Математическая биология и биоинформатика. 2013. Т. 8. № 2. С. 407–418.
URL: http://www.matbio.org/2013/Igolkina_8_407.pdf

========== **БИОИНФОРМАТИКА** ==========

УДК: 577.151.43 + 519.171.2 + 577.151.45

Метод поиска субстратспецифичных областей ферментов класса целлюлаз на основании их первичной и третичной структур

©2013 Иголкина А.А.[*1], Андронов Е.Е.[2], Порозов Ю.Б.[***3]**

[1] *Санкт-Петербургский государственный политехнический университет, Санкт-Петербург, 195251, Россия*
[2] *Государственное научное учреждение Всероссийский научно-исследовательский институт сельскохозяйственной микробиологии Российской академии сельскохозяйственных наук, Пушкин, Санкт-Петербург, 196608,Россия*
[3] *Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики (НИУ ИТМО), Санкт-Петербург, 197101, Россия*

Аннотация. В природе деградация растительной биомассы, основными компонентами которой являются целлюлоза и ее производные, выполняется микроорганизмами, синтезирующими ферменты класса целлюлаз (ФКЦ). Волокна клеточных стенок состоят из комплекса полисахаридов, расщепление которых осуществляют сложноустроенные ФКЦ. Они содержат один или несколько каталитических доменов, расщепляющих полисахариды, а также домены связывания с субстратами (carbohydrate binding module, CBM). Способность ферментов расщеплять полисахариды обуславливается конфигурацией каталитического центра в каталитическом домене (модуле), в частности наличием в нем комплементарного субстрату сайта связывания. В данной работе был разработан и апробирован комбинированный подход к выявлению сайтов связывания ФКЦ с растительными полисахаридными субстратами. Подход применялся к 90 белкам с выявленной целлюлазной активностью на основе данных, полученных M. Hess и др. В результате были получены две консенсусные последовательности областей связывания ФКЦ с полисахаридными субстратами Carboxymethyl Cellulose (CMC) и Xylan. На основании разработанного подхода было создано программное обеспечение, реализующее основные этапы поиска и обнаружения сайтов связывания. Созданный метод и программное обеспечение может быть применено при анализе больших групп белков, обладающих разнородной субстратспецифичностью с целью обнаружения функциональных участков.

Ключевые слова: *целлюлаза, граф, алгоритм поиска сайтов, структура белка, сайт связывания, биотопливо.*

ВВЕДЕНИЕ

В настоящее время разработано несколько вычислительных подходов к оценке макромолекулярных взаимодействий, в частности, широко используются разновидности докинга [1]. Этот метод молекулярного моделирования включает в себя расчеты на больших массивах входных данных, что требует значительных затрат

[*] igolkinaanna11@gmail.com
[**] eeandr@gmail.com
[***] porozov@ifc.cnr.it

времени и ресурсов. Поэтому попытки сузить область поиска межмолекулярных контактов до выполнения протоколов докинга представляются важным этапом оптимизации расчетов.

В работе описан подход для фильтрации субстратспецифичных сайтов связывания белков, состоящий из нескольких шагов. Его применение в качестве предварительного фильтра при виртуальном скрининге может позволить значительно ускорить процесс поиска областей связывания.

Целями данной работы являлись разработки метода поиска консенсусной последовательности сайтов связывания белков на основании анализа первичных структур и субстратспецифичностей ФКЦ. Применение метода поиска к имеющимся данным о ФКЦ [2] дает возможность сконструировать консенсусную последовательность предположительных сайтов связывания.

В настоящее время полный набор ферментов, необходимый для расщепления энергоемких злаковых растений и лигноцеллюлозной биомассы (целлюлоза, гемицеллюлоза и лигнин), неизвестен. Поиск и исследования таких ФКЦ направлены на повышение эффективности технологий получения с их помощью биотоплива [3–5]. С этой целью создаются штаммы микроорганизмов, трансформированные векторами, несущими гены различных ФКЦ [6, 7]. Такие модификации дают возможность микроорганизмам продуцировать сбалансированный комплекс ФКЦ, необходимый для расщепления лигноцеллюлозы. Однако известные на данный момент штаммы не синтезируют оптимальный комплекс ФКЦ, разлагающий биомассу полностью. В частности, это связано со способностью отдельных ферментов взаимодействовать только с определенными компонентами биомассы. Поэтому поиск генов и аминокислотных последовательностей натуральных или синтетических ФКЦ, учитывающий особенности структурной субстратспечифичности, ведет к увеличению производительности при получении биотоплива.

В природе расщепление растительных полисахаридов осуществляется несколькими классами микроорганизмов [2, 8], в частности, бактериями, живущими в пищеварительной системе травоядных. M. Hess и др. [2] было проведено исследование бактерий из рубца коровы, прилипших к волокнам помещенного в него растительного субстрата (просо). После секвенирования метагеномной ДНК бактерий и сборки геномов *de novo* были выбраны 90 генов белков, принадлежащих к одному из семи семейств гидролаз (GH3, GH5, GH8, GH9, GH10, GH26, GH48), часть из которых имела характерные домены связывания с карбогидратами (CBM_6, CBM_4_9). После тестирования продуктов экспрессии этих генов на ферментативную активность были получены данные о гидролитической активности каждого из 90 белков-целлюлаз на шести субстратах (CMC, ксилан, просо, мискантус, IL-Avicel, лихенин). Выборочно результаты эксперимента представлены в табл. 1.

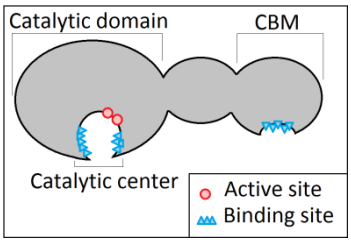

Рис. 1. Структура ФКЦ.

Использованные в тестировании субстраты имеют линейное полимерное строение и встречаются в качестве компонент растительной биомассы [3]. Для удержания и расщепления молекулы полисахарида структура ФКЦ должна обладать характерным строением каталитического домена: туннель или углубление (бороздка, «карман»), в котором расположены активный сайт и сайт связывания фермента с субстратом (Рис. 1) [4, 9, 10]. Разработанный в настоящей работе поэтапный метод поиска сайтов связывания учитывает как оговоренные особенности третичной структуры белков, так и различные варианты модели сайтов: строгая и нестрогая субстратспецифичность (специфичность одному или некоторой группе субстратов).

АНАЛИЗ И МЕТОДИКА ОБРАБОТКИ ИСХОДНЫХ ДАННЫХ

Анализ исходных данных

Результаты работы Matthias Hess и др. (Таблица 1) показали, что ФКЦ, принадлежащие одному доменному семейству гидролаз, имеют неодинаковую целлюлазную активность на наборе субстратов. Следовательно, активность ФКЦ зависит от конфигурации их каталитических центров, а не от принадлежности семейству гидролаз. Таким образом, одним из возможных факторов, определяющих неоднозначное поведения ФКЦ одного семейства на наборе субстратов, может быть делеция активных сайтов в каталитическом центре. Однако по результатам поиска в Pfam [11] оказалось, что почти все (71 из 90) белки имеют соответствующие их доменам активные сайты, при этом некоторые белки расщепляют субстраты, не имея предсказанных активных сайтов. Другим возможным фактором, влияющим на неоднозначное поведение ФКЦ, является недоступность активного сайта на молекулярной поверхности каталитического центра. В этом случае два известных активных сайта семейства GH5 могли бы дать не более четырех различных вариантов специфического связывания, что не подтвердилось экспериментом (табл. 1). Таким образом, был сделан вывод, что активность ФКЦ определяется наличием или отсутствием областей, необходимых для комплементарного связывания с субстратом.

Область контакта ФКЦ с субстратом состоит из набора сайтов связывания (одного или нескольких отдельных коротких отрезков последовательности белка) [9]. Анализ приведенных в работе [2] данных позволил сделать вывод о том, что области связывания для всех шести субстратов (CMC agar, Xylan, IL-Switchgrass, IL-Miscanthus, IL-Avicel, Lichenan) обладают присущими только им свойствами. Поэтому поиск областей связывания заключался в нахождении для каждого субстрата набора участков длины k аминокислот (k-меров), не имеющего общих k-меров с другими наборами.

Таблица 1. Свойства представителей семейства GH5 по результатам [2]

Gene ID	CAZy family	Hydrolytic Activity[3)]						Domain organization[4)]
		C	X	S	M	A	L	
458803_07710	GH5							
0_06533	GH5					✓		
3271578_13460	GH5						✓	
2698429_129360	GH5	✓						
1696514_56150	GH5	✓				✓	✓	
2395619_81340	GH5	✓			✓		✓	
3671981_90060	GH5	✓		✓	✓	✓		
3932955_213080	GH5	✓			✓	✓	✓	
3953955_160820	GH5	✓		✓	✓	✓	✓	
558318_19410	GH5	✓	✓	✓	✓	✓		

Методика обработки исходных данных

Предположительной областью связывания с любым из шести субстратов являлся k-мер, найденный во всех последовательностях белков, активных на этом субстрате, и не обнаруженный ни в одном из остальных белков. Поэтому простейший поиск областей

связывания может выглядеть следующим образом. Пусть продукты генов G1 и G2 активны на двух субстратах (C и X), а продукт гена G3 активен на одном субстрате (X), тогда общие *k*-меры для последовательностей генов G1 и G2, не найденные в последовательности гена G3, можно считать кандидатами на область связывания с субстратом C. Чем больше последовательности G1 и G2 отличаются от G3, тем больше будет найдено кандидатов на область связывания. Таким образом, для уменьшения “шума”, затрудняющего поиск областей связывания ФКЦ, предпочтительным являлось разделение генов ФКЦ на группы, представители которых мало отличались бы по аминокислотной последовательности, но имели неодинаковую активность на субстратах. Затем поиск *k*-меров проводился в каждой группе.

Таким образом, поиск был разбит на четыре этапа:

1. Группировка гидролаз на основании парного и множественного глобального выравнивания.
2. Группировка гидролаз, основанная на результатах построения филогенетических деревьев.
3. Алгоритм поиска областей связывания в группе.
4. Фильтрация с использованием моделирования структуры белков.

Группировка гидролаз на основании парного и множественного глобального выравнивания

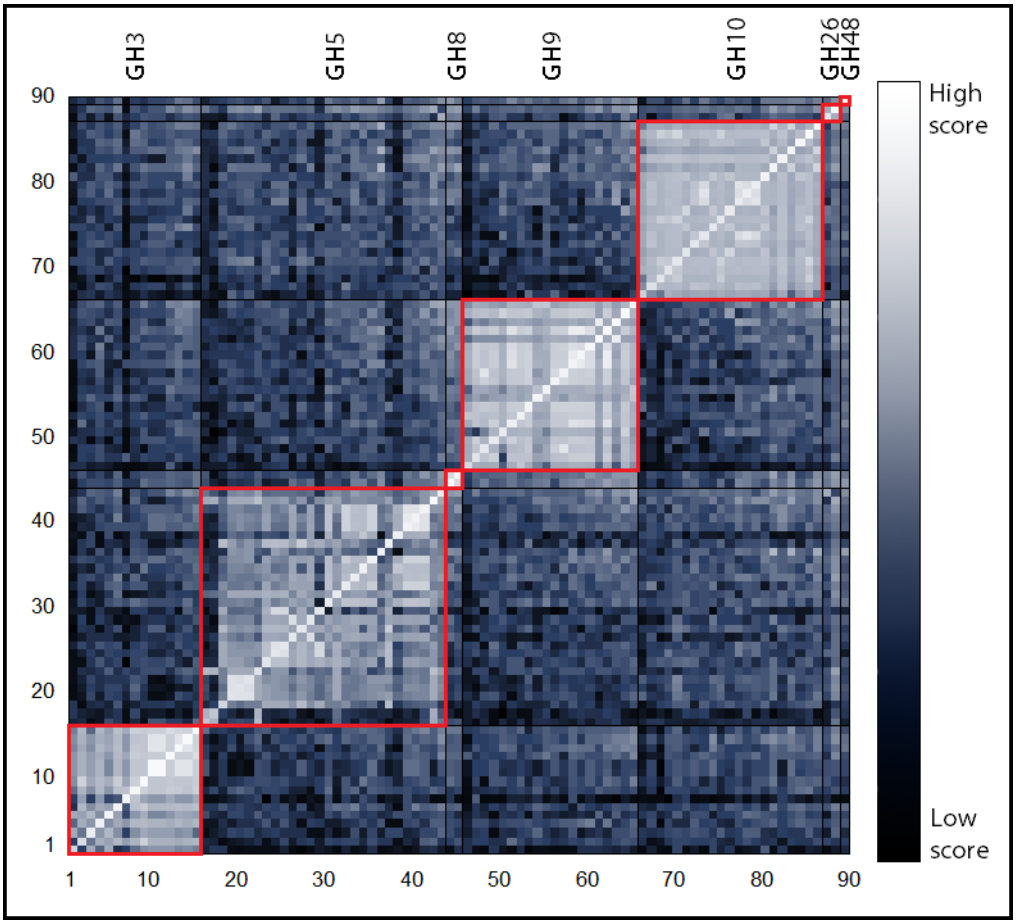

Рис. 2. Попарное сравнение 90 белков. Области для пар белков одного семейства (внутри красных квадратов) являются областями наибольшего сходства.

Среди имеющихся 90 представителей исследуемых гидролазных семейств было проведено попарное глобальное выравнивание всех белков с использованием

Needleall [12]. Результаты выравниваний подтвердили наличие характерных связей межу белками внутри одного семейства (Рис. 2). В то же время сходства между семействами гидролаз (GH3, GH5, GH8, GH9, GH10, GH26, GH48) не было обнаружено. Такое отсутствие сходства между последовательностями белков разных семейств означает, что области связывания для субстратов нужно искать в каждом семействе отдельно.

Дальнейший анализ каждого семейства белков проводился при помощи серии множественных глобальных выравниваний (ClustalW2 [13]) с варьированием количества белков в наборах для MSA и параметров выравнивания. Результат выравниваний не дал достаточно оснований для разделения белков в семействе на значимые группы. Это может быть связано с различной доменной структурой, внутридоменной изменчивостью и необходимостью применения «плавающих» параметров MSA.

Группировка гидролаз, основанная на результатах построения филогенетических деревьев

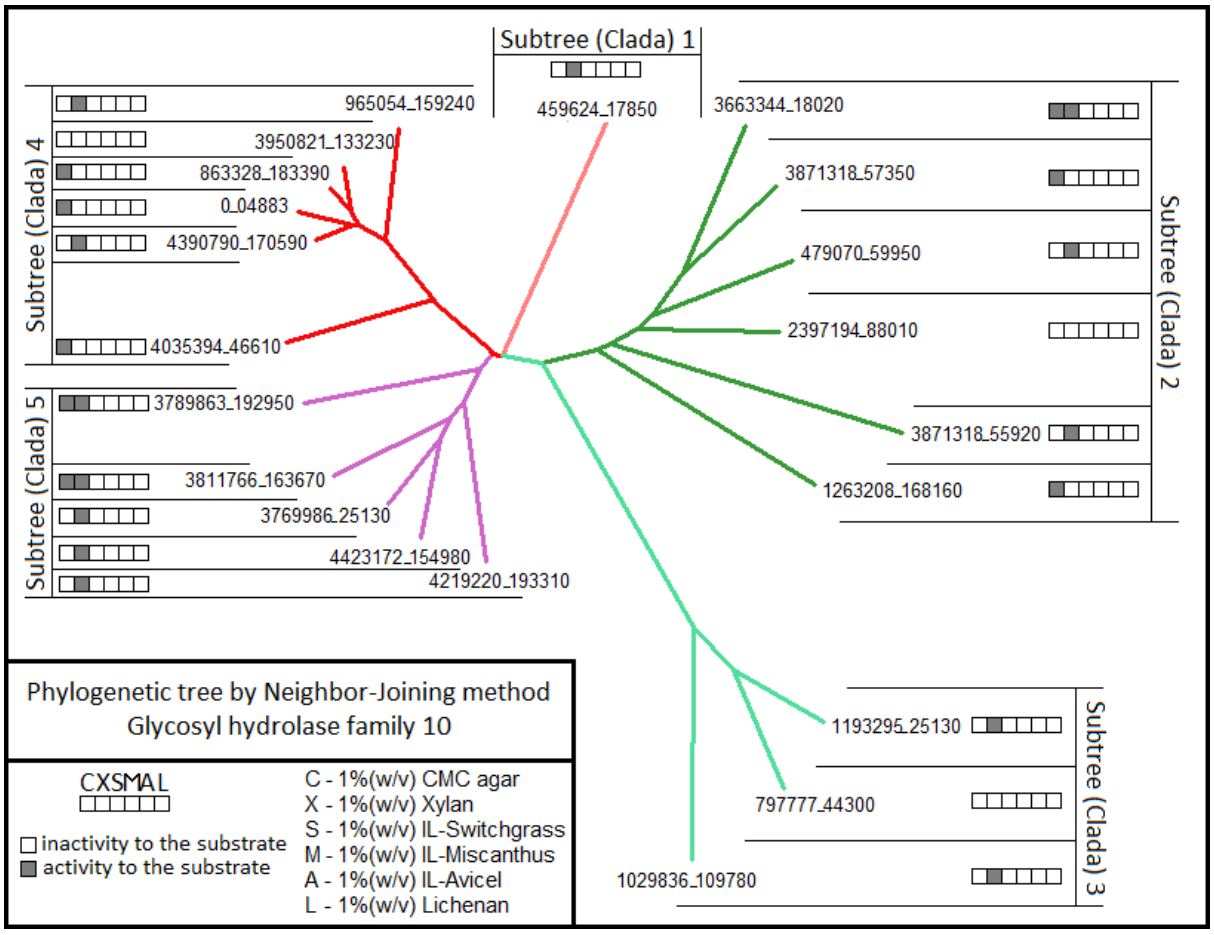

Рис. 3. Филогенетическое дерево, построенное для представителей семейства GH10. Для каждого представителя указана субстратспецифичность, а также выделены поддеревья (клады), в рамках которых проводится алгоритм поиска сайтов связывания. На листьях – Gene ID соответствующих генов.

Результаты множественного глобального выравнивания показали, что в белковых последовательностях достаточно много различий, которые мешают выявлять значимые группы (группы белков, отличающихся по активности к субстрату и обладающих минимальными различиями в аминокислотной последовательности). Поэтому был проведен анализ доменов или их частей, имеющихся в каждом белке, отвечающих за

функциональность (ФКЦ). Из каждого белка из [2] был выделен функциональный участок (домен или его часть) и проведено парное выравнивание этого участка с консенсусной последовательностью домена для соответствующего семейства с помощью Pfam [11]. Затем полученные парные выравнивания были объединены в одно множественное выравнивание для всех белков каждого семейства. Составные множественные выравнивания были получены благодаря наличию реперной доменной последовательности для каждого из 7 гидролазных семейств в Pfam.

После получения MSA функциональных участков белков каждого семейства были построены филогенетические деревья при помощи алгоритма Neighbor-Joining (пакет MEGA [14]). Для каждого дерева были определены несколько крупных клад, в которых проводился поиск областей связывания (Рис. 3). При делении дерева на клады принималось во внимание то, что внутри одной группы, с одной стороны, должны находиться различные по субстратспецифичности белки (по возможности), но, с другой стороны, принадлежность к кладе обеспечивает родство последовательностей.

Алгоритм поиска областей связывания в группе

Одним из условий группировки и образования клад было присутствие в одной кладе белков как с различной, так и с одинаковой субстратспецифичностью. Соблюдение этого условия дало возможность в рамках клады образовать несколько подгрупп, содержащих ФКЦ с одинаковой активностью. На основании такой группировки для каждой клады был построен граф активности, узлам которого соответствовали подгруппы (Рис. 4). Каждому узлу графа соответствовала совокупность объектов (свойства признаки атрибуты): набор белковых последовательностей членов подгруппы, активность на субстратах, одинаковая для всех представителей подгруппы, и некоторый список *k*-меров. Дополнительно в граф добавлялся узел (нулевая вершина), не содержащий последовательности белков и списка *k*-меров, но обладающий гипотетической активностью на всех субстратах. Данные по активности 90 ФКЦ дали возможность разделить узлы в графе на семь уровней. Номер уровня означает количество субстратов, на которых наблюдается активность белков соответствующей узлу подгруппы (от 0 до 6). Ребро в графе соединяет узел V_1 с узлом V_2, если выполняются все следующие условия:

- Активность, соответствующая узлу V_2, включает в себя активность, соответствующую узлу V_1.
- Узел V_2 располагается на соседнем с V_1 уровне. Если на соседнем уровне с V_1 не оказалось узла, удовлетворяющего предыдущему условию, то в качестве V_2 принимается нулевая вершина.

Алгоритм поиск кандидатов на область связывания состоял из двух обходов построенного графа. Обходы производились последовательно по уровням и начинались на уровне 6 (от нулевой вершины), заканчиваясь на уровне 0.

При первом обходе графа в каждом узле формировались списки *k*-меров, общих для всех последовательностей узла. Если на новом шаге этого обхода узел оказывался связанным ребром с корнем (Рис. 5,*A*), то в узле случайным образом выбиралась одна из соответствующих ему последовательностей, и из неё извлекались все различные *k*-меры. *k*-мер добавлялся в список рассматриваемого узла, если он был найден во всех последовательностях узла. Если на новом шаге обхода графа узел оказывался не связанным с корнем, то *k*-меры выбирались из списков всех узлов, соединенных с рассматриваемым и имеющих больший уровень. Аналогичным образом, если такой *k*-мер был найден во всех последовательностях узла, то он добавлялся в список узла. Если же *k*-мер не был найден ни в одной последовательности узла, то он считался подозрительным на область связывания с субстратом, который находится в активности узла, из которого был получен *k*-мер, но не находится в активности, соответствующей

узлу. (Рис. 5,*B*) Таким образом, после первого обхода графа в каждом узле сформировался список *k*-меров, и для некоторых субстратов были определены кандидаты на сайты связывания.

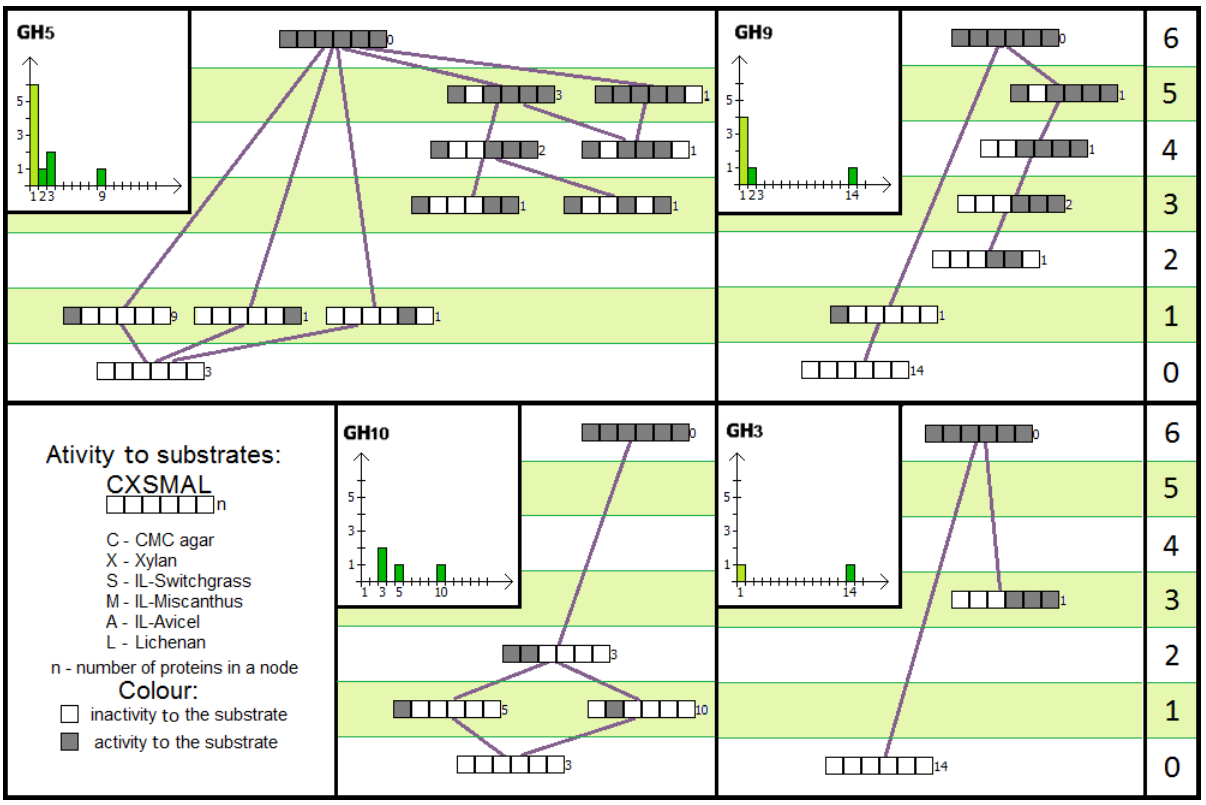

Рис. 4. Графы активностей, составленные из исследуемых представителей гидролазных семейств. Гистограммы распределения количества последовательностей в узлах.

Найденные после первого обхода графа *k*-меры могли ассоциироваться с областями связывания различных субстратов. Такой результат после первого обхода графа возникал из-за разветвленности графа и независимости формирования списков в узлах. В то же время для достоверного разделения *k*-меров по субстратам требовалось обнаружить соответствие один *k*-мер – один субстрат. Поэтому после первого обхода графа одинаковые *k* -меры, соответствующие разным субстратам, были исключены из дальнейшего рассмотрения (Рис. 5,*C*).

Как уже было замечено, при обходе графа *k* -меры могли быть получены из разных (независимых) путей от уровня 6 до уровня 0, потому был проведен второй обход графа, который проверял присутствие каждого найденного кандидата на область связывания во всех узлах, которые не участвовали в его формировании.

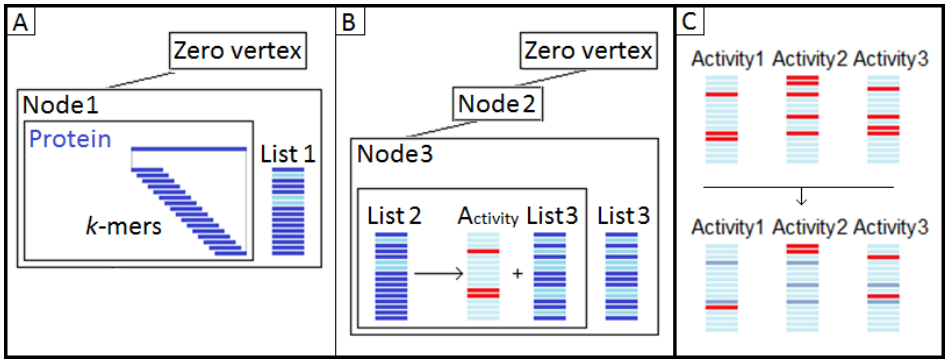

Рис. 5. Шаги алгоритма поиска областей связывания.

Фильтрация с использованием моделирования структуры белков

Была проведена дополнительная проверка найденных участков на трехмерной структуре белков. С помощью сервера ModWeb [15] для каждого из 90 белков были построены 3D-модели. Затем найденные *k*-меры были выделены на построенных структурах, и установлено их местоположение: внутри белка, на поверхности белка не в углублении, на поверхности белка в углублении. Участки, которые образовывали поверхность углублений, карманов или бороздок, принимались нами за достоверные области связывания. Характерные для каждого семейства активные сайты из PFam также были нанесены на построенные структуры (рис. 6).

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Разработанный метод поиска областей связывания был применен к данным об активности 90 белков семи гидролазных семейств (GH3, GH5, GH8, GH9, GH10, GH26, GH48). На этапе группировки белков по семействам оказалось, что среди исходных данных присутствует всего один белок семейства GH48, и по два белка – представителя семейств GH8 и GH26. Такое количество представителей в семействах было недостаточно для того, чтобы выявить характерные области связывания с субстратами. Кроме того, только у одного из пяти ФКЦ была выявлена активность хотя бы на одном из рассмотренных шести субстратов. Поэтому эти гены ФКЦ были исключены из рассмотрения.

На этапе группировки с помощью филогенетических деревьев (Рис. 3) оказалось, что разделение белков семейств на клады привело к образованию слишком маленьких групп, что было связано с недостаточным количеством первичных данных. Поэтому алгоритм поиска областей связывания применялся к графам, построенным для семейств GH3, GH5, GH9, GH10 в целом (Рис. 4).

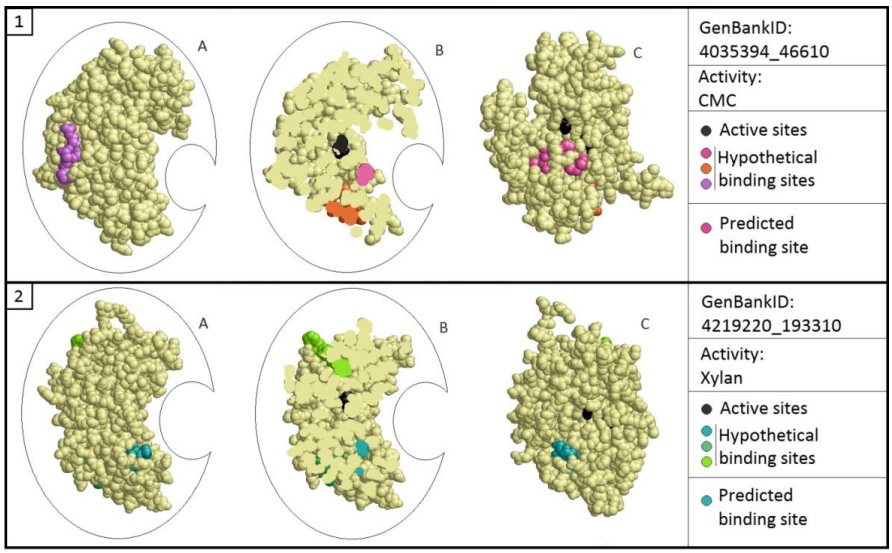

Рис. 6. Местоположение достоверных сайтов связывания (Predicted binding site) с субстратами CMC и Xylan на трехмерной структуре двух белков (1, 2) семейства GH10. Местоположение сайтов связывания с субстратами CMC и Xylan, найденных с помощью алгоритма поиска областей связывания (Hypothetical binding sites). *A*: вид на белок сбоку. *B*: вид на белок сбоку в сечении, на сечении видны располагающиеся в углублении каталитического центра активные сайты (Active sites) и достоверный сайт связывания. *C*: вид со стороны каталитического центра.

Большое количество узлов в графе, имеющих только один элемент, серьезно затрудняет правильный поиск областей связывания (*k*-меров), так как построенный алгоритм предполагает формирование в каждом узле списка общих характерных *k*-меров последовательностей узла. Таким образом, чем больше последовательностей

содержится в узле, тем *k*-меры, формирующие список, более специфичны. Напротив, если в узле находится только одна последовательность, то его список *k*-меров получается малоинформативным, что ведет к сильному зашумлению выходных данных. Поэтому, кроме графов активности семейств, были построены гистограммы распределения количества последовательностей в узлах (Рис. 4). Оказалось, что только граф семейства GH10 можно считать достаточно полным, поэтому дальнейшая работа проводилась только с ним.

Семейство GH10 содержит 21 ФКЦ с известной активностью на субстратах, при этом если представитель семейства проявил активность на субстрате, то это был субстрат CMC и/или Xylan. В этом семействе при помощи построенного алгоритма поиска областей связывания с параметром $k = 5$ были найдены три сайта связывания с субстратом CMC и три – с субстратом Xylan. Определение их местоположения на смоделированной структуре белков показало, что все найденные сайты связывания располагались на поверхности белков, но только один в каждой тройке был отмечен в углублении каталитического центра, содержащем активные сайты (Рис. 6). Опираясь на данные о выявленных областях связывания для каждого представителя семейства GH10 (Таблица 2), были получены консервативные последовательности для сайтов связывания ФКЦ с субстратами CMC и Xylan (Рис. 7).

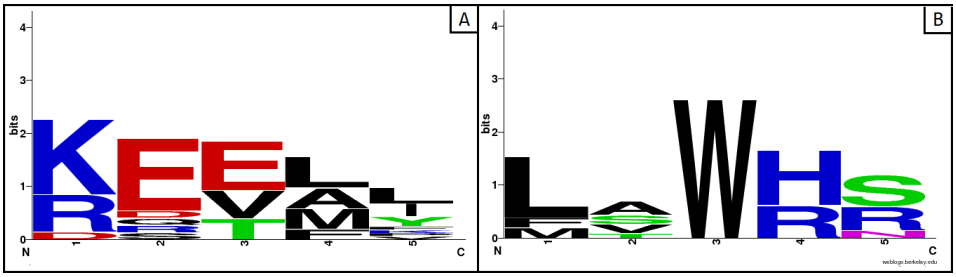

Рис. 7. Графическое представление консенсусов в последовательности сайтов связывания (Sequence Logo [16]). *A* – сайт связывания с CMC. *B* – сайт связывания с Xylan.

Однако, кроме сайтов, которые могут комплементарно связываться только с одним определенным субстратом, в каталитическом центре фермента могут располагаться сайты, способные формировать контакт с целым набором молекул. Кроме того, один фермент может иметь на поверхности каталитического центра несколько сайтов связывания, специфичных к одному субстрату (Рис. 2).

Таким образом, полный набор специфичных сайтов связывания ФКЦ с субстратами CMC и Xylan, возможно, не ограничивается найденными двумя. Исключение из рассмотрения остальных субстратспецифичных сайтов связывания (если такие имелись) могло произойти из-за способа построения графа активности – ребром соединялись два узла, находящиеся на соседних уровнях, если ни один из узлов не являлся нулевой вершиной. Если допустить, что ребром могут связываться узлы, находящиеся не только на соседних уровнях (в случае если среди них нет нулевой вершины), то это условие позволит добавить дополнительные ребра в граф активности, что повлечет за собой увеличение шагов алгоритма при обходе графа, а значит и увеличение количества *k*-меров на выходе алгоритма (рис. 8,*B*). Обнаруженные с помощью такого графа дополнительные участки, вероятно, могут расширить набор из найденных субстратспецифичных сайтов. Однако такие дополнительные *k*-меры должны будут пройти процедуру конкретизации специфичности, так как для них будет определена специфичность к одному из нескольких субстратов (рис. 8,*C*).

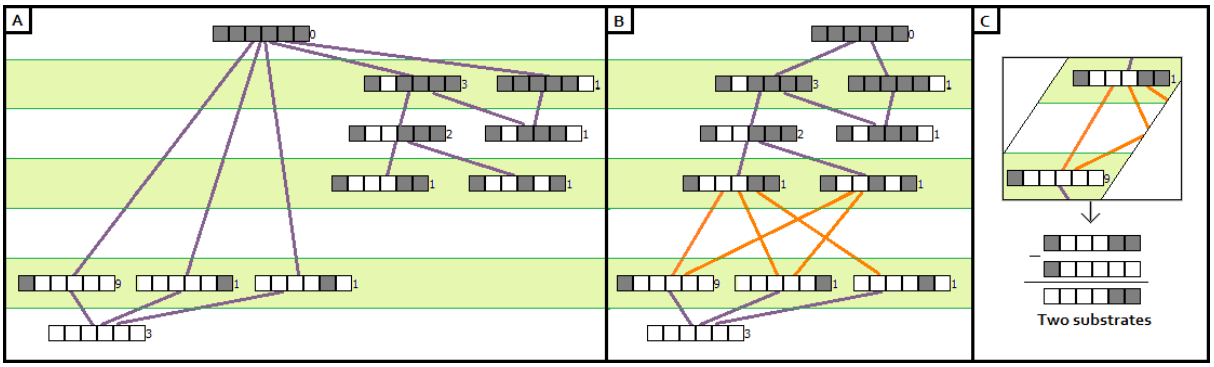

Рис. 8. Модификация графа активностей семейства для GH10 (описание см. в тексте). *A*: Граф активности, содержащий соединения узлов, находящихся только на соседних уровнях. *B*: Граф активности, допускающий соединение узлов, разница в номерах уровней которых может быть равной одному или двум. *C*: Участок графа активности. При прохождении алгоритма по выделенному ребру, отобранные *k*-меры будут кандидатами на область связывания не с определенным субстратом, а с одним из двух.

Таблица 2. Области связывания представителей семейства GH10 с субстратами CMC и Xylan.

Gene ID	Predicted binding sites to substrates		Hydrolityc Activity					
	CMC	Xylan	C	X	S	M	A	L
1263208_168160	Leu 102 - Asn 106		V					
4035394_46610	Leu 30 - Ser 34		V					
4423172_154980		Lys 27 - Tyr 31		V				
3789863_192950	Leu 89 - Ser 93	Lys 111 - Val 115	V	V				
4219220_193310		Lys 112 - Tyr 116		V				
3663344_18020	Leu 343 - Arg 347	Arg 148 - Leu 152	V	V				
965054_159240		Arg 172 - Ile 176		V				
0_04883	Leu 121 - Ser 125		V					
479070_59950		Arg 179 - Leu 183		V				
3769986_25130		Lys 160 - Tyr164		V				
3871318_57350	Met 378 - Arg 382		V					
863328_183390	Leu 104 - Ser 108		V					
4390790_170590		Lys 129 - Leu 133		V				
3950821_133230								
2397194_88010								
1193295_25130		Lys 359 - Leu 363		V				
459624_17850		Arg 183 - Ile 187		V				
1029836_109780		Pro 121 - Ile 125		V				
3871318_55920		Arg 143 - Leu 147		V				
3811766_163670	Leu 305 - Arg 309	Lys 152 - Pro 156	V	V				
797777_44300								

Разработанный метод поиска направлен на обнаружение сайтов, соответствующих только одному определенному субстрату, но с его помощью также можно искать и другие типы сайтов связывания. В предложенном алгоритме поиска областей связывания на определенном этапе (Рис. 5,*C*) исключались из рассмотрения *k*-меры, соответствующие сразу нескольким субстратам. Так как такие *k*-меры являлись общими для некоторых последовательностей ФКЦ, то можно предположить, что среди них находились области связывания, не обладающие избирательностью к определенным субстратам. Как было показано на рис. 7,*C*, добавление в граф активности новых ребер может привести к обнаружению сайтов, для которых субстратспецифичность не определяется явным образом. Таким образом, область

поиска сайтов связывания, не обладающих субстратспецифичностью, можно сузить, объединив два множества *k*-меров: активность которых проявлялась на нескольких субстратах и найденных на модифицированном графе, активность которых была неопределенной.

Анализ белковых последовательностей с помощью Pfam подтвердил, что ФКЦ имеют сложную доменную организацию, и кроме домена, определяющего семейство, фермент может содержать несколько доменов CBM. Так, три представителя семейства GH10 имеют в своей структуре домен CBM_6. Было замечено, что именно эти три ФКЦ образовали кладу 3 филогенетического дерева (Рис. 3). Таким образом, наличие большего количества исходных данных не только позволило бы провести алгоритм поиска в рамках клады филогенетического дерева, но также дало бы возможность найти области связывания CBM доменов с субстратом, в данном случае – области домена CBM_6.

Разработанный метод обнаружил субстратспецифичные сайты связывания даже в случае, когда исходная выборка ФКЦ мала. Данный подход можно модифицировать предложенными вариантами и применять к репрезентативным выборкам объектов, представляющих собой последовательности из некоторого алфавита, для выделения специфичных участков.

Работа поддержана Министерством образования и науки, ГК № 16.552.11.7085, и Соглашением № 14.B37.21.0562 в рамках федеральной целевой программы «Научные и научно-педагогические кадры инновационной России» на 2009–2013 годы.

СПИСОК ЛИТЕРАТУРЫ

1. Woodcock S., Henrissat B., Sugiyama J. Docking of Congo Red to the Surface of Crystalline Cellulose Using Molecular Mechanics. *Biopolymers.* 1995. V. 36. P. 201–210.
2. Hess M., Sczyrba A., Egan R., Kim T.W., Chokhawala H., Schroth G., Luo S., Clark D.S., Chen F., Zhang T. et al. Metagenomic Discovery of Biomass-Degrading Genes and Genomes from Cow Rumen. *Science.* 2011. V. 331. P. 463–467.
3. Rubin E.M. Genomics of cellulosic biofuels. *Nature*. 2008. V. 454. P. 841–845.
4. Himmel M.E., Ding S.Y., Johnson D.K., Adney W.S., Nimlos M.R., Brady J.W., Foust T.D. Biomass Recalcitrance:Engineering Plants and enzymesfor Biofuels Production. *Science*. 2007. V. 315. P. 804–807.
5. Samuel R., Pu Y., Foston M., Ragauskas A.J. Solid-state NMR characterization of switchgrass cellulose after dilute acid pretreatment. *Biofuels*. 2010. V. 1. P. 85–90.
6. Gilkes N.R., Kilburn D.G., Langsford M.L., Miller Jr R.C., Wakarchuk W.W., Warren R.A.J., Whittle D.J., Wong W.K.R. Isolation and Characterization of *Escherichia coli* Clones Expressing Cellulase Genes from Cellulomonas Jimi. *Journal of General Microbiology*. 1984. V. 130. P. 1377–1384.
7. Gilkes N.R., Langsford M.L., Kilburn D.G., Miller R.C.Jr., Warren R.A. Mode of Action and Substrate Specificities of Cellulases from Cloned Bacterial Genes. *The Journal of Biological Chemistry.* 1984. V. 259. № 16. P. 10455–10459.
8. Abu Bakar N.K., Abd-Aziz S., Hassan M.A., Ghazali F.M. Isolation and Selection of Appropriate Cellulolytic Mixed Microbial Cultures for Cellulases Production from OilPalm Empty Fruit Bunch. *Biotechnology*. 2010. V. 9. P. 73–78.
9. Koivula A., Reinikainen T., Ruohonen L., Valkeajärvi A., Claeyssens M., Teleman O., Kleywegt G.J., Szardenings M., Rouvinen J., Jones T.A., Teeri T.T. The active site of *Trichoderma reesei* cellobiohydrolase II: the role of tyrosine 169. *Protein Engineering.* 1996. V. 9. № 8. P. 691–699.
10. Knowless J., Lehtovaara P., Teeri T. Cellulase families and their genes. *Trends in biotech*. 1987. V. 5. P. 255–261.

11. Bateman A., Coin L., Durbin R., Finn R.D., Hollich V., Griffiths-Jones S., Khanna A., Marshall M., Moxon S., Sonnhammer E.L.L. et al. The Pfam protein families database. *Nucleic Acids Research*. 2004. V. 32. P. 138–141.
12. Rice P., Longden I., Bleasby A. EMBOSS: the European molecular biology open software suite. *Trends in Genetics*. 2000. V. 16. № 6. P. 276–277.
13. Larkin M.A., Blackshields G., Brown N.P., Chenna R., McGettigan P.A., McWilliam H., Valentin F., Wallace I.M., Wilm A., Lopez R. et al. Clustal W and Clustal X version 2.0. *Bioinformatics*. 2007. V. 23. № 21. P. 2947–2948.
14. Tamura K., Dudley J., Nei M., Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. *Molecular Biology and Evolution*. 2007. V. 24. P. 1596–1599.
15. Pieper U., Eswar N., Braberg H., Madhusudhan M.S., Davis F.P., Stuart A.C., Mirkovic N., Rossi A., Marti-Renom M.A., Fiser A et al. MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Research.* 2004. V. 32. P. D217–D222.
16. Crooks G.E., Hon G., Chandonia J.M., Brenner S.E. WebLogo: A Sequence Logo Generator. *Genome Research.* 2004. V. 14. P. 1188–1190.

Материал поступил в редакцию 24.04.2013, опубликован 16.07.2013.