDPsite.

Key words: structural genomics, functional site, prediction, sequence comparative analysis, specificity determinants, SDP.

Принятые сокращения: ПОС – определяющие специфичность позиции; КП – консервативные позиции; PDB – банк данных белковых структур; GOLD – база данных геномов; CDD – база данных консервативных доменов.

*Эл. почта: gelfand@iitp.ru

ВВЕДЕНИЕ

Экспоненциальный рост объемов баз данных, содержащих информацию о секвенированных последовательностях ДНК, значительно превышает экспериментальные возможности по аннотации (описанию функциональных характеристик) этих последовательностей. В настоящее время секвенировано 335 полных бактериальных, 41 полных эукариотических и 27 полных геномов археобактерий, еще 1596 таких проектов находятся на стадии исполнения (по данным базы данных GOLD [1]). Предварительная аннотация компьютерными методами стала в этих проектах частью рутинной процедуры. Кроме последовательности крайне важно для понимания функции белка знать его пространственную структуру. В 2000 г. стартовал международный Проект по структурной геномике [2], целью которого является расшифровка представительного набора пространственных структур белков различных изученных организмов. Основные этапы этого проекта следующие: 1) организовать все известные последовательности белков в семейства; 2) выбрать в качестве мишени одного или несколько представителей семейства; 3) расшифровать пространственную структуру мишени с помощью рентгеноструктурного анализа или ЯМР; и 4) построить модели пространственной структуры других представителей семейства. В результате реализации этого проекта будут получены структуры многих белков, для которых не только локализация их активных центров и/или других функциональных сайтов, но часто и сама функция неизвестны, и, более того, они не имеют подробно изученных гомологов. В таких случаях часто применяют различные вычислительные методы поиска функциональных сайтов.

Обзор некоторых методов, комбинирующих информацию о выравнивании последовательностей с информацией о пространственных структурах, приведен в [3]. Эти методы направлены на поиск областей на поверхности белка, важных для функции или специфичности белка. Например, анализ структуры белка Mj0577, расшифрованной в ходе Проекта по структурной геномике с помощью метода ConSurf [4], помог определить положение сайта связывания АТФ, а также показал функциональную важность поверхности контакта субъединиц гомодимера [5].

Некоторые методы поиска функционального сайта, такие как ConSurf [4] и метод Алой и соавт. [6], предполагают, что если позиция консервативна в выравнивании родственных последовательностей, она функционально важна. В методах [7, 8] вместо консервативности в последовательности используют более мягкую "структурную консервативность", но основная идея метода та же самая. Дель Соль Меса и соавт. [9] вводят понятие "коррелированных мутаций" в последовательности (одновременная мутация в двух далеких позициях одного выравнивания или разных выравниваний) и считают позиции, подвергавшиеся в ходе эволюции таким мутациям, функционально важными. Лихтарж и соавт. разработали метод "эволюционного следа" [10, 11], суть которого состоит в том, что белки выравнивания разбиваются на группы на разных уровнях сходства последовательностей, а потом для каждой группы выделяются консервативные в ней позиции. Последовательность консервативных в группе позиций называется ее "эволюционным следом". Эволюционные следы для разных групп сравниваются, и важными считаются те позиции, которые входят в эволюционные следы большого количества групп.

Ханненхалли и Расселл [12] и Мирный и Гельфанд [13] разработали методы поиска детерминант специфичности (к лиганду, ДНК, другому белку и т.п.) в выравнивании белковых последовательностей: выравнивание считалось заранее разбитым на группы белков одинаковой специфичности, детерминантами специфичности названы позиции, консервативные внутри групп специфичности, но различающиеся между группами. На основании этих методов мы предложили алгоритм SDPpred [14], который повторяет основные черты алгоритма [13], но технически более пригоден для анализа большого количества данных, так как имеет автоматическую процедуру выбора порога, а также лучше учитывает эволюционное расстояние между белками и сходство аминокислот.

В настоящей работе мы представляем новый алгоритм предсказания функциональных сайтов белка "SDPsite". Он сочетает в себе черты многих упомянутых выше методов: в выравнивании последовательностей находятся консервативные позиции; на основании выравнивания и филогенетического дерева предсказываются детерминанты специфичности (для этого разработана специальная процедура автоматического поиска групп специфичности); с использованием пространственной структуры одного из белков выравнивания выделяется наилучший кластер детерминант специфичности и консервативных позиций. SDPsite свободно доступен по адресу http://bioinf.fbb.msu.ru/SDPsite.

SDPsite протестирован на семействе бактериальных факторов транскрипции LacI, субтилизин-подобных протеаз и 68 доменах из базы данных консервативных доменов CDD.

ОПИСАНИЕ АЛГОРИТМА

Алгоритм предсказания функционального сайта состоит из трех частей: 1) предсказание детер-

минант специфичности (при этом применяется автоматическое разбиение выравнивания на группы специфичности); 2) предсказание консервативных позиций и 3) выделение лучшего кластера.

Предсказание детерминант специфичности

Алгоритм предсказания *позиций, определяющих специфичность* (ПОС), описан в [14]. В качестве входных данных для предсказания используется выравнивание аминокислотных последовательностей, в котором белки разделены на *группы специфичности*. Предполагается, что в одну группу специфичности входят белки с одинаковой специфичностью к субстрату, а у белков разных групп специфичность отличается. Вкратце алгоритм выглядит следующим образом. Отдельно рассматривается каждая позиция выравнивания. Чтобы оценить, является ли эта позиции ПОС, используется взаимная информация

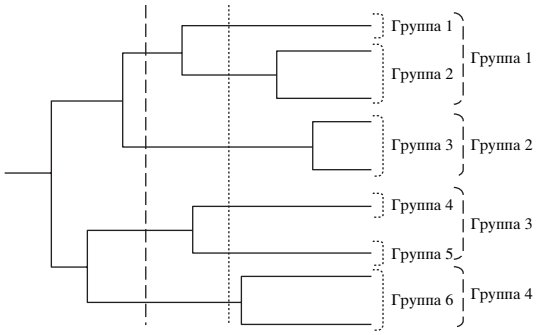

Рис. 1. Разбиение на группы в методе эволюционного следа. Две возможные группировки показаны пунктирной и прерывистой линиями.

$$I_p = \sum_{\text{по всем группам специфичности } i} \; \sum_{\text{по всем аминокислотам } \alpha} f_p(\alpha, i)\log\frac{f_p(\alpha, i)}{f_p(\alpha)f(i)}, \qquad (1)$$

где $f_p(\alpha, i)$ – частота аминокислоты α в позиции p в группе i, $f_p(\alpha)$ – частота аминокислоты α в позиции p во всей выборке, $f(i)$ – размер (доля) группы i.

При этом для учета особенностей реальных биологических данных вводится ряд поправок, подробно обсуждавшихся в [14]. С помощью случайных перемешиваний колонки вычисляется среднее и стандартное отклонение распределения ее ожидаемого информационного содержания $\langle I_p^{\exp}\rangle$ и $\sigma(I_p^{\exp})$, а затем статистическая значимость для каждой позиции:

$$Z_p = \frac{I_p - \langle I_p^{\exp}\rangle}{\sigma(I_p^{\exp})}. \qquad (2)$$

Для определения количества ПОС среди наиболее значимых позиций применяется оригинальная процедура, основанная на оценке Бернулли [15]. Сначала все позиции упорядочиваются по убыванию значимости Z_p. Далее выбирается такое значение k^*, для которого получение k^* значений Z, не меньших $Z_{(k^*)}$, наименее вероятно при условии нормального распределения Z (выбирается наименее вероятный в случайной ситуации набор позиций, "тяжелый хвост"; P – вероятность отсечки):

$$k^* = \arg\min_k P\{\text{существует, по крайней мере, } k \text{ наблюдений: } Z \ge Z_{(k)}\} = \arg\min_k\left(1 - \sum_{i=L-k+1}^{L} C_L^i q^i p^{L-i}\right), \qquad (3)$$

где

$$p = P(Z \ge Z_{(k)}) = \int_{Z_{(k)}}^{\infty} \frac{1}{\sqrt{2\pi}}\exp(-Z^2)dZ,$$

$$q = 1 - p.$$

Таким образом выделяется набор из k^* ПОС. Вероятность

$$P^* = P\{\text{существует, по крайней мере, } k \text{ наблюдений: } Z \ge Z_{(k)}\}, \qquad (4)$$

доставляющая этот минимум, называется *статистической значимостью набора из k* позиций.*

Автоматическое деление на группы специфичности

Чтобы широко применять алгоритм SDPsite, необходима процедура, позволяющая автоматически разделять выравнивание на группы специфичности. Мы использовали способ, аналогичный применяемому в методе эволюционного следа [10]. Мы рассматриваем неукорененное исходное дерево и предполагаем, что корень находится в середине самого длинного пути от одного листа к другому. После этого рассматривается набор группировок, получаемых рассечением дерева на определенном расстоянии от корня (рис. 1). При этом группы, содержащие меньше трех последовательностей, не рассматривались. ПОС находятся для каждой группировки, как описано в [14], и вычисляется статистическая значимость данного набора ПОС P^* по формуле (4). Лучшим считает-

ся тот набор, P^* для которого минимальна, т.е. *наименее вероятный набор* ПОС.

При этом вычисляемые в формуле (2) величины Z нуждаются в корректировке: если в выравнивание добавлять последовательности, не меняя количества групп специфичности, но равномерно увеличивая количество последовательностей в каждой группе, максимальная величина Z растет, и ее рост достаточно хорошо можно аппроксимировать логарифмической функцией (данные не показаны). Это соответствует представлению о статистической значимости с точки зрения здравого смысла. Однако увеличение статистической значимости для групп с большим количеством последовательностей не позволяет корректно сравнивать разделения одного выравнивания на разное количество групп: разделение на более "толстые" группы всегда выигрывает. Чтобы скомпенсировать указанный логарифмический рост, мы вводим поправку

$$Z := Z/\log(\text{средняя толщина группы}). \quad (5)$$

Предсказание консервативных позиций

Различные подходы к определению консервативных позиций описаны в обзоре [16]. В настоящей работе использована мера консервативности Сандера–Шнайдера [17]: консервативность позиции p вычисляется по формуле

$$C_p = \left(\sum_i^N \sum_{j>i}^N d(s_i, s_j) M(s_i(p), s_j(p))\right) / \left(\sum_i^N \sum_{j>i}^N d(s_i, s_j)\right), \quad (6)$$

где N – количество последовательностей в выравнивании; $d(s_i, s_j)$ – расстояние между последовательностями s_i и s_j, равное $1 - \frac{\%\ \text{идентичности}}{100}$; $s_k(p)$ – аминокислота, стоящая в последовательности s_k на позиции p; $M(\alpha, \beta)$ – матрица замен аминокислот, в данном случае использовалась матрица BLOSUM62 [18].

Как указано в обзоре [16], эта мера консервативности вполне удовлетворительна с точки зрения здравого смысла: ее область значений непрерывна и ограничена (отрезок от 0 до 1); она учитывает частоты аминокислот в столбце и частоты замен аминокислот и их физико-химические свойства с помощью включения матриц замен аминокислот; она нормализована с учетом вырожденности выравнивания (учитывает расстояния между последовательностями).

Для каждого значения C_p вычисляется его статистическая значимость. Мы вводим фоновое распределение C_p как консервативность колонок, составленных из случайных позиций каждой последовательности выравнивания. Таким образом, мы вычисляем для каждого C_p 10000 случайных значений консервативности C_p^{rand}, а затем вычисляем статистическую значимость

$$\tilde{Z}_p = \frac{C_p - \langle C_p^{\text{rand}} \rangle}{\sigma(C_p^{\text{rand}})}. \quad (7)$$

Здесь C_p^{rand} отвечает консервативности колонок в наборе невыравненных последовательностей. Поскольку выравнивание двух случайных последовательностей имеет ненулевой вес, мы центрируем полученную статистическую значимость и, в конце концов, получаем

$$Z_p = \tilde{Z}_p - \langle \tilde{Z}_p \rangle. \quad (8)$$

Далее мы используем такую же процедуру выбора числа значимых позиций, как и при предсказании ПОС (формулы 3 и 4).

Выделение лучшего кластера

Чтобы предсказать ПОС и консервативные позиции (КП), алгоритму требовалось только выравнивание последовательностей семейства белков и соответствующее ему филогенетическое дерево. Для выделения лучшего кластера нужна еще трехмерная структура одного из белков семейства. Если в семействе есть несколько белков с расшифрованными трехмерными структурами, то получаемый кластер может зависеть от выбора структуры. Однако тесты на реальных примерах показывают, что лучшие кластеры для разных структур достаточно сильно перекрываются (данные не показаны).

На заданной трехмерной структуре белка алгоритм находит остатки, соответствующие предсказанным ПОС и КП, и проводит их пространственную кластеризацию по алгоритму вложенных кластеров, основанному на плотности графа [19]. Вложенные кластеры строятся следующим образом. Сначала рассматриваются все вершины графа (в нашем случае они соответствуют множеству всех ПОС и КП на пространственной структуре) – кластер H_0. Для каждой вершины i вычисляется ее вес по формуле

$$\mu_i = \lambda_i \sum_j \omega_{ij}, \quad (9)$$

где j пробегает множество всех остальных вершин H_0, а ω_{ij} – вес ребра между вершинами i и j, вычисляемый по формуле

$$\omega_{ij} = \begin{cases} \frac{R}{d_{ij}}, & \text{если} \quad d_{ij} < D \\ 0, & \text{если} \quad d_{ij} \geq D, \end{cases} \qquad (10)$$

где d_{ij} – евклидово расстояние между ближайшими атомами аминокислот, соответствующих вершинам i и j, R = 5 Å – среднее расстояние между центрами атомов, при котором атомы находятся в контакте, D = 15 Å – расстояние, на которое распространяется влияние атома. R и D – константы, значения которых были подобраны из эмпирических и эвристических соображений. $\lambda_i = 0.5$, если вершина i соответствует КП, и 1 в противном случае. Таким образом, значимость КП искусственно занижается. Это сделано для того, чтобы алгоритм не выбирал геометрическое ядро (группу консервативных остатков, необходимых для образования правильной пространственной структуры белка) в качестве значимого кластера.

Далее находится множество вершин $F_0 \subset H_0$, для которых значение μ минимально и равно $\mu_0^{\min}$. Строится кластер $H_1 = H_0 \backslash F_0$. Эта процедура повторяется, пока на некотором шаге не будет получено пустое множество. Таким образом, будет построена серия вложенных кластеров $H_0 \supset \supset H_1 \supset \text{K} \supset H_N \supset \varnothing$. В качестве самого значимого кластера выбирается кластер n, для которого $\mu_n^{\min} = \max\{\mu_k^{\min} | k = 0, \ldots, N\}$. Этот кластер мы далее называем *лучшим кластером*.

В настоящей работе рассматривали два наиболее значимых кластера. Второй кластер находится по тому же самому алгоритму, с предварительным исключением из H_0 всех вершин, вошедших в первый кластер (*второй лучший кластер*).

Алгоритм предсказания функционального сайта был назван SDPsite и реализован в виде веб-сервера, доступного по адресу http://bioinf.fbb.msu.ru/SDPsite.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Алгоритм SDPsite протестирован на трех примерах. SDPsite применен к семейству бактериальных факторов транскрипции LacI, включающему в себя регуляторы катаболизма различных сахаров, а также некоторых других метаболических путей. Есть обширные данные о специфичности различных белков этого семейства [20] и влиянии мутации каждого остатка на функционирование белка [21]. Результаты применения SDPsite хорошо согласуются с имеющимися данными. Проведено сравнение работы SDPsite с другими методами предсказания функционального сайта, описанными в работе [22]. На рассмотренных в этой работе примерах (LacI и субтилизин-подобных протеазах) результаты SDPsite лучше, чем у других методов. SDPsite применен к большому количеству семейств из базы данных NCBI CDD (база данных консервативных доменов). Эта база данных содержит выравнивания белковых доменов, в которых некоторые позиции отмечены как "особенности" ("features") – активный центр, поверхность контакта с лигандом, сайт фосфорилирования и т.п. Мы предполагаем, что эти "особенности" и являются функционально важными позициями. Несмотря на то, что при таком подходе мы неизбежно недооцениваем собственные результаты (не отмеченные "особенностями" позиции также могут быть функционально важными, а набор "особенностей" включает позиции, которые не подходят под определение функционального сайта, такие как сайты фосфорилирования, гликозилирования и т.п.), SDPsite дает удовлетворительные результаты.

Таблица 1. Группы специфичности, выделяемые при автоматической группировке

Группа при автоматической группировке	Соответствующие группы, выделенные в [20]
Группа 1	CcpA
Группа 2	CytR
Группа 3	GntR
Группа 4	FruR + ScrR
Группа 5	MalR
Группа 6	GalR
Группа 7	RbsR(PP)
Группа 8	PurR + RbsR(EC)

Применение SDPsite к семейству бактериальных факторов транскрипции LacI

Рассмотрено выравнивание некоторых регуляторов семейства LacI, содержащее 125 последовательностей, разбитых на следующие группы специфичности, различающиеся типом эффектора и операторной последовательности ДНК: PurR, ScrR, RbsR(EC), GntR, RbsR(PP), GalR, MalR, CytR, CcpA, FruR. Эта группировка получена с помощью анализа геномного контекста и регуляторных сайтов методами сравнительной геномики [20]. Эволюционные отношения белков и их разделение на группы представлены на филогенетическом дереве (рис. 2*a*). Мы использовали структуру PurR из *E. coli* (идентификатор банка пространственных структур белков PDB – 1bdh) для визуализации предсказаний и нахождения кластеров.

Приведенное разделение на группы специфичности не использовано при предсказании ПОС, напротив, группы выделяли автоматически. При

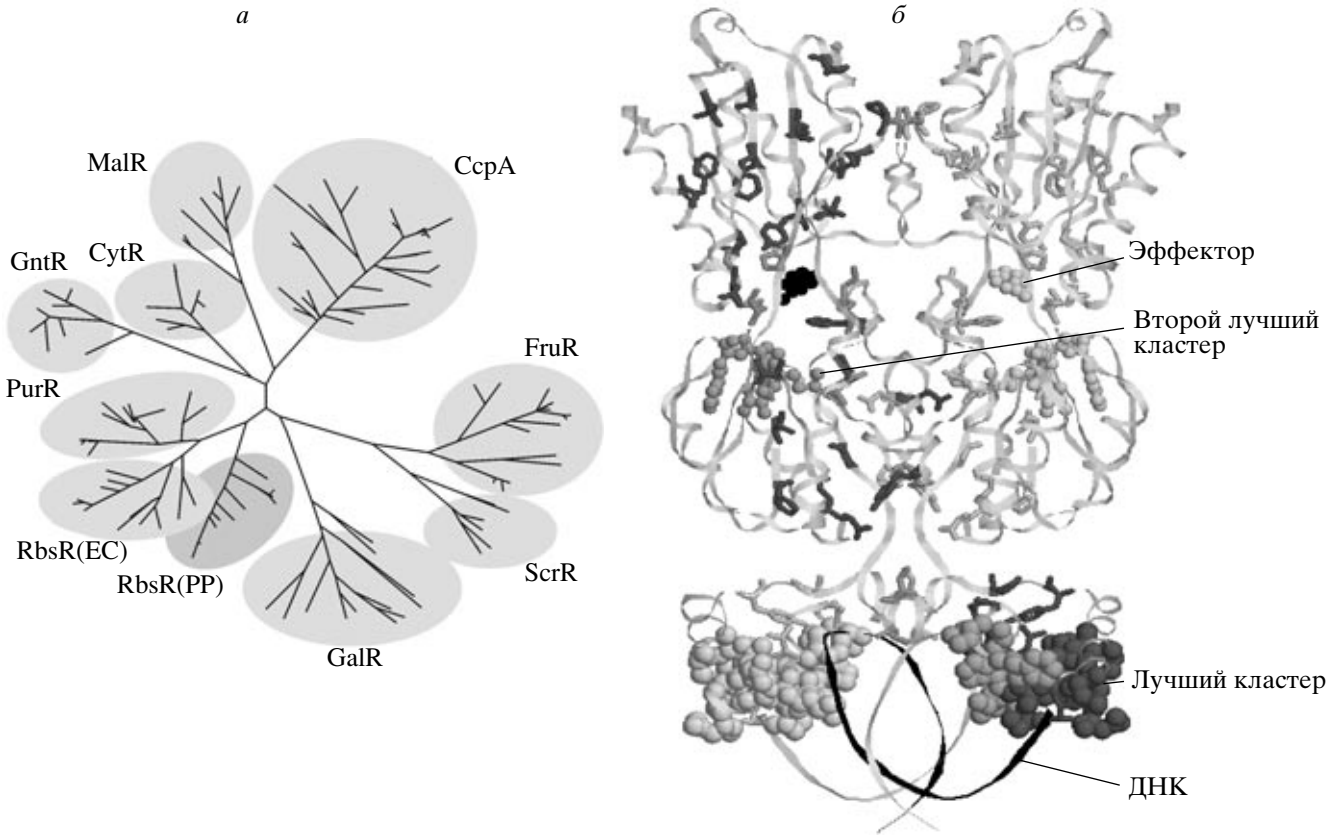

Рис. 2. *а* – Филогенетическое дерево семейства LacI. *б* – ПОС (светло-серый) и КП (темно-серый) на структуре PurR из *E. coli* (идентификатор PDB 1bdh).

этом выделяемые группы практически совпадают с исходными (табл. 1). Предсказанные позиции приведены в табл. 2 (нумерация по PurR из *E. coli*) и на рис. 2*б*. Как видно, большинство предсказанных ПОС расположено либо в области контакта с эффектором, либо в области контакта с ДНК, либо на поверхности контакта между субъединицами. Это соответствует интуитивным представлением о том, что именно в этих областях должны быть расположены аминокислотные остатки, отвечающие за специфическое взаимодействие белка с лигандами или другой субъединицей. КП также встречаются в этих областях (особенно в области контакта с ДНК), однако их значительно больше внутри белковой глобулы, где они не доступны для растворителя, следовательно, не могут непосредственно участвовать в функционировании белка, а служат, вероятно, для стабилизации пространственной структуры. В двух наиболее важных функциональных сайтах этого белка (ДНК-связывающий домен и карман, связывающий эффектор) расположены два лучших кластера, найденных SDPsite.

Влияние мутации каждого аминокислотного остатка последовательности LacI на функцию

Таблица 2. Предсказанные позиции для семейства LacI (нумерация по PurR из *E. coli*)

Тип позиции	Количество предсказанных позиций	Номера по PurR из *E. coli*
ПОС	20	5, 15, 16, 20, 25, 27, 53, 55, 91, 96, 123, 144, 145, 146, 147, 160, 162, 284, 294, 323
КП	40	3, 6, 7, 8, 11, 12, 13, 14, 17, 19, 23, 28, 32, 35, 36, 45, 47, 63, 74, 82, 90, 117, 118, 141, 143, 158, 161, 181, 186, 200, 242, 244, 248, 253, 266, 271, 274, 285, 287, 298
Лучший кластер	19	5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 17, 19, 20, 25, 27, 28, 32, 35, 36
Второй лучший кластер	4	144, 145, 146, 147

белка описано Саков и соавт. [21]. Все остатки разбиты на классы в зависимости от того, может ли этот остаток быть заменен, и каков эффект от возможной замены. На основании этих классов мы разделили все аминокислотные остатки на пять групп: 1) замена остатков не влияет на функционирование белка; 2) замена без потери функции возможна только на малые аминокислоты; 3) остатки, замена которых нарушает связывание эффектора или передачу сигнала, не контактирующие с эффектором непосредственно; 4) аминокислотные остатки, которые нельзя заменить без потери функции, не контактирующие с эффектором или ДНК непосредственно; и 5) остатки, контактирующие с эффектором или ДНК, которые нельзя заменить без потери функции. Распределение всех аминокислотных остатков белка, а также предсказанных позиций по этим группам показано на рис. 3. Видно, что в двух наиболее важных для функции группах, (4) и (5), доля ПОС и КП выше, чем в среднем, а доля кластеров еще выше. С другой стороны, чем более значима группа для функции белка, тем больше в ней относительное количество предсказанных позиций, и большее количество их концентрируется в кластерах.

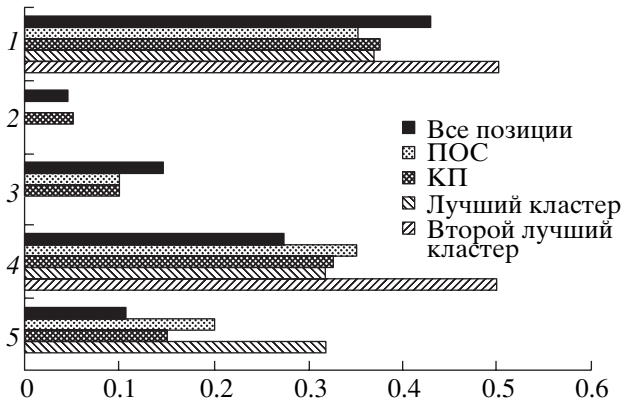

Рис. 3. Распределение предсказанных позиций по группам важности для функции в случае выборки из [20]. *1* – Остатки можно заменить без ущерба для функции; *2* – остатки можно заменить только на малые аминокислоты; *3* – замена этих остатков нарушает связывание эффектора или передачу сигнала, но они не контактируют с эффектором непосредственно; *4* – остатки нельзя заменить, но они не контактируют с эффектором или ДНК непосредственно и *5* – остатки контактируют с эффектором или ДНК, их нельзя заменить. Доля групп 4 и 5 выше среди остатков, формирующих кластеры.

Сравнение SDPsite с другими методами

Сойер и Голдстейн сравнивают несколько методов поиска функционального сайта на примере семейств LacI и субтилизин-подобных протеаз [22]: вычисление частоты наиболее частой аминокислоты [16]; индекс консервативности, основанный на энтропии [16]; индекс консервативности Валдар–Торнтон [16]; метод эволюционного следа [10]; ConSurf [4]; логарифм правдоподобия, вычисленный с помощью PAML [23]; модель эволюции по классам сайтов [22]. Мы сравнили SDPsite с результатами, приведенными в этом исследовании.

Как указано выше, для LacI имеются полные данные о влиянии замены каждого остатка белка на его функцию [21]. В случае субтилизина также имеются обширные (приблизительно для половины остатков белка) данные о влиянии мутаций различных остатков на функцию [24]. Остальные позиции могут быть либо не важными для функции, либо никогда не исследовались. В дальнейшем мы предполагаем, что они не важны и, возможно, искусственно занижают качество результатов наших предсказаний.

Выравнивания для этих тестов построены так же, как в [22]. С помощью программы BLAST для LacI проведен поиск белков из *E. coli* (P03023), похожих на LacI, по базе данных SwissProt с условием E-value > 0.001. Получено 75 последовательностей. После того как были удалены последовательности с большими концевыми делециями, осталось 70 последовательностей из 24 бактериальных геномов со средним сходством 23.5% с LacI из *E. coli*. В случае субтилизин-подобных протеаз мы взяли семейство HBG020722 из базы данных HOBACGEN, выпуск 10 [25]. Оно содержит 80 последовательностей со средним сходством 35% с субтилизином из *Bacillus amyloliquefaciens*. Мы удалили из выравнивания 10 последовательностей, содержащих большие концевые делеции, и рассмотрели только часть, соответствующую активному ферменту.

Предсказанные позиции, наложенные на структуру PurR из *E. coli* в случае LacI и субтилизин из *B. amyloliquefaciens* в случае субтилизин-подобных протеаз, показаны на рис. 4. Видно, что в LacI потерян кластер в ДНК-связывающем домене, однако в целом предсказание остается хорошим. В случае субтилизина ПОС предсказаны слабо, поэтому лучший кластер состоит, в основном, из КП.

В работе [22] для оценки качества предсказания использована мера "отношение чувствительности к перепредсказанию". Чувствительность вычисляется по формуле $TP/(TP + FN)$, где *TP* (true positives) – остатки, предсказанные методом и действительно являющиеся важными; *FN* (false negatives) – остатки, не предсказанные методом, но являющиеся важными. Перепредсказание вычисляется по формуле $FP/(FP + TN)$, где *FP* (false positives) – остатки, предсказанные методом, но не являющиеся важными; *TN* (true negatives) – остатки, не предсказанные методом и не являющиеся важными. Все исследованные методы не предсказывают кластера важных остатков, а предоставляют непрерывный список остатков,

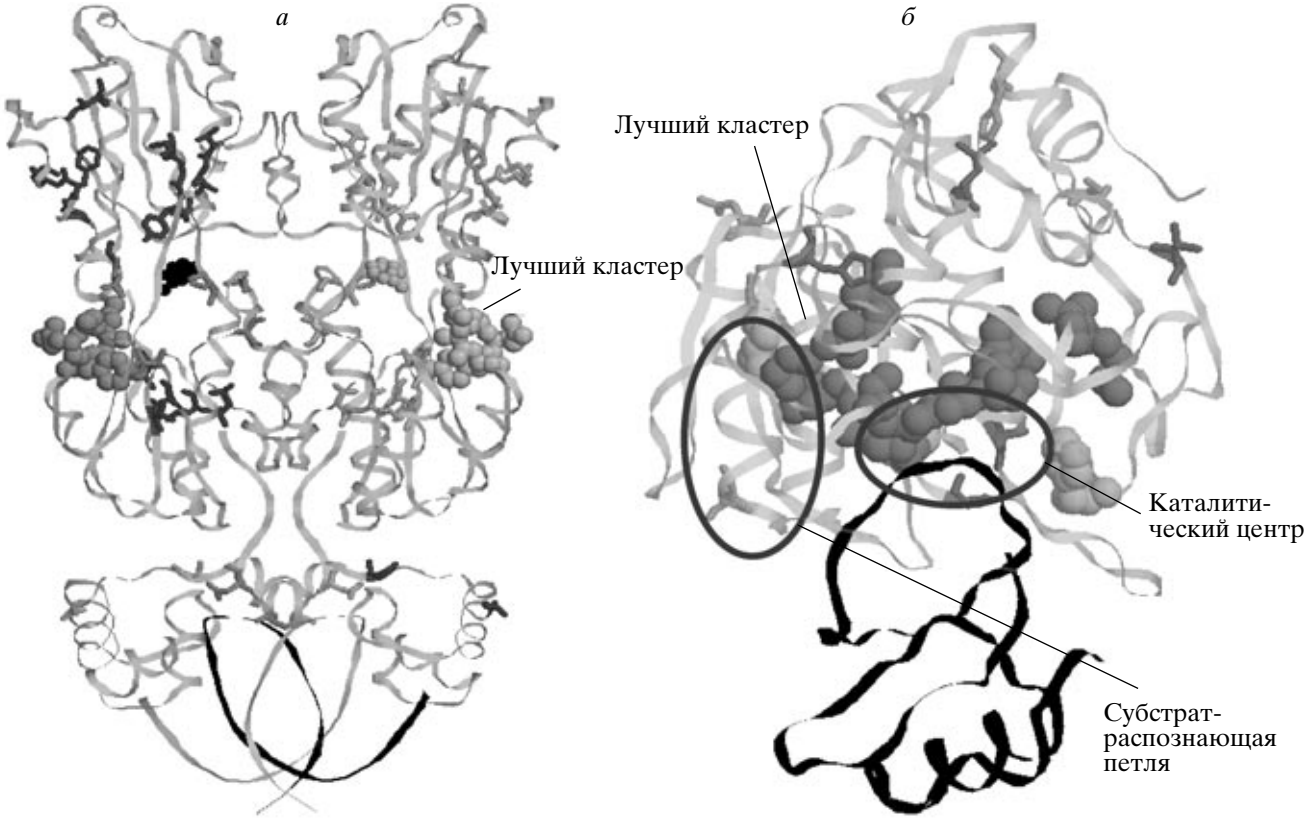

Рис. 4. Предсказанные позиции на структуре PurR из *E. coli* (идентификатор PDB 1bdh) (*а*) и субтилизина из *B. amyloliquefaciens* (идентификатор PDB 1to2) (*б*). ПОС показаны светло-серым, КП показаны темно-серым.

отсортированных по степени их предсказанной важности. Поэтому отношение $TP/(TP + FN)$ vs. $FP/(FP + TN)$ представляет собой не точку, а график, называемый кривая ROC (receiver operating characteristic, операторное принимающее устройство). Для LacI все методы дают удовлетворительные предсказания – на большом диапазоне чувствительность значительно превосходит уровень перепредсказания, – тогда как для субтилизина все предсказания не лучше случайных. Авторы [22] предполагают, что такие плохие результаты для последнего семейства могут быть связаны с плохим выравниванием или с общим низким сходством среди этого типа протеаз.

Таблица 3. Результаты работы SDPsite в тестах [22]

	LacI			Субтилизин		
	перепредсказание	чувствительность	специфичность	перепредсказание	чувствительность	специфичность
A	0.007	0.06	0.75	0.024	0.17	0.57
B	0.05	0.07	0.5	0.043	0.128	0.43

Примечание. A – широкий набор значимых позиций (LacI – группы 2–5, субтилизин – все, для которых наблюдалось изменение активности). B – узкий набор значимых позиций (LacI – группа 5, субтилизин – остатки, участвующие в каталитической активности или распознавании субстрата).

В результате анализа с помощью SDPsite предсказывается не только относительная значимость каждой позиции, но и оптимальное их количество, поэтому его результаты выглядят как одна точка на графике кривой ROC. Значение чувствительности, уровня перепредсказания и чувствительности для обоих семейств показаны в табл. 3. Во всех четырех случаях SDPsite попадает в левый нижний угол графика кривой ROC, по крайней мере, не ниже остальных представленных методов. SDPsite показывает довольно хорошую специфичность (отношение *TP* к общему количеству предсказанных позиций). Положение в левом нижнем углу означает, что по этим тестам SDPsite имеет достаточно низкую чувствительность. Это может быть связано с тем, что важными в тестах считались достаточно большие наборы позиций, не все из которых прямо задействованы в функции. Так, для LacI важными считаются, в частности, все консервативные позиции, отвечающие за стабилизацию общей пространственной структуры белка. Следует отметить, что большинство методов, рассмотренных в [22], при хо-

Таблица 4. Тестовая выборка из базы данных CDD

Имя домена	Длина выравнивания	Количество последовательностей	Идентификатор PDB	Цепь	Имя домена	Длина выравнивания	Количество последовательностей	Идентификатор PDB	Цепь
35EXOc	112	55	2KZM	A	LIGANc	253	44	1DGS	A
53EXOc	194	56	1EXN	B	LMWPc	94	72	1D1P	B
ACTIN	296	45	1NM1	A	MADS	59	91	1MNM	B
ADF	108	51	1COF	A	MYSc	472	43	2MYS	A
aklPPc	301	29	1ELZ	A	PI3Kc	299	21	1E8X	A
Aminopeptidase	59	18	1B65	A	PIPKc	264	16	1BO1	A
AP2	59	23	1GCC	A	PLCc	184	14	1GYM	A
AP2Ec	188	40	1QUM	A	PNPsynthase	230	18	1HO4	A
Arfaptin	194	9	1I4D	A	POLXc	294	10	2BPF	A
BPI	123	31	1BP1	A	PP2Ac	235	19	1AUI	A
C2	72	100	1DQV	A	PP2Cc	158	100	1A6Q	A
CAP_ED	83	100	1RGS	A	PRCH	224	11	1PRC	H
CASc	197	52	1CP3	A	PROF	107	26	1D1J	D
CBM9	144	19	1I82	A	PTB	90	51	2NMB	A
CH	82	53	1AOA	A	PTPc	180	99	2SHP	A
DED	62	32	1A1Z	A	PTS_IIA_fru	107	99	1A6J	B
DEXDc	82	100	1D9X	A	PTS_IIA_lac	97	27	1E2A	A
DSPc	112	51	1VHR	A	PTS_IIA_man	97	43	1PDO	A
DSRM	52	100	1DI2	A	PTS_IIB_glc	74	88	1IBA	A
ENDO3c	115	100	1MUY	A	RA	74	48	1EF5	A
eu-GS	442	10	2HGS	A	RhoGAP	138	75	1AM4	A
fer2	60	100	1B9R	A	S4	51	100	1DM9	B
FGF	107	31	1QQK	A	SAM	53	99	1B0X	A
FH	59	48	1E17	A	Sec7	165	34	1PBV	A
G-alpha	302	63	1AZT	B	SEC14	123	100	1AUA	A
GMPK	93	57	1GKY	A	SERPIN	239	91	1OVA	A
GYF	55	21	1GYF	A	SH2	54	100	1AYA	A
H15	79	70	1HST	A	SNc	91	30	2SNS	A
HDc	85	100	1F0J	A	TBOX	174	32	1XBR	A
HECTc	313	48	1C4Z	A	TNF	96	33	1A8M	A
HELICc	104	100	1D2M	A	Topo6_Spo	245	25	1D3Y	B
HPT	87	71	1QSP	A	UBCc	129	70	2UCZ	A
HTH_ASER	66	100	1SMT	B	vWFA	83	100	1DZI	A
KISc	224	52	3KAR	A	XPG	249	34	1A76	A

рошей чувствительности показывают также и высокий уровень перепредсказания. SDPsite, напротив, специально спроектирован так, чтобы давать, по возможности, наименьшее перепредсказание.

Применение SDPsite к выборке из базы данных CDD

При анализе базы данных CDD мы рассмотрели тот же набор доменов, который рассмотрен в работе [26]. Эти домены имеют одну или несколько "особенностей", а соответствующие выравнивания содержат хотя бы один белок, для которого известна пространственная структура. После того как были удалены выравнивания, длина которых была меньше 50 аминокислотных остатков, или структура дерева для которых не позволяла выделить, по меньшей мере, две группы из трех или более последовательностей, осталось 68 доменов (табл. 4). Только позиции, помеченные од-

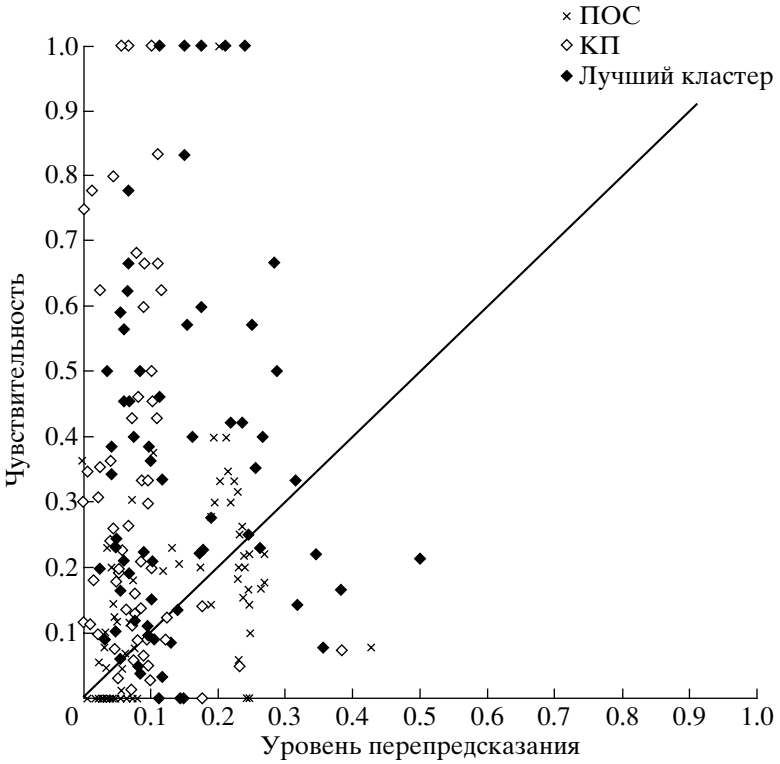

Рис. 5. Отношение чувствительности к перепредсказанию для ПОС, КП и лучшего кластера из рассмотренных доменов в CDD. Диагональ соответствует случайному выбору позиций.

ной из "особенностей", считались функционально значимыми. Это дает нижнюю границу оценки качества для метода, поскольку некоторые остатки, не помеченные "особенностью", могут также быть важны, а среди "особенностей" есть такие, которые не удовлетворяют интуитивному определению функционального сайта, например, сайты модификации аминокислотных остатков (фосфорилирования, гликозилирования и т.п.). Тем самым, настоящая чувствительность метода не ниже, а перепредсказание не выше, чем оцененная таким образом.

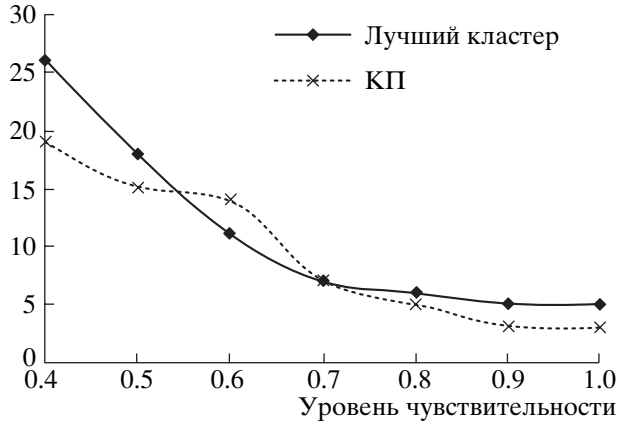

Рис. 6. Количество доменов CDD, для которых SDPsite предсказывает покрытие больше заданного уровня чувствительности.

Отношение чувствительности $TP/(TP + FN)$ (ось ординат) к перепредсказанию $FP/(FP + TN)$ (ось абсцисс) показано на рис. 5. Видно, что, несмотря на указанные выше дефекты такой оценки, для КП и кластеров это отношение, в среднем, лучше, чем для случайного выбора позиций (диагональ). Это не совсем очевидно для ПОС. Возможно, это связано с тем, что не все рассмотренные семейства на самом деле содержат группы различной специфичности, или с тем, что большинство аннотированных "особенностей" по смыслу должно быть консервативно во всем семействе. КП показывают достаточно хорошее отношение чувствительности к перепредсказанию (большая часть точек в верхнем треугольнике), однако кластеры, в среднем, показывают самую лучшую чувствительность. На рис. 6 показано количество семейств, для которых предсказание имеет чувствительность больше определенного порога. Видно, что кластеры опережают КП практически на всем диапазоне интересных предсказаний (чувствительность больше или равна 0.4). При этом уровень перепредсказания ни в одном случае не превосходит 0.3. Средняя чувствительность для кластеров равна 0.35, и для 5 семейств чувствительность равна 1, а для КП средняя чувствительность 0.30, и количество семейств с чувствительностью 1 равно 3.

ВЫВОДЫ

В настоящей работе представлен новый метод поиска функционального сайта SDPsite. Этот метод сочетает в себе многие черты ранее разработанных методов для поиска функционального сайта, такие как поиск консервативных позиций, выделение лучшего кластера на структуре белка. Отличительной особенностью данного метода является предсказание детерминант специфичности (ПОС) и связанное с ним автоматическое разделение выравнивания на группы специфичности.

Мы сравнили этот метод, во-первых, с экспериментальными данными (на примере LacI), во-вторых, с другими аналогичными методами (на примере LacI и субтилизина), а также провели массовый тест на большом количестве семейств из базы данных CDD.

Основная сложность при оценке методов для предсказания функциональных сайтов – отсутствие надежных контролей. Если в случае семейства LacI мы имеем практически полные данные о влиянии мутаций разных остатков на функцию белка, для других семейств это не так. В настоящей работе принято решение считать все не описанные в исходных данных остатки не значимыми для функции, а также не разделять различные типы функциональных остатков, что может существенно занижать оценку качества предсказания. Несмотря на это, результаты предсказаний с помощью SDPsite достаточно хорошо согласуются с контрольными данными.

Анализируя предсказания, полученные для семейств LacI и субтилизин-подобных протеаз, можно сделать вывод, что в случаях, когда в семействе имеются выраженные группы специфичности (как для LacI), ПОС предсказываются хорошо, и они играют первостепенную роль при выделении лучшего кластера. Соответственно, предсказанный функциональный сайт оказывается в области специфического взаимодействия. В случаях, когда группы специфичности выражены слабо (как для протеаз), основную роль при формировании лучшего кластера играют КП.

При сравнении с другими методами оказывается, что по соотношению чувствительности к перепредсказанию SDPsite работает на уровне, а возможно, и точнее лучших методов. Однако при этом он демонстрирует достаточно низкую чувствительность. Частично это может быть связано с идеологией SDPsite: довольно большое количество предсказанных на первом этапе ПОС и КП отметается при формировании наилучшего кластера, с тем чтобы минимизировать перепредсказание. Однако возможно и другое объяснение: при исследовании влияния мутаций на функцию важными признано большое количество позиций, напрямую на функцию не влияющих. Косвенно это подтверждает тот факт, что при сужении класса важных позиций для LacI до незаменимых и участвующих в непосредственных контактах с эффектором или ДНК доля предсказанных позиций вырастает до трети.

При анализе данных, полученных для доменов из CDD, обращает на себя внимание достаточно большое количество результатов на уровне случайного шума (левая нижняя четверть). Особенно плохие результаты дает рассмотрение только ПОС. Это может быть связано с тем, что многие рассмотренные выравнивания состоят из небольшого количества последовательностей и не содержат белков разной специфичности. В этом случае предсказание ПОС не имеет смысла. Низкие результаты для КП и кластеров могут объясняться свойствами некоторых "особенностей": например, сайты фосфорилирования довольно плохо консервативны в родственных белках. В случае, когда это единственная аннотированная "особенность", SDPsite, вероятнее всего, найдет лучший кластер в какой-то другой области белка, что приведет к очень плохому предсказанию (чувствительность = 0). С другой стороны, достаточно большое количество предсказаний имеет чувствительность выше 0.4 и перепредсказание ниже 0.3, что можно считать хорошим результатом.

Целью структурной геномики является расшифровка и функциональное описание возможно большего количества белков из разнообразных организмов. Поскольку для расшифровки структуры часто выбираются белки из плохо изученных семейств, не имеющие близких гомологов с известными структурами, функциональная аннотация их имеющимися методами (поиск похожих хорошо изученных последовательностей или структур) затруднена. SDPsite, на наш взгляд, может быть успешно применен для поиска функциональных сайтов в таких структурах и, как следствие, полезен для их аннотации.

Авторы благодарны проф. А.А. Миронову за ценные замечания, высказанные в ходе работы над проектом, и проф. А.В. Финкельштейну за критическое прочтение рукописи и ценные комментарии.

Работа поддержана грантами Российского фонда фундаментальных исследований (04-04-49438), Howard Hughes Medical Institute (55005610), INTAS (05-1000008-8028, 04-83-3704) и программой Российской академии наук "Молекулярная и клеточная биология".

СПИСОК ЛИТЕРАТУРЫ

1. Liolios K., Tavernarakis N., Hugenholtz P., Kyprides N.C. 2006. The Genome OnLine Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res*. **34**, D332–D334.

2. Chandonia J.-M., Brenner S.E. 2006. The impact of structural genomics: expectations and outcomes. *Science*. **311**, 347–351.
3. Russell R.B., Alber F., Aloy P., Davis F.P., Korkin D., Pichaud M., Topf M., Sali A. 2004. A structural perspective on protein-protein interactions. *Curr. Opin. Struct. Biol*. **14**, 313–324.
4. Glaser F., Pupko T., Paz I., Bell R.E., Bechor-Shental D., Martz E., Ben-Tal N. 2003. ConSurf: Indentification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*. **19**, 163–164.
5. Bell R.E., Ben-Tal N. 2003. *In silico* identification of functional protein interfaces. *Comparative and Functional Genomics*. **4**, 420–423.
6. Aloy P., Querol E., Aviles F.X., Sternberg M.J.E. 2001. Automated structure-based prediction of functional sites in proteins: application to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol*. **331**, 395–408.
7. Ma B., Elkayam T., Wolfson H., Nussinov R. 2003. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc. Natl. Acad. Sci. USA*. **100**, 5772–5777.
8. Landgraf R., Xenarios I., Eisenberg D. 2001. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol*. **307**, 1487–1502.
9. del Sol Mesa A., Pazos F., Valencia A. 2003. Automatic methods for predicting functionally important residues. *J. Mol. Biol*. **326**, 1289–1302.
10. Lichtarge O., Bourne H.R., Cohen F.E. 1996. An evolutionary trace method defined binding surfaces common to protein families. *J. Mol. Biol*. **257**, 342–358.
11. Yao H., Kristensen D.M., Mihalek I., Sowa M.E., Shaw C., Kimmel M., Karvaki L., Lichtarge O. 2003. An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J. Mol. Biol*. **326**, 255–261.
12. Hannenhalli S.S., Russell R.B. 2000. Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol*. **303**, 61–76.
13. 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.
14. 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.
15. Vinogradov D.V., Mironov A.A. 2002. SiteProb: yet another algorithm to find regulatory signals in nucleotide sequences. In: *Proc. 3rd Int. Conf. on Bioinformatics of Genome Regulation and Structure BGRS'2002*. Novosibirsk, Russia, pp. 28–30.
16. Valdar W.S.J. 2002. Scoring residue conservation. *Proteins*. **48**, 227–241.
17. Casari G., Sander C., Valencia A. 1995. A method to predict functional residues in proteins. *Nature Struct. Biol*. **2**, 171–178.
18. Henikoff S., Henikoff J. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci*. **89**, 10915–10919.
19. Mirkin B., Muchnik I. 2002. Layered clusters of tightness set functions. *Appl. Math. Lett*. **15**, 147–151.
20. Laikova O.N. 2003. The LacI family of bacterial transcriptional regulators and the evolution of sugar utilization regulons in bacteria. In: *Proc. 1st Int. Moscow Conference on Computational Molecular Biology MCCMB'03*. Moscow, Russia, pp. 121–122.
21. Suckow J., Markiewicz P., Kleina L.G., Miller J., Kisters-Woike B., Muller-Hill B. 1996. Genetic studies of the Lac Repressor XV: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure. *J. Mol. Biol*. **261**, 509–523.
22. Soyer O.S., Goldstein R.A. 2004. Predicting functional sites in proteins: site-specific evolutionary models and their application to neurotransmitter transporters. *J. Mol. Biol*. **339**, 227–242.
23. Ko J., Murga L.F., Andre P., Yang H., Ondrechen M.J., Williams R.J., Agunwamba A., Budil D.E. 2005. Statistical criteria for the identification of protein active sites using theoretical microscopic titration curves. *Proteins*. **59**, 183–195.
24. Bryan P.N. 2000. Protein engineering of subtilisin. *Biochim. Biophys. Acta*. **1543**, 203–222.
25. Perriere G., Duret L., Gouy M. 2000. HOBACGEN: database system for comparative genomics in bacteria. *Genome Res*. **10**, 379–385.
26. Panchenko A.R., Kondrashov F., Bryant S. 2004. Prediction of functional sites by analysis of sequence and structure conservation. *Protein Sci*. **13**, 884–892.