Биомедицинская химия, 2004, том 50, вып. 6, с.554-565

БИОИНФОРМАТИКА

УДК 612.853-205

СТРУКТУРНО-ФУНКЦИОНАЛЬНЫЕ МОТИВЫ СТЕРОЛОВЫХ 14-АЛЬФА - ДЕМЕТИЛАЗ (CYP51)

А.В.Лисица[1], С.А.Гусев[1], Ю.В.Мирошниченко[1], Г.П.Кузнецова[1], В.Н. Лазарев[2], В.С.Скворцов[1], И.И.Карузина[1], В.М.Говорун[2], А.И.Арчаков[1]

[1]Государственное учреждение Научно-исследовательский институт биомедицинской химии им. В.Н.Ореховича РАМН, 119121 Москва, Погодинская ул., 10; тел.(095) 2473960; факс (095)2450857; эл. почта: fox@ibmh.msk su
[2]НИИ физико-химической медицины МЗ РФ, 119828 Москва, ул. Малая Пироговская, 1а

Ферменты, объединяемые в семейство цитохромов Р450 - семейства CYP51, катализируют деметилирование стероидных субстратов по положению 14-альфа. В отличие от остальных представителей надсемейства цитохромов Р450, члены семейства CYP51 встречаются во всех царствах живой природы, при этом сохраняя достаточно высокую консервативность Поэтому предполагается, что именно семейство CYP51 является родоначальником всех современных форм цитохромов Р450. В ходе настоящего исследования при помощи кластерного анализа, множественного выравнивания и статистического метода выявления структурно-функциональных мотивов было проанализировано 36 полноразмерных последовательностей аминокислотных остатков представителей семейства стероловых деметилаз. Предложен статистический критерий, согласно которому наличие или отсутствие структурно-функциональных мотивов в консенсусной последовательности множественного выравнивания трактуется как информационное содержание. Поскольку множественное выравнивание зависит от группы выравниваемых последовательностей, а состав группы, в свою очередь, определяется топологией дендрограммы кластерного анализа, то информационное содержание позволяет также уточнять структуру дендограммы. Высоким информационным содержанием характеризовались консенсусные последовательности для следующих кластеров: деметилазы грибкового происхождения, ферменты животных + ферменты растений, ферменты растений + ферменты простейших и группа белков бактерий. Указанные группы использовались для получения консенсусной последовательности для всего семейства CYP51. В консенсусной последовательности семейства были выявлены статистически достоверные участки консервативности. Эти области (мотивы) совмещались с данными об элементах вторичной структуры и местах субстратного узнавания, установленных для CYP51 из *Mycobacterium tuberculosis* CYP51 (MT). Выявлено 7 мотивов, являющихся, по-видимому, обязательными для каждого белка семейства CYP51. Для оценки роли каждого мотива в обеспечении специфичности реакции 14-альфа-деметилирования, вычислялась вероятность обнаружения мотивов в других цитохромах Р450, не входящих в семейство стероловых деметилаз Некоторые из обнаруженных мотивов оказались абсолютно специфичны для стероловых 14-альфа-деметилаз, в то время как другие являлись общими для различных форм цитохромов Р450.

Ключевые слова: стерол-14-альфа-деметилаза (CYP51), *Mycobacterium*, структурно-функциональный мотив, консенсусная последовательность, субстрат-связывающие участки

ВВЕДЕНИЕ Цитохромы Р450, катализирующие отщепление 14-альфа-метиловой группы стероидных субстратов, классифицируются в семейство CYP51 стероловых 14-альфа-деметилаз [1]. В настоящее время это семейство объединяет 42 фермента. Распределение представителей семейства CYP51 в природе не равномерное. большинство ферментов выделено из грибов, тогда как простейшие представлены только единственным цитохромом Р450 из генома трипаносом

В последние десятилетия представители семейства CYP51 являлись объектом интенсивного исследования благодаря ключевой роли этих ферментов в метаболизме грибов В грибах CYP51 участвует в синтезе эргостерола из его предшественника - ланостерола [2]. Эргостерол, грибковый аналог холестерина млекопитающих, выполняет важные функции в обеспечении соответствующей консистенции клеточной мембраны. Ингибирование биосинтеза эргостерола азольными ингибиторами подавляет рост патогенных грибков.

Присутствие стероловой 14-альфа-деметилазы в представителях всех царств, возможно, свидетельствует о том, что именно семейство CYP51 является прародителем современного многообразия форм цитохромов Р450 [3]. Эволюционное исследование стероловых 14-альфа-деметилаз [4], основанное на сравнительном анализе белковых и генных структур, показало, что топология эволюционного древа согласуется с филогенией основных царств живой природы: бактерии, растения, грибы, млекопитающие. Это также подтверждает гипотезу о прокариотическом происхождении цитохромов Р450.

Методом рентгено-структурного анализа определена структура бактериального CYP51 MT (CYP51 из *Mycobacterium tuberculosis*) [5]; те же авторы впервые доложили о субстрат-узнающих участках (SRS-substrate Recognition Site) стероловых 14-альфа-деметилаз [6] Предсказания были сделаны путем наложения данных рентгено-структурного анализа CYP51MT (код в Protein Data Bank - 1ea1) на выровненные аминокислотные последовательности семейства CYP51. На выравнивание также были нанесены точечные мутации, остатки, контактирующие с докированным субстратом, и субстрат-узнающие участки, предсказанные ранее для членов семейства CYP2 [7].

Учитывая все вышеизложенное, представлялось интересным выявить в первичных структурах стероловых деметилаз локальные консервативные участки, ответственные за обеспечение функциональной специфичности данного фермента - т.е. найти *структурно-функциональные мотивы*. В соответствии с оригинальным подходом мы рассматриваем природную последовательность аминокислотных остатков как объект чрезмерно высокой степени сложности. Выравнивая несколько последовательностей друг с другом и вычленяя общую (консервативную) часть, можно искусственно снизить сложность анализируемого объекта - (консенсусной) последовательности. С точки зрения исследователя, снижение сложности приводит к повышению информативности (информационного содержания) [8].

Информационное содержание консенсусной последовательности напрямую связано с узлами дендрограммы кластерного анализа. Действительно, выбор конкретного узла определяет группу белков в кластере, а, следовательно, определяет и картину множественного выравнивания, саму консенсусную последовательность и ее информационное содержание. Выбирая различные узлы дерева, можно выявить консенсусные последовательности с относительно более высоким информационным содержанием. Указанный подход, во-первых, помогает уточнить топологию дендрограммы, а, во-вторых, позволяет обнаружить структурно-функциональные мотивы.

Предложенный непараметрический статистический критерий позволяет дать оценку информативности консенсуса. Критерий чувствителен к распределению участков консервативности вдоль консенсусной последовательности, при этом более высокие значения критерия соответствуют тем участкам последовательности, где консервативные позиции *сконцентрированы*. В ходе кластеризации выявлялось такое разбиение выборки на группы, которое обеспечивало бы наиболее высокое суммарное информационное содержание соответствующих консенсусных последовательностей. В результате,

консенсусные последовательности, полученные для выявленных кластеров белков, содержат компактные участки консервативности - мотивы.

Мотивы можно использовать для определения того, какие структурные особенности данного белка являются обязательными для отнесения его в данное семейство, а какие - нет. Чтобы различить обязательные и факультативные мотивы, проводилась оценка их специфичности путем поиска соответствующих паттернов в последовательностях цитохромов Р450, не входящих в семейство CYP51.

В данной работе приводится поэтапное описание процедуры нахождения структурно-функциональных мотивов. Процедура включает в себя кластерный анализ, множественное выравнивание, отбор консенсусных последовательностей с высоким информационным содержанием, выявление мотивов в отобранных консенсусах и оценка специфичности мотивов.

МЕТОДИКА. *Основные термины*. В дальнейшем мотив рассматривается как участок консервативности среди первичных структур в группе функционально родственных белков (в данном случае, стероловых 14-альфа-деметилаз). С определенными допущениями можно предполагать, что мотив ответственен за обеспечение ферментативной функции белка и/или за правильную сборку его пространственной структуры (фолдинг). Формально, мотив может выявляться в большинстве белков в группе, тогда как *паттерн* должен присутствовать в каждом белке. На основе исходного множественного выравнивания мотив может быть преобразован в паттерн с учетом всех вариаций, встречающихся в консервативных позициях мотива. В данной работе сначала выявляются мотивы - т.е. конструкции, допускающие некоторую степень вариабельности, - а затем мотивы преобразуются в паттерны с тем, чтобы оценить их специфичность. Избранный подход оправдан, поскольку некоторый допуск на уровень консервативности мотива обеспечивает устойчивость к незначительным точечным отклонениям в анализируемых первичных структурах. Эти отклонения (по типу точечных мутаций) могут появиться, вообще говоря, и в результате ошибок прочтения последовательности. Для оценки специфичности сформированный паттерн ищется в первичных структурах цитохромов Р450, не принадлежащих к семейству CYP51. Определив в результате поиска в базе данных число ложно-положительных результатов (FP-False Positives), содержащих (N) последовательностей, специфичность мотива можно рассчитать как (N - FP)/N.

Консенсусная последовательность и ее информационное содержание. Несколько выровненных (методом итеративной оптимизации [9]) белковых последовательностей могут быть представлены в виде одной консенсусной последовательности (консенсуса). Консенсусная последовательность формируется с учетом двух параметров, задаваемых пользователем: во-первых, порога консервативности и, во-вторых, редуцированных аминокислотных остатков алфавита.

Порог консервативности определяет минимальную частоту встречаемости аминокислотного остатка в колонке выравнивания, необходимую для внесения данного остатка в строку консенсуса. Редуцированный алфавит позволяет объединять в группы остатки, сходные по своим физико-химическим (или другим) характеристикам. В работе используется объединение остатков в пять групп. [n]: L, I, M, V; [a]: F, Y, W; [s]: A, S, T, G; [+]: K, R, H и [=]: D, E, Q, N.

Таким образом, консенсусная последовательность состоит из следующих элементов: вариабельных позиций (обозначаемых точками или символом "x" в мотивах), вставок (тире) и консервативных позиций (или аминокислотный остаток, или символ редуцированного алфавита, обозначающий группу остатков). В зависимости от выбранного уровня консервативности и заданного редуцированного алфавита относительное содержание консервативных позиций в консенсусе меняется, влияя на его плотность. Плотность выражается как отношение числа консерватиных позиций к общей длине консенсусной последовательности.

Цель использования консенсусов состоит в том, чтобы искусственно снизить сложность природных белковых последовательностей, сравнивая их друг с другом и вычленяя общую часть. Имеет смысл сравнить консенсус с математическим

ожиданием, а группу белков - с выборкой случайных величин, на основании которых дается оценка среднего. Участки консервативности рассматриваются как нечто существенно необходимое для адекватного функционирования белка. При расширении группы анализируемых структур, плотность консенсусной последовательности понижается, что делает эти необходимые участки более заметными. Однако, если двигаться дальше в этом направлении, т.е. добавлять в группу все новые и новые белки, то рано или поздно наступит ситуация вырожденного консенсуса, в котором присутствуют только вариабельные позиции и совсем нет консервативных. Очевидно, что вырожденный консенсус не несет никакой существенной информации. Итак, между двумя одинаково лишенными смысла крайностями - слишком сложной природной белковой последовательностью и чрезмерно упрощенным вырожденным консенсусом - должен быть оптимум информационного содержания консенсусной последовательности.

Тривиальный оптимум можно было бы ожидать при 50% плотности консенсуса, когда достигается баланс между "потерянной" и "сохраненной" информацией. Однако, при таком подходе игнорируется характер распределения консервативных позиций вдоль консенсуса. Действительно, консервативные позиции могут равномерно чередоваться с вариабельными, или, напротив, консервативные позиции могут образовывать компактные кластеры. Принимая за данность наличие в структуре белка мотивов, следует большие значения информационного содержания ассоциировать с теми консенсусными последовательностями, в которых консервативные позиции сгруппированы в этом случае, как исходная природная последовательность, так и вырожденная консенсусная строка - обе получат низкую оценку, потому что ни одна из этих структур не содержит в себе мотивы.

При последовательном добавлении к множественному выравниванию новых (родственных) последовательностей, можно наблюдать постепенный рост информационного содержания, т. к. мотивы начинают обосабливаться на фоне случайных мутационных флуктуаций. В определенный момент консенсус достигнет оптимального разделения на мотивы, и добавление к выравниванию новых белковых последовательностей только лишь ухудшит результат. Чтобы оценивать информативность консенсуса в терминах содержания мотивов, применялся статистический критерий Шермана [10].

Связь между древовидным представлением результатов кластеризации и информационным содержанием консенсуса. Вообще говоря, желательно найти такое объединение белков в кластеры, при котором суммарное информационное содержание всех консенсусных последовательностей этих кластеров было бы максимальным. Эвристическое решение этой задачи может быть достигнуто путем использования представления результатов кластеризации в виде дендрограммы (использовался метод объединения ближайших соседей, реализованный в пакете программ "pfaat" [11]). Начиная со стадии, когда каждая индивидуальная белковая последовательность занимает отдельный кластер, будем продвигаться вглубь дендрограммы по направлению к завершающей стадии единственного кластера. Для каждого узла бифуркации, встречающегося на этом пути, можно обособить соответствующую группу белков, провести их множественное выравнивание и получить консенсус. Для полученного консенсуса содержание информации оценивалось как описано выше, и полученное значение отображалось в зависимости от плотности консенсуса.

Очевидно, что каждая точка на полученном графике "плотность-информативность" соответствует определенному узлу дендрограммы. Известно, что чем более глубоко расположены ветви дендрограммы, тем менее ясна их взаимная топология. Формально это отображается низкими оценками бутстрапа. Бутстрап - это подход, при котором вносятся незначительные флуктуации в исходные данные (в последовательности), и фиксируются изменения топологии дендрограммы. Численная оценка бутстрапа показывает, сколько раз определенная топология появлялась после внесения флуктуаций.

Для ветвей (узлов) с низкими значениями бутстрап-оценки целесообразным является опробовать различные альтернативные варианты топологии и отобрать те из них, которые порождают консенсусы с относительно более высокой информативностью. Анализ альтернативных разветвлений привносит на график плотность-информативность дополнительные точки, для которых нет соответствующих узлов в первоначальной дендрограмме.

Из графика информативность-плотность видно, что низкие значения информационного содержания присущи консенсусным последовательностям как высокой, так и низкой плотности. На определенном участке кривая информационного содержания достигает максимума. Конкретное положение максимума зависит от состава группы анализируемых белков. Консенсусные последовательности групп белков в районе максимума отбирались для дальнейшего анализа. Одновременно, точки с высоким информационным содержанием привносили изменения в картину кластеризации.

Для построения кривой информативность-плотность применялись консенсусные последовательности, полученные при 100% уровне консервативности - т.е. инвариантные консенсусы. Данное ограничение необходимо для адекватного решения задачи оптимального разбиения на кластеры. Однако, непосредственно для выявления точных границ мотивов, можно варьировать пороговым значением консервативности (или даже манипулировать с редуцированным алфавитом, что, однако, не делалось в данной работе). В случае, если консенсусная последовательность изобилует вариабельными позициями, то уровень консервативности целесообразно снизить с отметки 100%. Снижение порога имеет смысл проводить только до тех пор, пока растет информационное содержание. На максимуме информативности получается консенсус, оптимальным образом отражающий исходные данные.

Базы данных. Белковые последовательности стероловых 14-альфа-деметилаз были загружены из базы данных по цитохромам P450 (http://cpd.ibmh.msk.su) [12]. Из 42 депонированных в базе данных первичных структур, было отобрано 36 полноразмерных (с длиной не менее 350 аминокислотных остатков). Неиспользованные фрагментарные структуры представляют собой участки гиперконсервативных областей, соответствующих альфа-спиралям I и K - эти участки обычно используются для конструирования праймеров при клонировании. Сформированная выборка приведена в таблице 1. Средняя длина отобранных последовательностей составляла 506±31 аминокислотных остатков; средняя идентичность 41+15%. Указанный процент идентичности формально позволяет рассматривать отобранную совокупность белков в качестве семейства, поскольку превышает стандартный 40%-й порог. В тоже время, идентичность наиболее удаленных гомологов в семействе (например, идентичность между деметилазами бактерий и простейших) составляет всего лишь 20%.

Специфичность мотивов оценивалась путем поиска их в белковых последовательностях, депонированных в базе данных по цитохромам P450. База данных содержит 1066 полноразмерных белковых последовательностей, аннотированных Номенклатурным комитетом к середине 2002 года.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ. На рисунке 1 приведена дендрограмма, полученная для белков семейства CYP51. Белковые последовательности из семейств CYP7 и CYP5 использовались в качестве объектов внешней группы. При бутстрап-анализе, в 100 случаях из 100, последовательности CYP51 группировались в кластеры, соответствующие царствам, поэтому ветвление внутри этих групп не приводится. Между кластерами царств значения бутстрап-оценки в большинстве случаев не превышают 60. Кроме того, один из бактериальных белков - CYP51 из *S. coelicolor* - обнаруживается в составе внешней группы, т. е. стандартный метод кластеризации оказался не в состоянии выявить родство между этим белком и остальными членами семейства стероловых деметилаз. Дальнейший анализ мотивов показал, что CYP51 из *S. coelicolor* действительно должен быть отнесен к группе бактериальных деметилаз, поскольку только при объединении всех бактериальных цитохромов P450, включая

Таблица 1. Список полноразмерных белков семейства CYP51

Животные (6 последовательностей)
Человек, свинья, крыса: мышь; курица, *Fugu rubripes*
Грибы (20 последовательностей)
Schizosaccharomyces pombe; Ustilago maydis; Uncinula necator; Erysiphe graminis; Blumeria graminis; Mollisia yalludae, Mollisia acuformis; Penicillium italicum; Penicillium digitatum; Candida albicans, Candida tropicalis; Candida krusei[1,2]*; Candida glabrata; Saccharomyces cerevisiae; Filobasediella neoformans; Aspergills fumigatus; Aatergillus nidulans; Boryotinia fuckeliana: Neurospora crassa; Venturia inaequalis*
Растения (5 последовательностей)
Сорго (хлебный злак); пшеница; рис[1]; томат; *Arabidopsis thaliana*
Простейшие (1 последовательность)
Trypanosoma brucei
Бактерии (4 последовательности)
Mycobacterium tuberculosis, Methylococcus capsulatus, Streptomyces coelicolor, Mycobacterium smegmatis

Примечание: [1]C-конец отсутствует; [2]N-конец отсутствует

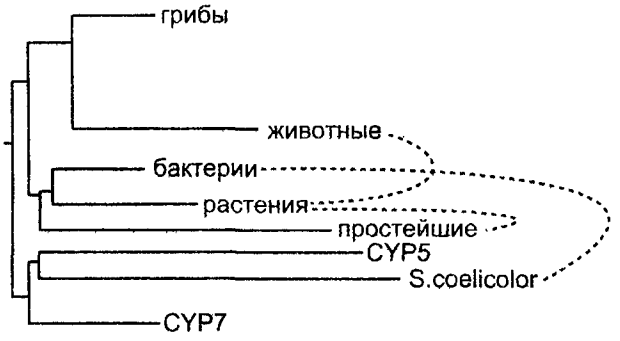

Рисунок 1
Дендрограмма, полученная для семейства CYP51.
Показаны только ветви с низкими значениями бутстрап-оценки. Пунктирные кривые показывают альтернативные топологии, которые характеризуются консенсусными последовательностями с относительно более высоким информационным содержанием.

и белок из *S. coelicolor*, удается получить консенсус с достоверно высоким информационным содержанием.

Для каждого узла иерархического дерева была получена консенсусная последовательность (см. раздел "методика"). На основе полученных консенсусов, был построен график зависимости информативности от плотности (рис. 2). Каждая точка на графике сопоставлена с определенным узлом дендрограммы. Консенсусные последовательности высокой плотности обладают низким информационным содержанием, что видно на примере кластеров животных или растительных цитохромов P450. Также многие из внутренних узлов в пределах кластеров царств характеризуются консенсусными последовательностями высокой плотности и низкого информационного содержания. Индивидуальные белковые последовательности имеют плотность, равную единице и нулевую информативность. С другой стороны, вырожденные консенсусные последовательности также демонстрируют низкое информационное содержание

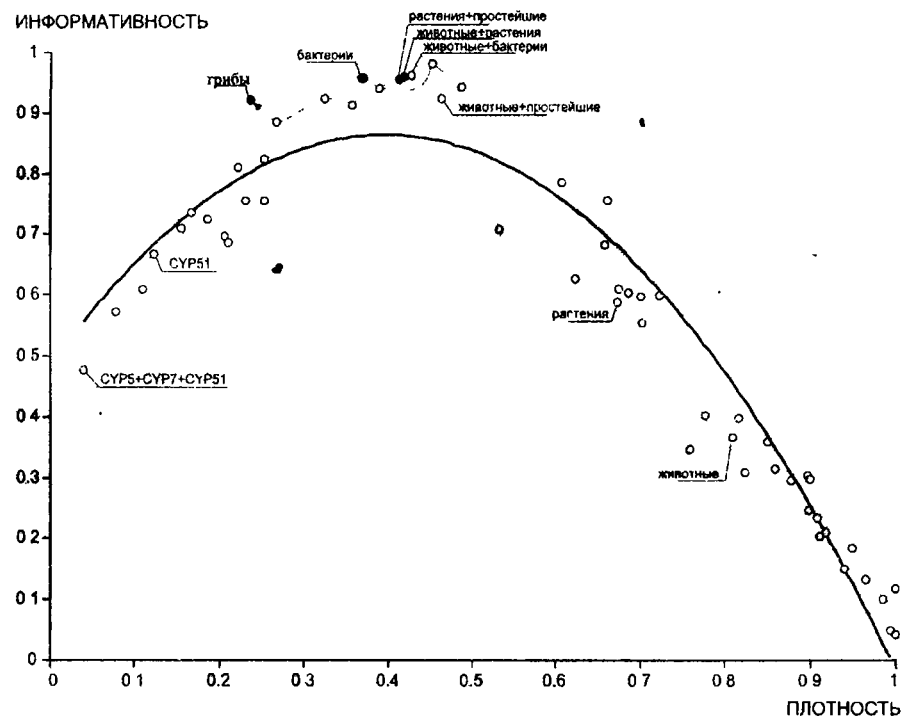

Рисунок 2.

Кривая плотность-информативность. Пунктирная линия соединяет точки, которые соответствуют внутренним узлам кластера грибов: все они объединяются в один кластер (отмечено стрелкой) без существенной потери информативности. Сплошные кружки соответствуют кластерам с высоким информационным содержанием. Соответствующие этим кластерам консенсусные последовательности были отобраны для дальнейшего анализа.

из-за их низкой плотности Как крайний случай можно отметить общую консенсусную последовательность, полученную на основе множественного выравнивания всех анализируемых белков (обозначенную на графике как CYP5+CYP7+CYP51). При этом значительная часть информации потеряна, но информационное содержание отнюдь не равняется нулю, поскольку все последовательности принадлежат одному и тому же надсемейству. Как и любой цитохром P450, консенсус CYP5+CYP7+CYP51 содержит домены высокой консервативности· гем-пептид, альфа-спирали I и K Консенсусная последовательность семейства CYP51 также имеет относительно низкое информационное содержание и таким образом не может непосредственно использоваться для обнаружения мотивов.

В верхней части графика расположены точки, соответствующие консенсусным последовательностям с высоким информационным содержанием. Многие из них соответствуют внутренним узлам кластера дрожжей, эти точки соединены пунктирной линией. Справедливо заключить, что все внутренние узлы дрожжевых CYP51 могут быть объединены в единый кластер без значительной потери информации о мотивах Другой кластер, также имеющий сравнительно высокую оценку информационного содержания, объединяет формы стероловых деметилаз бактериального происхождения. Однако, бактериальная группа обладает высоким информационным содержанием только в том случае, если последовательность *S. coelicolor* соединяется с остальными бактериальными белками (как показано пунктиром на рис 1).

Следующей задачей являлось получение высокоинформативных консенсусов для групп белков растений, животных и простейших Консенсусные последовательности первых двух групп имеют высокую плотность, тогда как группы простейших не имеют таковой: имеется единственная последовательность

CYP51 из *Trypanosoma brucei*. Были проанализированы различные возможные комбинации из указанных групп: для каждой комбинации вычислялось информационное содержание. Видно, что группировка последовательности трипаносомы с группой белков животных имеет относительно более низкое информационное содержание, чем при группировке простейшего с группой растений По той же самой причине (то есть из-за высокой информативности) должна быть отобрана комбинация животные+растения (пунктир на рис 1). Причина, по которой следует исключить комбинацию животные+бактерии, состоит в том, что бактериальный кластер сам по себе имеет высокое информационное содержание и, таким образом, будучи объединенным с любым другим кластером, может служить искусственным фильтром для увеличения информационного содержания. В результате оценки всех комбинаций, наиболее предпочтительными оказываются следующие четыре группы белков, характеризующиеся консенсусными последовательностями высокой информативности: грибы, бактерии, растения+простейшие и животные+ растения. Каждый из белков семейства CYP51 вносит равный вклад в эти консенсусные последовательности, и каждая из этих консенсусных последовательностей несет достаточную информацию о структурно-функциональных мотивах.

Проведя множественное выравнивание четырех отобранных консенсусов, можно получить общую консенсусную последовательность для всего семейства стероловых деметилаз. Очевидно, она будет иметь низкое информационное содержание (рис 2, CYP51), но это можно скорректировать, снизив консервативность до 75%. Это невозможно было сделать при выравнивании индивидуальных белков семейства, потому что в этом случае результирующая консенсусная последовательность отражала бы особенности наиболее широко представленной группы, т.е. особенности цитохромов P450 грибкового происхождения. С другой стороны, используя иерархический подход и двигаясь к консенсусу семейства через консенсусы групп, удается обеспечить равное участие каждого из исходных белков в процессе формирования мотивов.

Общая консенсусная последовательность представлена на рисунке 3 Консенсус содержит 585 позиций, 66 (11%) из них являются консервативными. Заряженные остатки встречаются более редко (16%), чем неполярные (32%). Интересно, что доля неполярных остатков (32%) превышает суммарную фоновую частоту встречаемости остатков лейцин, валин, изолейцин и метионин в исходных белках (26%). Это может указывать на важную роль, которую гидрофобные остатки играют в образовании глобулярного фолда белка

Применение статистического критерия Шермана позволило вычленить 7 структурно-функциональных мотивов (MTF, motif). Мотивы отмечены на рисунке 3 прямоугольниками В таблице 2 представлена информация по каждому мотиву. Первый мотив расположен в области бета-структуры b1-1. Мотив встречается в последовательностях 28 цитохромов P450, не входящих в семейство стероловых 14- альфа-деметилаз Ложно-положительные результаты получены для цитохромов P450 подсемейств CYP2D, CYP4F и CYP7A млекопитающих, CYP33E от C. elegans, для растительного CYP97 и бактериального CYP101A. Во всех этих последовательностях MTF1 локализован на N-конце (позиции 50-100). Однако, в некоторых случаях, мотив был обнаружен и на С-конце, приблизительно в районе 500 остатка. Также MTF1 встречается дважды в одном из белков семейства CYP51 (из *U. necator*). Вторая копия начинается с позиции 506 в области бета-структуры 4.

Мотив MTF2 приходится на участок петли между альфа-спиралями B и C. Предполагается, что BC-петля участвует в формировании канала доступа субстрата в активный центр фермента, закрывая этот канал сверху [5,6]. MTF2 содержит остатки, которые контактируют с субстратом, в то время как сам мотив расположен в пределах первого субстрат-узнающего участка (SRS). Сигнатура MTF2 очень специфична для стероловых деметилаз, ложно-положительные результаты встречаются только у шести цитохромов P450, половина из которых принадлежит семействам CYP71 и CYP52. Ни один из ложных результатов не содержит сигнатуру

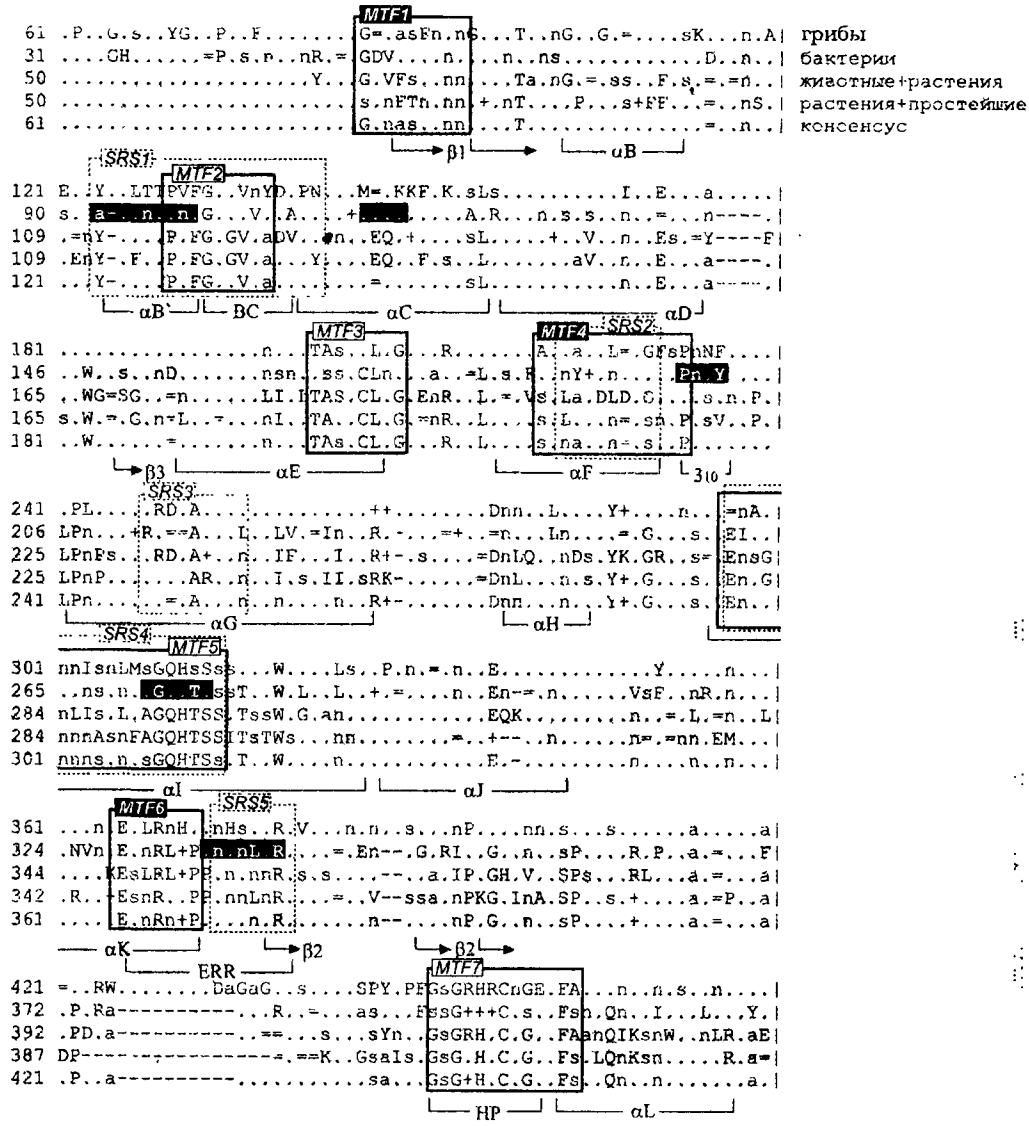

Рисунок 3.

Структурно-функциональная карта семейства CYP51. Множественное выравнивание четырех консенсусных последовательностей высокой информативности (грибы, бактерии, животные+растения, растения+простейшие) представлено в виде консенсуса всего семейства стероловых 14-альфа-деметилаз. Консервативность консенсуса семейства задана на уровне 75% с тем, чтобы обогатить его информационное содержание. Прямоугольники, обведенные сплошной линией, и обозначенные "MTF" изображают мотивы, найденные при помощи критерия Шермана. Прямоугольники, обведенные пунктирной линией и обозначенные "SRS", изображают субстрат-узнающие участки [7]. Черным фоном на бактериальной последовательности показаны кластеры остатков, которые входят в контакт с докированным субстратом [6]. Элементы вторичной структуры размечены под общей консенсусной последовательностью. BC - BC-петля; MDR-извилина; ERR - ERR-триада; HP - гем-пептид. Редуцированный алфавит аминокислотных остатков [n]: L, I, M, V, [a]: F,Y,W; [S]: A,S,T,G; [+]: K,R,H и [=]: D,E,Q,N.

MTF2 в области BC-петли. В соответствующей области этот мотив не встречается ни у одного из белков других семейств. С другой стороны, мотив дублируется в последовательности одного из членов семейства CYP51 (*V. inaequalis*).

MTF3 показал только 3 ложно-положительных результата (в семействах CYP60 и CYP71), т. е. он практически абсолютно специфичен для стероловых деметилаз.

Таблица 2 Мотивы, выявленные в консенсусных последовательностях семейства CYP51

#	Мотивы	Локализация[1]	Специфичность (%)
1	[GS]x[VILC][AGFYV][RSTWE]xx[LIMC][VLIMGAF]	β1-1	77
2	[GP]x[FL]Gxx[VG]x[FYA]	BC-loop	99
3	[TV][AS][ASTGC]x[CAST]Lx[GR]	αE	99,5
4	[AS]x[LIVMY][FYL]xx[LM][DEY]x[GSR]xx[PVH]	αF	99,3
5	[DE][LIV]xx[MLQI][MLVFI][IV][ASTG]x[LMI]x[APG]G[SHQ][HE][TS][SI][AS]	αI	100
6	Ex[LIM]R[LIMVR][RHD][PTSALHM]	αK	75
7	[GS][AG]G[RKV][HR]xCx[GS]xxF[ASG]	heme-peptide	98

Примечание: [1]Элементы вторичной структуры приведены согласно данным рентгено-структурного анализа CYP51MT [5]

Этот мотив кодирует как раз ту часть альфа-спирали E, которая совместно с BC-петлей формирует стенки канала доступа субстрата. Однако, ни один из остатков внутри или около MTF3 не контактирует с докированным субстратом

MTF4 приходится на альфа-спираль F. Пространственно эта спираль одним концом приближена к спирали E, а другим переходит в FG-петлю. FG-петля формирует второй канал доступа субстрата в активный центр. Примечательно, что MTF4 совпадает с SRS2; однако данные молекулярного докинга не указывают, что MTF4 обеспечивает распознавание субстрата. В контакт с субстратом входит следующая за F-спиралью FG-петля, имеющая 3_{10}-спиральную конформацию. Расположение мотива 4 в районе альфа-спирали F достаточно специфично для стероловых деметилаз.

MTF5 - самый длинный мотив, он описывает консервативную и наиболее протяженную спираль I. MTF5 не дает ни одного ложно-положительного результата, очевидно из-за своей длины.

Следующий мотив (MTF6) соответствует альфа-спирали K - второму (после спирали I) высококонсервативному участку первичной структуры цитохромов P450. Поэтому не удивительно, что MTF6 встречается во многих других цитохромах P450, и, следовательно, характеризуется низкой специфичностью.

Гем-пептид представлен мотивом MTF7. На седьмой позиции в этом мотиве находится цистеин, являющийся пятым аксиальным лигандом гема цитохромов P450. Паттерн гем-связывающего участка является общим для многих форм фермента; обычно именно это место используется в качестве зонда при поиске цитохромов P450 в геномах. Тем более неожиданно то, что при поиске MTF 7 обнаруживается только 21 ложно-положительный результат (казалось бы, этот мотив должен встречаться во многих цитохромах P450) Для объяснения феномена необычно высокой специфичности, были исследованы несколько альтернативных вариантов MTF7 (табл.3).

За основу был взят мотив, депонированный в базе данных PROSITE (http://www.expasy.ch/cgi-bin/nicedoc.pl?PDOC00081), поскольку этот мотив отвечает практически всем цитохромам P450, т.е. его специфичность равна нулю. Различия между мотивом PROSITE и мотивом MTF7 наблюдались в двух позициях - (-3) и (+5) - если за начало отсчета брать гем-связанный цистеин Каждая из этих позиций тестировалась с точки зрения специфичности. Сначала, положительно заряженный остаток помещали в позицию (-3) мотива 7, затем вводили ароматический остаток в позицию (+5). Ни одна из этих модификаций, взятая отдельно, не вела к высокой специфичности, тогда как их комбинация оказалась крайне специфичной (43 ложно-положительных результата, большинство из которых было обнаружено в семействе CYP71). Избирательность MTF7 становится почти абсолютной, если положительный заряд в положении (-3) представлен только аргинином.

ЗАКЛЮЧЕНИЕ. Выявление структурно-функциональных мотивов (областей консервативности) зачастую является одним из первых этапов анализа последовательностей аминокислотных остатков. Выявленные в семействе CYP51 мотивы локализуются в функционально важных областях По литературным

Таблица 3. Специфичность гем-связывающего участка CYP51

Позиции:	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	S(%)
MTF7	G	AG	G	RV	H	x	C	LIMVP	G	x	x	F	AG	98
PROSITE	SGNH	X	GD	X	RKHPT	x	C	LIMVFAP	GAD	x	x	x	x	5
Положительно заряженный остаток(-3)	SGNH	X	GD	RKV	RKHPT	x	C	LIMVFAP	GAD	x	x	x	x	46
Ароматический остаток(+5)	SGNH	X	GD	X	RKHPT	x	C	LIMVFAP	GAD	x	x	FYW		74
Положительно заряженный остаток (-3) и и ароматический остаток (+5)	SGNH	X	GD	RKV	RKHPT	x	C	LIMVFAP	GAD	x	x	FYW		95

Примечание: За начало отсчета (нулевая позиция) принят остаток цистеина, являющийся 5-ым аксиальным лигандом гема в CYP51 из *M. tuberculosis* - позиция 394 и позиция 470 в CYP51 из *C. albicans*. "x "- обозначает любой остаток, "S"- специфичность

данным, мотивы 5, 6 и 7 ответственны за фиксацию гема, а мотив 2 представляет собой участок, обеспечивающий узнавание субстрата. Высокая специфичность мотивов 2, 3 и 4 свидетельствует в пользу того, что данные конструкции участвуют в обеспечении функционирования стероловых деметилаз.

Основываясь на результатах молекулярного моделирования и докинга, можно предположить наличие у стероловых 14-альфа-деметилаз двух каналов доступа к активному центру. Первый канал образован BC-петлей, этот участок проявился как MTF2. Второй канал формируется FG-петлей, по-видимому этот канал связан с участком консервативности MTF4. В структуре CYPBM3 (CYP450 из *B.megaterium*) FG-петля прикрывает канал доступа, образованный бета-структурами b1 [13]. Отметим, что фрагмент b1-структуры появляется в структурно-функциональном отображении как MTF1.

Таким образом, при сравнительном анализе белковых последовательностей семейства CYP51 обнаруживаются два типа мотивов. Первый тип - это мотивы *общности* - т.е. мотивы, которые можно найти в структурах всех цитохромов P450, а не только в стерловых деметилазах. К мотивам общности следует отнести MTF5, 6 и 7. Мотивы второго типа - мотивы *частного* - обеспечивают специфичность цитохромов P450 конкретного семейства. Для стероловых 14-альфа-деметилаз мотивы частного - это MTF2, MTF4 и, возможно, MTF7.

Строго говоря, наличие таких образований, как структурно-функциональные мотивы на настоящий момент не является доказанным принципом строения белковых молекул. В то же, время, наиболее распространенные методы функционального аннотирования новых генов базируются именно на поиске участков локального сходства последовательностей. При этом локальность консервативных участков как отличительный признак либо закладывается в сам алгоритм (например, алгоритм BLAST [14]), либо вводится косвенно, путем задания штрафа за вставку при глобальном выравнивании. Используя критерий оценки информационного содержания, удалось сделать более явственными требования к компактному характеру расположения консервативных позиций в выравнивании, а также применить метод поиска локальных участков консервативности к результатам множественного выравнивания. Предложенный подход для выявления структурно-функциональных мотивов может послужить инструментом для создания надежной методики автоматической классификации белков в составе надсемейств. С другой стороны, углубленное исследование мотивов отдельного семейства, подобно проведенному в отношении стероловых деметилаз, позволяет планировать и направлять генно-инженерные эксперименты по клонированию белков с химерными функциями. Наконец, данные о мотивах позволяют оптимизировать процесс конструирования трехмерных моделей белков [15].

Работа поддержана грантами РФФИ № 04-04-48030 и НШ №325.2003.4.

ЛИТЕРАТУРА

1 *Nelson D., Koymans L., Kamataki T., Stegeman J., Waxman D., Waterman M., Gotoh O., Coon M, Estabrook R., Gunsalus I., Nebert D* (1996) Pharmacogenetics, **6**, 1-42

2 *Lamb D., Kelly D., Manning N., Hollomon D., Kelly S* (1998) FEMS Microbiol Lett, **69**, 369- 373

3 *Nelson D.* (1999) Arch Biochem Biophys, **369**, 1-10

4 *Yoshida Y., Noshiro M., Aoyama Y., Kawamoto T., Horiuchi T., Gotoh O.* (1997) J Biochem (Tokyo), **122**, 1122-1128

5 *Podust L., Poulos T., Waterman M.* (2001) Proc Natl Acad Sci USA, **98**, 3068.

6 *Podust L., Stojan J., Poulos T, Waterman M* (2001) J Inorg Biochem, **87**, 227-235

7 *Gotoh O.* (1992) J Biol Chem, **267**, 83-90

8 *Lisitsa A., Archakov A., Lewi P* (2003) Meth Find. Exper Clin Pharmacol, **25**(9), 733-745

9 *Gotoh O.* (1996) Adv Biophys, **36**, 159-206

10 *Sneath P* (1998) Bioinformatics, **14**, 608-616

11 *Johnson J., Mason K, Moallemi C., Xi H., Somarroo S., Huang E.* (2003) Bioinformatics, **19**, 544- 545

12 *Lisitsa A, Gusev S, Karuzina I., Archakov A., Koymans L* (2001) SAR QSAR Environ Res, **12**, 359-366

13 *Graham S., Peterson J.* (2002) Methods Enzymol, Elsevier, **357**, pp 15-28

14 *Altschul S., Gish W, Miller W., Myers E.W., Lipman D.J* (1990) J Mol Biol, **215**, 403- 410

15 *Chakrabarti S., John J, Sowdhamini R* (2004) J. Mol Model, **10**, 69-75.

Поступила 01.06 2004

STRUCTURAL-FUNCTIONAL MOTIFS OF STEROL 14-ALPHA DEMETHYLASES (CYP51)

A.V. Lisitsa[1], S.A. Gusev[1], Y.V. Miroshnichenko[1], G.P. Kuznetsova[1], V.N. Lasarev[2], V.S. Skvortsov[1], I.I. Karuzina[1], V.M. Govorun[2], A.I. Archakov[1]

[1]Orechovich Institute of Biomedical Chemistry, Russian Academy of Medical Sciences, 10 Pogodinskaya Str., Moscow, 119121, Russia, e-mail fox@ibmh msk su, fax +7-(095)-245-0857
[2]Institute of Physico-Chemical Medicine, 1a Malaya Pirogovskaya Str, Moscow, 119828 Russia

CYP51 family of cytochromes P450 (sterol 14-alpha-demethylases) comprises the representatives from different kingdoms of living world, thus positioning itself as the most ancient member of the superfamily In the course of the present research the collection of 36 full-length CYP51 amino acid sequences was submitted to cluster analysis Each node of the clustering dendrogram corresponds to the groups of proteins, located on the branches descending from the node By making the multiple alignment of each group of protein sequences we obtained the node-specific consensus sequences The informational content of the consensus was defined as the presence of the compact conserved sites, the motifs The assessment of informational content was computed using Sherman's non-parametric statistical criterion The high informational content was observed for the 100% conserved consensus sequences of the following CYP51,s groups fungi, animal+plant, plant+protista and bacteria These selected consensus sequences were next aligned all together to get the final consensus for the whole family To enrich the informational content of the CYP51 consensus the level of its conservation was dropped to 75% Regions of statistically significant conservation were unraveled in the CYP51 consensus sequence These regions (motifs) were then correlated with the information on secondary structure elements and substrate recognition sites reported for CYP51 from *Mycobacterium tuberculosis* Seven motifs appeared to be obligatory for every CYP51 protein The motifs thus obtained were searched for among all the known cytochrome P450 proteins Some motifs were found to be absolutely specific for 14-alpha-demethylases, whereas others were common to different species of cytochromes P450

Key Words 14-alpha-demethylase (CYP51), Mycobacterium, structural-functional motif, consensus sequence, substrate recognition site