

МИРОШНИЧЕНКО Юлиана Викторовна

**ОБЩЕЕ И ЧАСТНОЕ В СТРУКТУРНОЙ ОРГАНИЗАЦИИ БЕЛКОВ  
НАДСЕМЕЙСТВА ЦИТОХРОМОВ P450**

03.00.28 – биоинформатика

**АВТОРЕФЕРАТ**

диссертации на соискание ученой степени  
кандидата биологических наук

Москва 2006

Работа выполнена в Государственном учреждении Научно-исследовательском институте биомедицинской химии имени В.Н. Ореховича Российской академии медицинских наук

Научный руководитель:

кандидат биологических наук

Лисица Андрей Валерьевич

Официальные оппоненты:

доктор биологических наук, профессор

Коротков Евгений Вадимович

доктор биологических наук, профессор

Иванов Алексей Сергеевич

Ведущая организация:

Государственное учреждение Научно-исследовательский вычислительный центр Московского университета имени М.В. Ломоносова

Защита состоится «19» октября 2006 года в 11:00 часов на заседании Диссертационного совета Д 001.010.01 при ГУ НИИ биомедицинской химии им. В.Н. Ореховича РАМН по адресу: 119992, Москва, ул. Погодинская, 10.

С диссертацией можно ознакомиться в библиотеке ГУ НИИ биомедицинской химии имени В.Н. Ореховича РАМН по адресу: 119992, Москва, ул. Погодинская, 10.

Автореферат разослан «18» сентября 2006 года.

Ученый секретарь Диссертационного совета

кандидат биологических наук

Былинкина В.С.

## **1. Общая характеристика работы**

### **1.1. Актуальность проблемы**

Надсемейство цитохромов P450 представляет собой актуальный объект для исследования вычислительными методами. С практической точки зрения, интерес к изучению цитохромов P450 обусловлен ведущей ролью ферментов этой группы в метаболизме лекарственных препаратов и ксенобиотиков. Более 60% существующих ныне лекарств окисляются с участием цитохромов P450.

Моноксигеназная реакция, катализируемая цитохромами P450, заключается во внедрении в липофильную молекулу субстрата атома кислорода (1-й этап биотрансформации). В результате монооксигенирования растворимость окисленного вещества повышается и, после конъюгации, вещество выводится из организма. Таким образом, следует отметить, что цитохромы P450 играют значимую роль в обеспечении постоянства внутренней среды организма. Понимание гомеостатической роли цитохромов P450 может быть расширено за счет включения функции регулирования уровня гормонов: гемовые монооксигеназы надпочечников, простаты, щитовидной железы, эпителия ЖКТ участвуют как в синтезе, так и деградации гормонов.

Цитохромы P450 представляют собой потенциальные мишени для действия лекарственных препаратов. На основе ингибиторов цитохромов P450 создано целое поколение противогрибковых препаратов. Ведутся исследования в области компьютерного конструирования ингибиторов форм цитохромов P450 семейства CYP1A, чья активность сопряжена с развитием онкологических заболеваний, в частности, рака легких.

Уникальность реакции монооксигенирования, катализируемой цитохромами P450, обуславливает своеобразие молекулярно-эволюционных процессов в надсемействе. В настоящее время известно более 3 тыс. форм цитохромов P450, эти белки выявлены в геномах животных, растений, грибов. В геноме человека насчитывается 62 гена, кодирующих цитохромы P450, в геноме растений генов цитохромов P450 – более 200. Если функции более 70% цитохромов P450 человека известны, то для растений экспериментальной информации значительно меньше: изучена каталитическая функция менее чем для 5% растительных форм фермента.

Многообразие форм цитохромов P450, по мнению исследователей, является естественным депозитарием «заготовок» для использования в биотехнологии. Известны случаи, когда цитохромы P450 принимают участие в катализе реакций биосинтеза

противоопухолевых препаратов. Экстракты цитохромов P450 (микросомальные фракции тканей растений и животных) используются в качестве биореакторов для получения новых химических веществ. Интенсивно ведутся работы в области создания химерных форм цитохромов P450 с программируемой каталитической активностью.

Функциональное разнообразие надсемейства цитохромов P450 сочетается с существенными различиями в первичных структурах этих белков. Идентичность последовательностей аминокислотных остатков, входящих в надсемейство, в среднем составляет 26%. В то же время, все известные в настоящее время пространственные структуры цитохромов P450 характеризуются единообразным фолдом.

Начиная с 1989 года поддерживается систематическая номенклатура надсемейства цитохромов P450. В семейство выделяются белки, гомологичные на 40%; группы белков, гомологичные более чем на 46% объединяются в подсемейство. Наряду с формальными признаками сходства последовательностей, при создании номенклатуры авторы использовали дополнительную информацию о сходстве строения генов и об особенностях функциональной активности.

По-видимому, одним из основных недостатков существующей классификации следует считать ее искусственность. Пользуясь традиционной систематикой невозможно сделать выводы ни в отношении эволюционного, ни в отношении функционального сходства ферментов, входящих в семейства и подсемейства. Формально указанный недостаток выливается в списки исключений, указывающих на отнесение к классификационной группе белка, который в рамках определенного функционального контекста должен принадлежать другому подразделу классификации.

Неоднократно предпринимались попытки пересмотреть принципы классификации цитохромов P450. Предлагаемые решения основывались на применении методов кластерного анализа и множественного выравнивания последовательностей. В частности, основанный на иерархическом выравнивании подход инвентаризации надсемейства позволил получить консенсусную структуру цитохромов P450, объясняющую особенности строения белков надсемейства. При этом использовались методики, анализирующие первичную структуру в целом, без дифференцированного анализа составляющих её элементов. С другой стороны, уже в 1992 году было показано, что, несмотря на общее структурное разнообразие, в строении цитохромов P450 можно выделить локальные участки, несущие особую функциональную нагрузку. Это наблюдение до сих пор не потеряло своей актуальности: предложенное в 1992 г. понятие участков узнавания субстрата используется во многих работах, посвященных изучению структуры и функции цитохромов P450. В тоже время, обобщенная математическая

модель классификации надсемейства с учетом локальных структурно-функциональных элементов отсутствует, несмотря на то, что были показаны частные случаи успешного применения алгоритмов поиска мотивов для анализа надсемейства цитохромов P450.

Таким образом, актуальной является задача совершенствования алгоритмической методики выявления структурно-функциональных элементов (мотивов) в надсемействе цитохромов P450. Для решения этой задачи привлекается концепция наличия элементов общего и частного в структурах белков надсемейства. В основу концепции легли положения «островной гипотезы», рассматривающей термодинамические ограничения белкового фолдинга.

«Островная» гипотеза строения белков основывается на предпосылке о незначительной доле термодинамически выгодных конформаций белка по отношению к общему количеству гипотетических последовательностей, которые можно получить из 20-ти аминокислотных остатков. Следствием этой предпосылки является неоднородность участков первичной структуры с точки зрения их вклада в обеспечение пространственной конформации белка.

**Целью работы** являлась разработка подхода для выявления в аминокислотных последовательностях белков надсемейства цитохромов P450 формальных элементов, определяющих структурную общность и функциональную специфичность различных форм этого фермента. В рамках достижения указанной цели решались следующие **задачи**:

1. Предложить алгоритмический метод выявления структурно-функциональных мотивов в группе белков (цитохромов P450) и исследовать его свойства.
2. Выявить мотивы структурной общности для всего надсемейства белков.
3. Выявить мотивы частного в отдельной группе функционально родственных белков.
4. Разработать методику классификации цитохромов P450 с учетом структурно-функциональных мотивов и сравнить результаты с традиционной номенклатурой.

## **1.2. Научная новизна и практическая значимость**

В работе впервые рассматриваются принципы формализации подходов к классификации надсемейства цитохромов P450. Предложена алгоритмическая методика для оптимизации результатов кластерного анализа на основе структурно-функциональных мотивов, выделяемых в первичной структуре кластеризуемых белков. Показано, что

выявляемые в группах белков консервативные элементы соответствуют участкам белка, имеющим значение для обеспечения общности структуры фолда и для реализации специфичной ферментативной активности.

Полученные результаты могут быть использованы для прогнозирования функций новых белков надсемейства цитохромов P450. В практическом плане, созданная методология может быть применена для решения актуальных задач биотехнологии: синтез новых химических соединений и конструирование ферментов-монооксигеназ с заданной функцией.

### **1.3. Апробация работы**

Основные положения диссертационной работы докладывались и обсуждались в ходе следующих конференций:

- Российский Национальный Конгресс «Человек и лекарство» (Москва, 2003);
- 13-я Интернациональная конференция по цитохромам P450 (Прага, 2003);
- 2005 г. «Системная биология и биоинженерия» (Звенигород).

### **1.4. Публикации**

Материалы диссертационной работы отражены в 5 публикациях, из них: 3 в общероссийских рецензируемых изданиях, 1 публикация в сборнике трудов международной научной конференции, 1 тезисы доклада.

### **1.5. Объем и структура диссертации**

Диссертация изложена на 126 страницах машинописного текста, включая 15 таблиц, 27 рисунков. Состоит из глав: «Введение», «Обзор литературы», «Материалы и методы», «Результаты и обсуждение», «Выводы», «Список литературы».

### **1.6. Основные положения, выносимые на защиту**

1. Предложенный алгоритмический метод, основанный на применении статистики Шермана к результатам множественного выравнивания, позволяет выявлять в наборах последовательностей цитохромов P450 локальные консервативные участки (мотивы).
2. Выявляемые участки локального сходства отвечают элементам структуры цитохромов P450, обеспечивающим его общую пространственную организацию (мотивы общности) и/или специфическую функциональную активность (мотивы частного).
3. Учет структурно-функциональных мотивов позволяет корректировать результаты кластерного анализа, что повышает уровень соответствия состава формальных групп (кластеров) с общепринятыми номенклатурными подразделами – семействами.

## **2. Материалы и методы**

### **2.1. Выборка**

С использованием базы данных по цитохромам P450 была сформирована выборка, в которую вошли представители различных семейств цитохромов P450. В выборку включались только те семейства, для которых известно не менее 5 форм цитохромов P450, причем для каждой формы известна полноразмерная последовательность аминокислотных остатков ( $450 \pm 50$  аминокислотных остатков для микросомальных и митохондриальных форм,  $350 \pm 50$  для бактериальных форм). Состав сформированной выборки приведен в таблице 1.

**Таблица 1.** Выборка белков надсемейства цитохромов P450.

	Животные	Растения	Грибы	Бактерии	<b>ВСЕГО</b>
Семейства	21	26	6	3	<b>56</b>
Подсемейства	84	78	27	31	<b>220</b>
Белки	425	382	61	39	<b>907</b>

Средняя идентичность последовательностей в сформированной выборке составляла  $24 \pm 3\%$ .

Исследование мотивов частного проводилось на выборке цитохромов P450 семейства стероловых деметилаз CYP51. Всего в эту выборку вошло 36 последовательностей, из них: 6 форм животного происхождения, 5 - растительного, 20 - низшие грибы, 4 - бактерии и 1 - простейшие. Средняя идентичность последовательностей в семействе составляет  $35\pm 7\%$ . Для анализа полученных результатов использовались данные о пространственной структуре цитохрома P450 из микобактерии туберкулеза (CYP51MT, код PDB - 1E9X).

## 2.2. Локальное выравнивание

Для проведения локального выравнивания использовалась программа BLAST. При расчете гистограмм счетов применялась версия программы, установленная на сервере ГУ НИИ БМХ РАМН, поддерживающая режимы пакетной обработки запросов. Оценка локального сходства между парой последовательностей проводилась при помощи Интернет-версии программы bl2seq. Во всех случаях использовались следующие параметры локального выравнивания:

- подпрограмма: blastp, матрица замен: BLOSUM62
- штраф за открытие вставки: 11, штраф за продолжение вставки: 1, длина слова: 3
- ограничение на вероятность случайного совпадения (expectation): 10.0

## 2.3. Кластерный анализ

Кластерный анализ проводился методом невзвешенных средних на основе матрицы попарных сходств между последовательностями белков. Сходство между парой последовательностей оценивалось при помощи идентичности глобального выравнивания. Глобальное парное выравнивание рассчитывалось с использованием программы ALN со следующими параметрами:

- матрица замен: BLOSUM62;
- штраф за открытие вставки:  $8\pm 3$ , штраф за продолжение вставки  $4\pm 2$ ; конкретные значения подбирались с использованием случайных последовательностей.

*Определение границ кластеров.* Для определения границ кластеров использовалось унифицированное правило, применимое для всех кластеров в составе надсемейства. В зависимости от специфики задачи использовалась одна из следующих методик:

- Метод «колена» – оценка зависимости скорости агломерации (образования кластеров) от шага кластерного анализа;
- Индекс Джаккарда – оценка соответствия состава кластеров с семействами и подсемействами, выделяемыми согласно номенклатурной систематике надсемейства. Оптимальным считается уровень отсечения, при котором достигается наилучшее соответствие между формальными кластерами и номенклатурными группами – семействами и подсемействами.
- Коррекция границ кластеров с использованием критерия структурно-функциональных мотивов осуществляется при помощи адаптированного статистического критерия Шермана.

#### 2.4. Иерархическое выравнивание

В основе процедуры иерархического выравнивания, применяемого в работе, лежит алгоритм множественного выравнивания, последовательно применяемый к группам сходных последовательностей. Группы формируются методом кластерного анализа. В данной работе использовался итерационный метод оптимизации множественного выравнивания, реализованный в программе PRRP. Использовалась матрица замен PAM250, дополненная специальными символами, обозначающими группы аминокислотных остатков (табл. 2). Группировка остатков производилась путем кластерного анализа матрицы замен PAM250. Алгоритм множественного выравнивания был адаптирован для работы со специальными символами, что позволило выравнивать не только последовательности аминокислотных остатков, но и консенсусные последовательности.

**Таблица 2.** Специальные символы, используемые для обозначения групп аминокислотных остатков

Символ*	[a]	[+]	[=]	[n]	[s]
Состав группы	FYW	KRH	DEQN	LIMV	ASTG

\* специальный символ [.] использовался для обозначения любого аминокислотного остатка

Результаты множественного выравнивания представлялись в виде консенсусной последовательности. Консенсусная последовательность содержит консервативные остатки, характерные для большинства (или для всех) последовательностей в множественном выравнивании. Изменяя уровень консервативности, регулируется состав

консенсуса, т.е. – количество выносимых в консенсусную строку консервативных символов. Плотность консенсусной последовательности рассчитывалась как отношение числа консервативных позиций к общей длине консенсуса. Оптимизация результатов множественного выравнивания достигалась путем 10-15-ти кратного перезапуска алгоритма при варьировании трех параметров: штраф за открытие вставки, штраф за продление вставки и порядок следования выравниваемых последовательностей. Рандомизация параметров проводилась до тех пор, пока плотность консенсусной последовательности не достигала максимального значения и больше не менялась.

В работе используется общепринятая номенклатура для обозначения систематических идентификаторов форм цитохромов P450. После префикса “СУР” указывается числовое обозначения семейства, далее следует буквенное обозначение подсемейства, затем – числовое обозначение гена. После систематического идентификатора может указываться видовая специфичность формы.

## 2.5. Выявление мотивов в консенсусной последовательности

Выявление структурно-функциональных мотивов в составе консенсусной последовательности производилось при помощи статистического критерия Шермана. Указанный критерий позволяет для заданной битовой строки (см. рис. 1) получить оценку вероятности, описывающую характер распределения значений (нулей или единиц) в строке. Большие значения вероятности отражают тенденцию группировки значений в компактные кластеры.

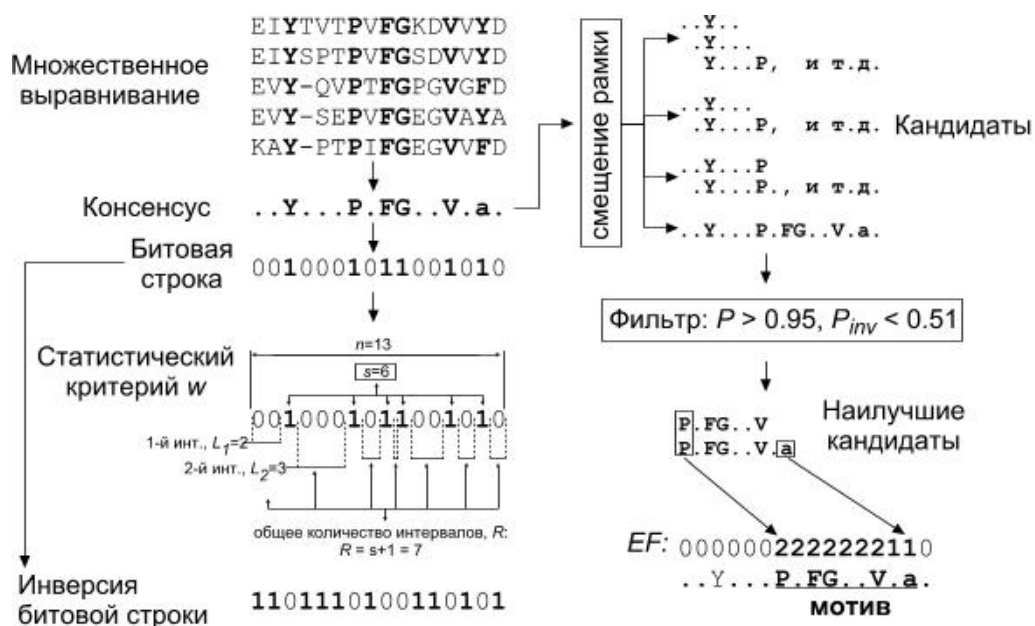
На первом шаге алгоритма, консенсусная строка перекодировалась в битовое представление, согласно которому 0 соответствовал варибельной позиции, а 1 соответствовала консервативной позиции консенсуса. С использованием заданного окна (как правило, его размер брался равным 20 остаткам), консенсус разбивался на перекрывающиеся подстроки. Для каждой подстроки вычислялась оценка вероятности  $P\{w > w_{est}\}$ ; где критерий  $w$  рассчитывается по формуле:

$$w = 0,5 \sum |L_k - \frac{1}{d+1}| \quad (1),$$

а  $w_{est}$  – оценка для случайного характера распределения значений в битовой строке.

Кроме того, битовая подстрока инвертировалась и для инвертированной строки рассчитывалась оценка  $P_{inv}$ .

Из всего множества подстрок, сгенерированных на основе консенсусной последовательности, отбирались удовлетворяющие критериям:  $P > 0,95$  и  $P_{inv} < 0,51$ . Для каждой позиции  $i$  в консенсусе рассчитывался счет  $S_i$  вхождения в состав мотива как количество раз, которое указанная позиция встречается в отобранном наборе подстрок. Непрерывные последовательности отличных от нуля оценок длиной более 5 символов рассматривались как мотивы.



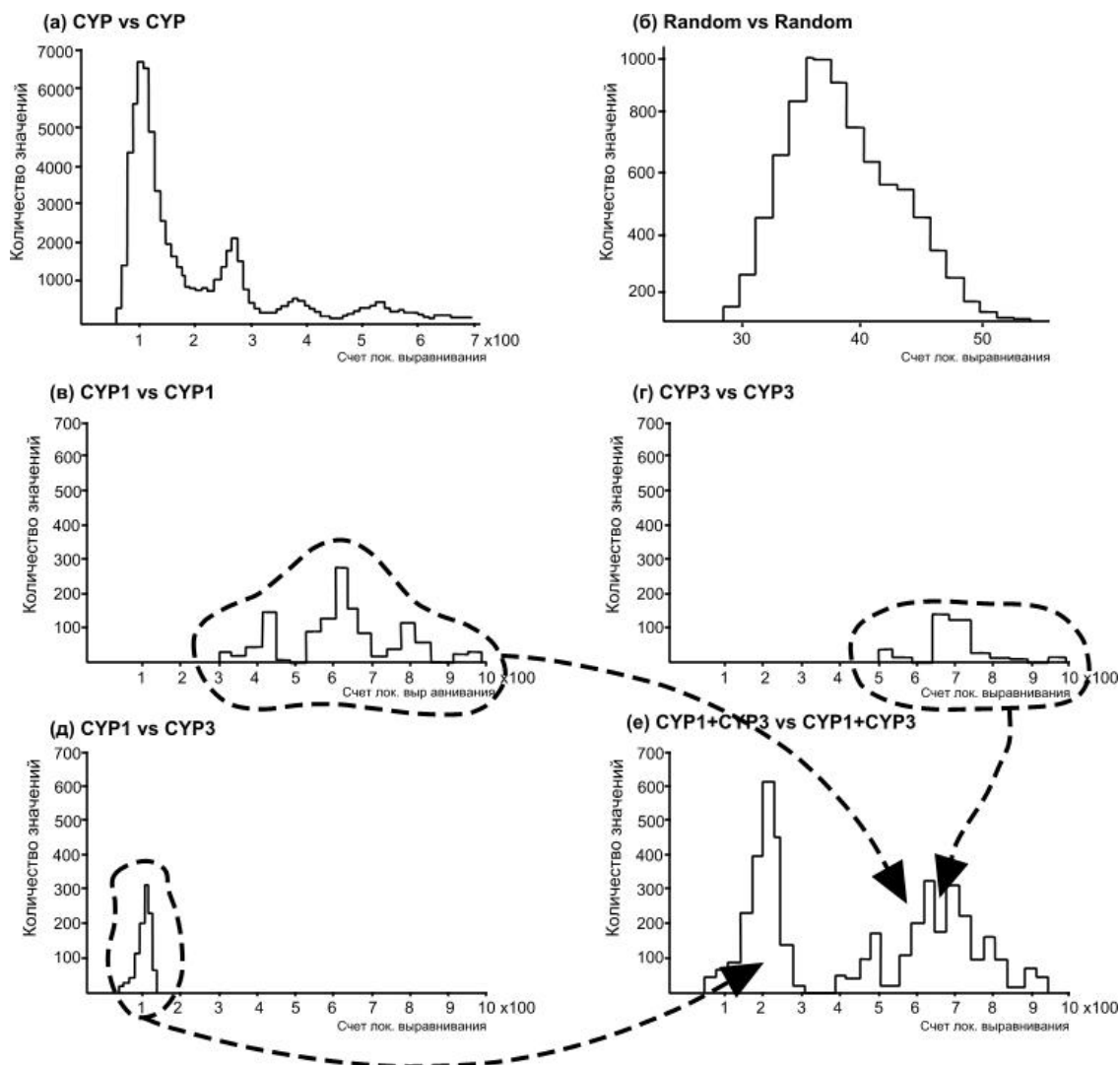
**Рис. 1.** Алгоритм выявления структурно-функциональных мотивов в составе консенсусной последовательности.

### 3. Результаты и обсуждение

#### **3.1. Предпосылки наличия мотивов в надсемейства и семействах цитохромов P450**

В основу данной работы легли результаты, полученные в ходе анализа надсемейства цитохромов P450 при помощи программы локального выравнивания BLAST. Сущность эксперимента, результаты которого приведены на рис. 2, заключается в проведении сравнения всех последовательностей в отобранной группе белков друг с другом (кросс-сравнение). В качестве группы может фигурировать как все надсемейство целиком, так и отдельные подгруппы белков в его составе – семейства.

По результатам сравнения групп белков строятся гистограммы распределения счетов локального выравнивания.



**Рис. 2.** Гистограммы распределения счетов локального выравнивания, полученных в результате кросс-сравнения выборок: (а) надсемейство, (б) случайно сгенерированные последовательности символов, (в) семейство CYP1, (г) семейство CYP3, (д) семейство CYP1 с семейством CYP3, (е) группа CYP1+ CYP3.

Результаты, представленные на гистограмме 2а свидетельствуют о том, что для большинства пар последовательностей надсемейства характерны участки локального сходства. Счет, вычисляемый в результате выравнивания этих локальных участков, в среднем равен 100 битам, что определяет положение наиболее выраженного пика на гистограмме. Минимальное значение счета, получаемого при выравнивании цитохромов P450, соответствует 50-60 битам, что обеспечивает возможность достоверно отличить членов надсемейства от случайно сгенерированных строк (см. рис. 2б).

Кроме наиболее выраженного, первого пика, гистограмма на рис. 2а содержит три минорных пика. Наличие минорных пиков нарушает общий вид кривой, описывающей статистическое распределение счетов локального выравнивания. Для объяснения наблюдаемого феномена необходимо обратиться к гистограммам 2в-2е.

На рис. 2в и 2г приведены гистограммы счетов локального выравнивания, накопленных в результате кросс-сравнения последовательностей семейств СУР1 и СУР3. Оба семейства насчитывают практически одинаковое количество белков, при этом средняя идентичность последовательностей семейства СУР3 ( $68,9 \pm 8,6\%$ ) превышает такую последовательностей семейства СУР1 ( $55,4 \pm 16,5\%$ ).

Сравнение структурно близких белков, входящих в одно семейство, приводит к повышению счета локального выравнивания, как это показано на гистограммах 2в и 2г смещением пиков гистограммы к отметкам 600 и 700 бит соответственно. Интересно, что смещение в большую сторону (к отметке 700) наблюдается для семейства СУР3, характеризующегося более высоким сходством входящих в него белков.

В то же время, в результате кросс-сравнения семейств друг с другом наблюдается единственный пик, соответствующий отметке 100 бит. Очевидно, что этот пик определяет элементы сходства, общие для всего надсемейства в целом. Этот пик, соответствует первому пику на гистограмме 2а.

Рис. 2е позволяет утверждать, что локальные элементы сходства, присутствующие в определенных группах в составе надсемейства цитохромов Р450, приводят к отклонению гистограммы распределения от канонической формы распределения экстремальных значений. Кроме доминирующего пика 100 бит, присутствуют дополнительные пики, обусловленные элементами сходства членов семейств СУР1 и СУР3.

Таким образом, на основе анализа гистограмм распределения счетов локального выравнивания в надсемействе и отдельных семействах можно сделать вывод о наличии в структуре последовательностей цитохромов Р450 участков локального сходства, которые в дальнейшем обозначаются как *мотивы*. Причем можно выделить две группы мотивов: мотивы общности, характерные для всех белков надсемейства, и мотивы частного, присутствующие только в отдельных семействах.

Рассмотренные выше доказательства наличия мотивов в надсемействе цитохромов Р450 основываются на традиционной систематике этих белков. В то же время, правомочна обратная формулировка проблемы, в рамках которой предлагается найти такое разделение надсемейства белков на группы, для которого некая интегральная оценка локального сходства будет максимальна. Проблема может быть решена при наличии алгоритма,

способного оценивать локальное сходство для групп белков, а не для выбранной пары белков (как это делается при построении гистограмм). Одновременно, следует решить задачу непосредственного выявления мотивов (а не только констатации факта их присутствия), для того чтобы иметь возможность сравнить их с имеющимися экспериментальными данными о структурно-функциональных особенностях цитохромов P450.

### 3.2. Алгоритм выявления структурно-функциональных мотивов

Метод выявления мотивов путем статистического анализа характера распределения консервативных остатков в составе консенсусной последовательности обладает существенным ограничением, не позволяющим применять его в вышеизложенной форме для решения задач данной работы. Таким ограничением является невозможность применения критерия для сравнения между собой консенсусов, полученных для различных групп белков, или при использовании различных параметров (уровня консервативности, типа используемого редуцированного алфавита). Для снятия указанного ограничения в работе предлагается метод, основанный на оценке информационного содержания консенсусной последовательности.

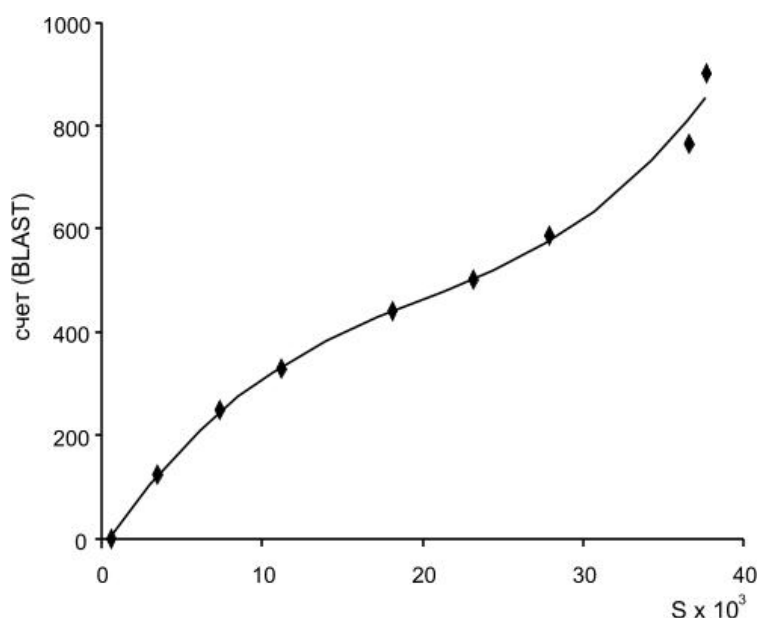
Основы этого подхода заложены в алгоритмической схеме, рассмотренной выше. Действительно, для каждой подстроки в составе консенсуса рассчитывается два критерия:  $P$  и  $P_{inv}$ . Критерий  $P$  определяет насколько компактно расположены в консенсусе консервативные остатки, критерий  $P_{inv}$  – наоборот, отражает компактность расположения переменных остатков. Применение фильтра, учитывающего оба критерия, приводит к тому, что алгоритм выявляет в составе консенсусной последовательности компактные кластеры консервативных остатков ( $P > 0,95$ ) и при этом запрещает в составе этих кластеров наличие протяженных слитных участков переменности ( $P_{inv} < 0,51$ ). В случае наличия таких участков, установленные пороговые значения критериев  $P$  и  $P_{inv}$  разделяют подстроку на два мотива.

Особенности вычислительной схемы позволяют использовать ее для выявления участков локального сходства. Учитывая, что в составе консенсусной последовательности могут присутствовать несколько участков локального сходства, введем меру  $S$  как сумму значений, которые принимает счет ( $S_i$ ) в каждой позиции консенсусной последовательности:

$$S = \sum S_i \quad (2)$$

Следует отметить наличие взаимосвязи между стандартным методом оценки локального сходства, используемого в программе BLAST, и предложенной суммарной оценкой  $S$ . На рис. 3 показана корреляция между значениями  $S$  и счетом локального парного выравнивания, рассчитанного программой BLAST. Для расчета величины  $S$  использовались консенсусные последовательности, полученные при задании различных уровней консервативности. Расчет битового счета локального выравнивания производился путем выравнивания консенсусной последовательности с самой собой, при этом переменные участки заменялись случайно сгенерированными символами.

Взаимосвязь между счетом локального выравнивания и предлагаемым критерием для выявления мотивов в консенсусной последовательности, отображенная на рис. 3, свидетельствует об адекватности реализованного алгоритмического решения. В то же время, принципиальным отличием предложенного критерия от стандартной процедуры вычисления счета локального выравнивания является возможность разложить его на индивидуальные потенциалы, приписанные каждому остатку в консенсусе. Оценка  $S_i$  зависит от окружения остатка  $i$ , точнее от «предрасположенности» этого окружения к формированию участка локального сходства – т.е. мотива.



**Рис. 3.** Зависимость между счетом локального выравнивания (BLAST) и оценкой мотивов ( $S$ ).

Представление результатов множественного выравнивания в виде консенсусной последовательности используется довольно часто. Цель использования консенсусов

заключается в том, чтобы искусственно снизить сложность природных белковых последовательностей, сравнивая их друг с другом и вычлняя общую часть. Предполагается, что эта общая часть более содержательна в информационном плане, нежели любая другая отдельно взятая часть первичной структуры белка. Основанием для этого предположения являются молекулярно-эволюционные гипотезы, в рамках которых предполагается, что функционально значимая часть гена наиболее устойчива к спонтанным мутациям.

В данной работе базовые постулаты молекулярной эволюции развиваются в рамках оригинальной математической модели. Основу модели составляет стихийно сложившееся интуитивное представление, бытующее в литературе, о наличии взаимосвязи между понятиями информационного содержания первичной структуры белка и соответствующей функциональной активностью (в частности, речь идет о широко используемом постулате, что структура белка определяет его функцию). В более широком смысле, вместо функциональной активности используется понятие термодинамических барьеров фолдинга. В практическом плане постулат «структура определяет функцию» может трактоваться в том смысле, что элементы структуры с высоким содержанием информации с большей вероятностью определяют особенности функционирования белка. С этой точки зрения использование консенсусных последовательностей является попыткой вычлнить наиболее информационно-насыщенную часть в наборе первичных структур белков.

В работе предлагается для оценки информационного содержания консенсусной последовательности использовать величину  $S$ , отражающую присутствие локальных участков выраженной гомологии (мотивов). Однако для того, чтобы оценить информационное содержание консенсусной последовательности, следует использовать не только величину  $S$ , но и комплиментарную ей величину:

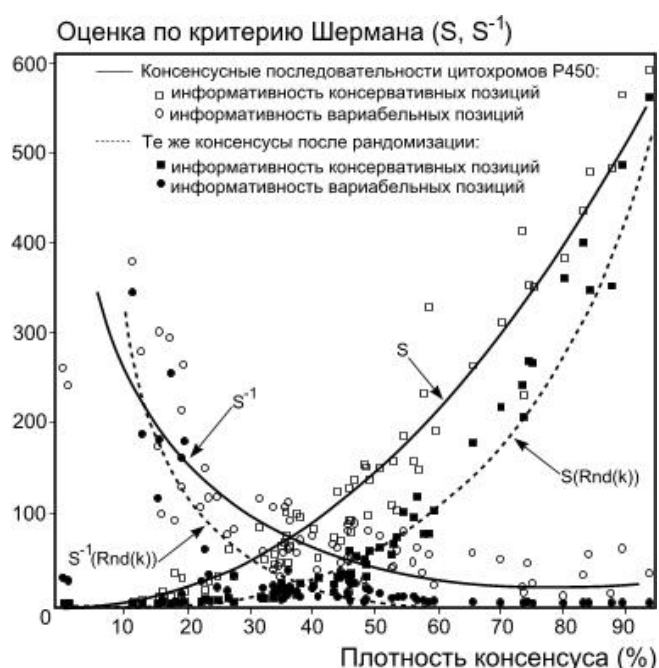
$$S^{-1}(bs) = S(\overline{bs}) \quad (3)$$

где  $bs$  обозначает битовую строку (консенсус). Введение комплиментарной величины  $S^{-1}$ , рассчитываемой для инвертированного консенсуса ( $\overline{bs}$ ), обусловлено тем, что редукция информации в форме консенсусной последовательности только до некоего порогового уровня, после которого консенсус вырождается в строку, состоящую исключительно из переменных позиций. Пороговое значение вводится как равенство величин  $S$  и  $S^{-1}$ . При этом достигается максимальное значение оценки информационного содержания консенсусной последовательности  $I$ , рассчитываемое по формуле:

$$I = S/S^{-1}, \quad S < S^{-1}$$

$$I = S^{-1}/S, \quad S \geq S^{-1} \quad (4)$$

Формула (4) задает область определения величины  $I$  от 0 до 1. Значения близкие к нулю соответствуют двум случаям: консенсус перегружен консервативными остатками (гиперконсервативный консенсус) или консенсус перегружен переменными позициями (гипервариабельный консенсус). Значения оценки информационного содержания близкие к единице отражают оптимальное состояние консенсуса с точки зрения наличия в нем структурно-функциональных мотивов.

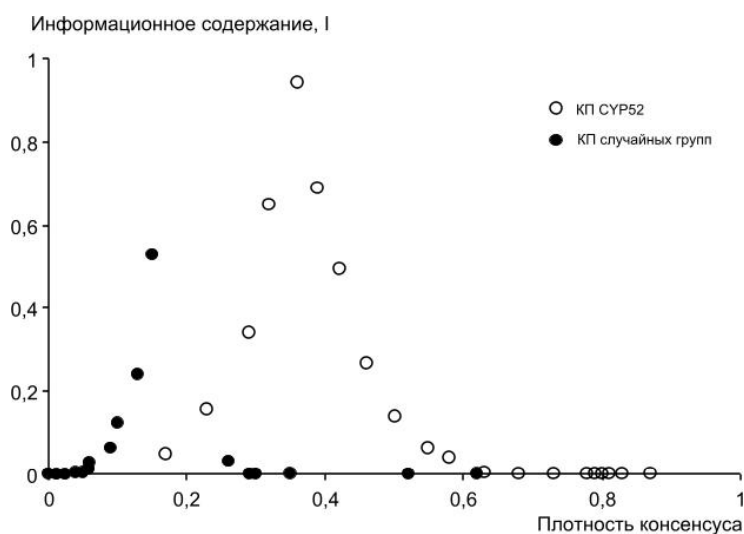


**Рис. 4.** Интегральная оценка  $S$  наличия мотивов в консенсусной последовательности и комплементарная ей величина  $S^{-1}$  в зависимости от плотности консенсуса.

На рис. 4 представлены результаты, полученные путем расчета параметров  $S$  и  $S^{-1}$  для консенсусов различных семейств и подсемейств цитохромов P450.

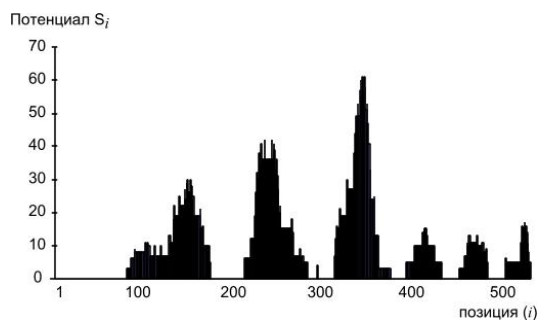
Величина оценки  $S$  принимает максимальные значения при 100% плотности консенсуса. Плотность 100% означает, что консенсус состоит только из консервативных позиций, переменные участки отсутствуют. Такая ситуация может возникать при высокой степени гомологии между анализируемыми последовательностями, что не позволяет извлечь информацию о функционально важных участках – т.е. информационное содержание такого консенсуса стремится к нулю. По мере того как увеличивается степень разнообразия исследуемого набора первичных структур, падает плотность консенсуса и соответственно снижается величина  $S$ . Снижение величины  $S$  монотонно происходит до

нуля, что приблизительно соответствует 10% уровню плотности консенсуса. Последний, при этом, представляет собой последовательность переменных позиций и так же не несет содержательной информации.



**Рис. 5.** Зависимость оценки информационного содержания от плотности консенсусной последовательности. Данные получены для семейства СУР52 при варьировании уровне консервативности консенсуса.

Рассматривая возрастание величины  $S$  по мере увеличения плотности консенсуса можно отметить, что с точки зрения информационного содержания процесс на определенном уровне достигает максимума величины  $I$ , а затем она неуклонно снижается (см. рис. 5). Формализовать это наблюдение удастся, если привлечь величину  $S^{-1}$ , которая увеличивается по мере уменьшения плотности консенсусной последовательности (рис. 4). Предлагается считать (и правомочность этого допущения доказывается в данной работе), что баланс между величинами  $S$  и  $S^{-1}$  соответствует максимально информативному консенсусу. В его составе можно выделить локальные мотивы, используя значения потенциала  $S_i > 0$ , где  $i$  – номер позиции в консенсусной последовательности (рис. 6).



**Рис. 6.** Значения потенциала  $S_i$  для позиций консервативных последовательностей семейства СУР51. Пики указывают расположение и границы структурно-функциональных мотивов.

Для сравнения оценок информационного содержания, полученных для консенсусов разных групп, вводится нормировочный фактор, отражающий отношение между величиной  $S(k)$ , полученной для исследуемого консенсуса  $k$ , и  $S(rnd(k))$ , полученной для консенсуса, в котором порядок следования консервативных и переменных позиций нарушен случайным образом (при этом, очевидно, плотность консенсуса не меняется):

$$I_{abs} = I * S(k) / S(rnd(k)) \quad (5)$$

### 3.3. Мотивы общности

Для выявления мотивов общности использовалась процедура инвентаризации в алгоритмической реализации. Методика инвентаризации включает в себя кластерный анализ, определение границ кластеров, построение консенсусов для кластеров методом множественного выравнивания при заданном уровне консервативности 75%, и построение общего консенсуса для надсемейства путем множественного выравнивания консенсусов кластеров. В данной работе, с использованием алгоритма оценки информационного содержания консенсуса с точки зрения наличия структурно-функциональных мотивов, был пересмотрен критерий оценки уровня консервативности консенсусных последовательностей.

Ранее использовался фиксированный критерий, установленный на уровне 75%. Указанное значение применялось для всех семейств и подсемейств, без учета особенностей включаемых в них последовательностей цитохромов P450. Негативным эффектом унификации уровня консервативности консенсуса являлось то, что информация о значимых структурно-функциональных элементах выпадала из состава консенсуса. С другой стороны, в ряде случаев консенсусы (особенно уровня подсемейства) характеризовались высокой плотностью, что в дальнейшем приводило к разбалансировке множественного выравнивания на следующем уровне иерархии.

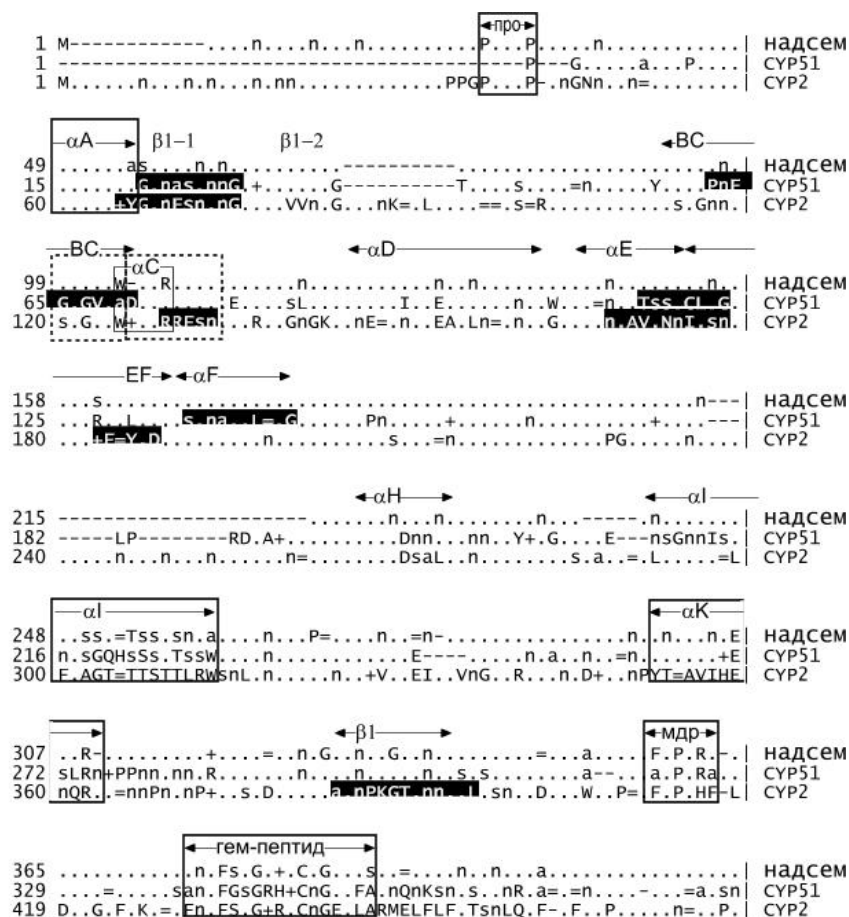
Особый случай представляли собой номенклатурные подгруппы (кластеры), включающие только один белок (т.н. синглтоны). В отсутствие формальных критериев оценки информационного содержания, в предыдущих работах единственная последовательность подгруппы фигурировала в качестве полноценной анализируемой единицы наряду с действительными консенсусами, полученными для подгрупп из нескольких белков.

В ходе процедуры инвентаризации надсемейства, результаты которой рассматриваются в данной работе, синглетоны обрабатываются следующим образом: каждый из уже построенных консенсусов, имеющий наивысшую возможную информативность, поочередно выравнивается с последовательностью-синглетоном. На основании проведенных попарных сравнений строятся вторичные консенсусы, из которых отбирается тот, который обладает наибольшим информационным содержанием. Этот консенсус в дальнейшем используется в качестве представителя синглетона для проведения множественного выравнивания на следующем уровне иерархии надсемейства.

Результаты проведенной инвентаризации отражены на рис. 7 в виде схематического изображения консенсусной последовательности надсемейства цитохромов P450. В его составе можно выделить следующие элементы общности.

В консенсусе надсемейства присутствуют пять мотивов, характерных для всех цитохромов P450. К ним относятся, альфа-спирали С, I, К, меандр (извилина) и гем-пептид. Так называемая триада ERR является элементом спирали К. Второй аргинин, входящий в состав триады совместно с извилиной образуют сетку водородных связей, обеспечивающую пространственную организацию консервативных участков при формировании белкового фолда. В свою очередь, меандр, в ансамбле со спиралью I и гем-пептидом принимает участие в фиксации гема. Спираль I сочетает структурную роль (фиксацию гема) и функциональное назначение – фиксацию молекулярного кислорода в относительной близости от каталитического центра (атом железа в составе гема). Таким образом, мотивы общности, входящие в С-концевую часть консенсуса надсемейства обеспечивают формирование фолд-детерминирующего ядра цитохромов P450 и поддерживают функцию монооксигеназного катализа. Рассмотренные выше мотивы общности проявляют наибольшую консервативность в ходе молекулярной эволюции белков надсемейства цитохромов P450.

N-концевая часть консенсуса свидетельствует о значительной вариабельности этого участка структуры цитохромов P450. Единственный выявленный мотив приходится на центр спирали С, и может быть описан в виде универсального паттерна как [WH]xxR. Однако, детальное исследование N-концевого фрагмента консенсуса позволяет выявить еще два дополнительных консервативных элемента, которые, в силу своей незначительной протяженности формально не фигурируют в качестве структурно-функциональных мотивов. Среди них – пролиновый кластер, выявляемый только в структуре микросомальных цитохромов P450 формирует узнаваемую сигнатуру RxxP. Функционально этот кластер отграничивает жесткой конструкцией белковую глобулу от трансмембранного якоря.



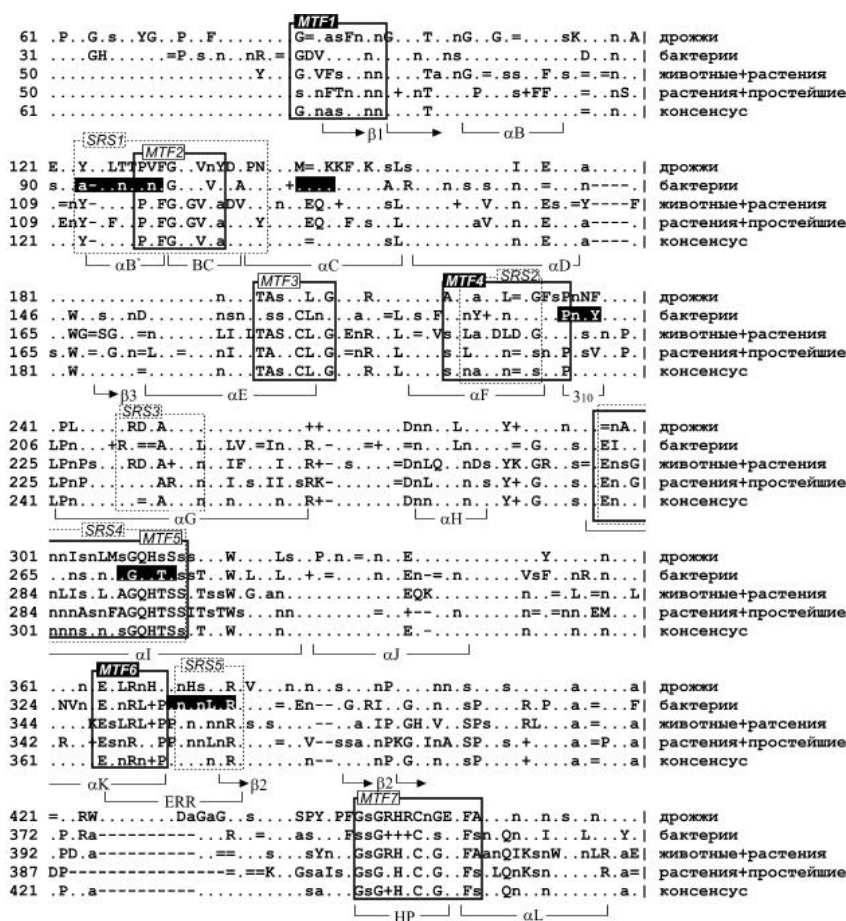
**Рис. 7.** Консенсус надсемейства цитохромов P450, полученный в результате оценки информационного содержания результатов множественного выравнивания в ходе процедуры инвентаризации. мдр – меандр; про – пролиновый кластер. В выравнивание также включены консенсусные последовательности для семейств CYP2 и CYP51.

### 3.4. Мотивы частного в семействе стероловых деметилаз

Семейство стероловых деметилаз (CYP51) уникально для надсемейства цитохромов P450. Особенность этой группы белков заключается в том, что его члены встречаются в представителях всех царств живой природы. При этом, следует отметить высокую консервативность входящих в семейство CYP51 белков и высокую субстратную специфичность, ориентированную исключительно на метаболизм соединений стеролового ряда.

Анализ семейства проводился в соответствии с общей схемой инвентаризации. Результаты кластерного анализа показали, что в составе семейства присутствуют 5 групп белков. В группы объединились белки, принадлежащие к одному царству. Для каждой группы было проведено множественное выравнивание и построены соответствующие

консенсусные последовательности. С целью выравнивания информационного содержания консенсусных последовательностей, границы кластеров были скорректированы, в результате чего возникли следующие группы: грибы, бактерии, растения+простейшие и животные+растения. Консенсусы групп были в свою очередь выровнены между собой и скомпонованы в единый консенсус семейства СУР51. Одинаковый уровень информационного содержания составляющих консенсусов обеспечивает равный вклад каждого белка семейства стероловых деметилаз в общую консенсусную последовательность.



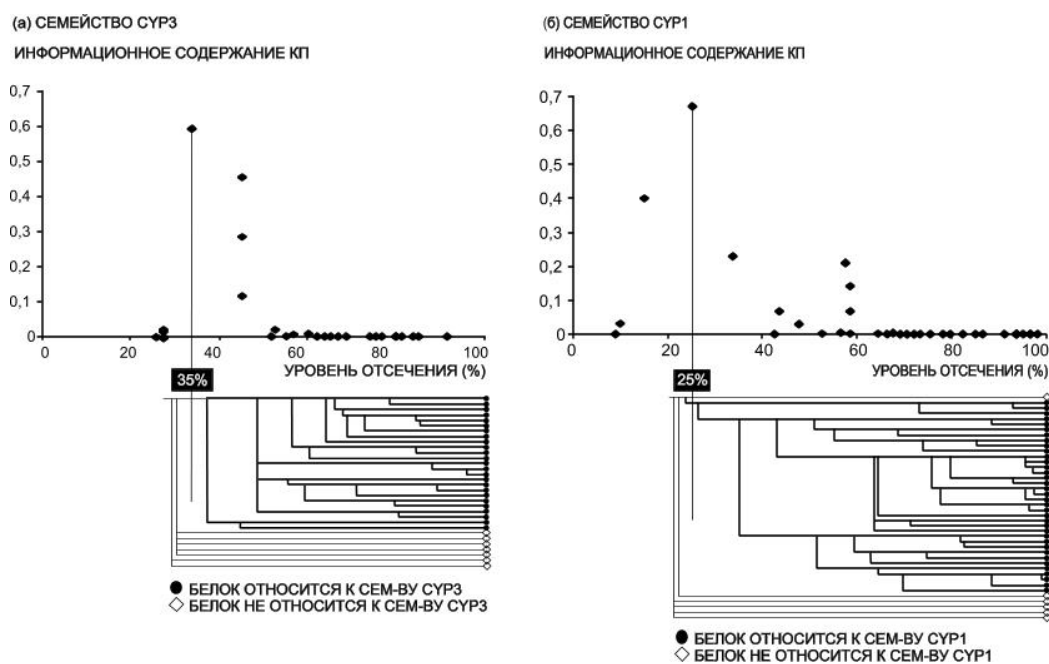
**Рис. 8.** Мотивы частного в составе структурно-функциональной карты семейства стероловых деметилаз. MTF – мотив; SRS – участок узнавания субстрата.

Структурно-функциональная карта семейства СУР51 приведена на рис. 8. На рисунке отображена не только общая консенсусная последовательность, но и составляющие консенсусы сформированных групп белков. Отмечены мотивы среди которых путем сравнения с литературными данными можно выделить 2 типа. К первому принадлежат участки, обеспечивающие функциональную специфичность стероловых деметилаз – MTF1, MTF2, MTF3, MTF4. Мотивы же, обозначенные как MTF5- MTF7,

совпадают с мотивами консенсуса надсемейства (см. рис. 7), т.е. отвечают элементам общности.

### 3.5. Использование критерия оценки мотивов для корректировки границ кластеров

В ходе кластерного анализа, белки объединяются в группы (кластеры) на основе меры сходства. На каждом шаге либо новые белки вливаются в состав уже существующего кластера, либо образуются новые кластеры. Как только в ходе кластеризации обособляется группа белков, можно провести множественное выравнивание соответствующих последовательностей аминокислотных остатков, построить консенсус и оценить уровень его информационного содержания. Таким образом, каждый шаг кластерного анализа может быть сопоставлен с соответствующей величиной информационного содержания (рис. 9).



**Рис. 9.** Уточнение границ кластеров семейств СУР1 и СУР3 с использованием оценки информационного содержания консенсусных последовательностей. КП – консенсусная последовательность.

По мере того, как белки объединяются в группы, информационное содержание консенсуса этих групп возрастает. Прохождение каждого нового узла дендрограммы приносит новые объекты к кластеру, что снижает плотность консенсуса и увеличивает его информационное содержание. На одном и том же шаге агломерации может

образоваться несколько новых кластеров, соответственно на диаграмму зависимости информационного содержания от шага кластеризации наносится ряд точек, количество которых соответствует числу узлов. Возрастание информационного содержания кластера продолжается до тех пор, пока в его состав не войдут «чужеродные» структуры, в которых отсутствуют мотивы, специфические для генов кластера. Это приводит к снижению оценки информационного содержания и служит критерием останова агломерации.

Уровень останова неодинаков для различных семейств. Так, в случае семейства СУР3 максимум информационного содержания приходится на кластер, формирующийся на шаге кластеризации уровня 35% средней идентичности. Для семейства СУР1 этот порог на 20% ниже и составляет 15%.

Рассмотренный частный пример позволяет заключить, что не существует универсального порога кластеризации для всего надсемейства цитохромов Р450. В каждом случае границы кластеров (семейств) подлежат уточнению с учетом структурных особенностей объединяемых белков. Если для уточнения границ кластеров используется предлагаемый критерий, основанный на оценке наличия общих структурно-функциональных мотивов в кластере, то в этом случае удастся добиться существенного повышения уровня соответствия между составом кластеров и традиционными номенклатурными подгруппами – семействами (табл. 3).

*Табл. 3. Значения критерия соответствия Джаккарда, полученные при нахождении уровня отсечения различными методами.*

Метод нахождения границ кластеров	Уровень отсечения (средний % идентичности белков в кластере)	Соответствие состава кластеров и семейств, % (индекс Джаккарда)
индекс Давида-Болдина	39	67
наилучшее совпадение с номенклатурой	35	80
метод «колена»	39	68
критерий структурно-функциональных мотивов	15-43	84

Видно, что предлагаемый критерий останова значительно превосходит по показателю соответствия состава стандартные подходы – метод «колена» и индекс Давида-Болдина. Из этого можно заключить, что критерий, основанный на счете информационного содержания консенсусной последовательности, действительно позволяет получать функционально родственные группы белков. Следовательно и сами

мотивы, лежащие в основе расчета информационного содержания, представляют собой структурно или функционально значимые участки.

#### **4. Выводы**

4.1. Разработан алгоритмический метод, позволяющий выявлять участки локальной консервативности для заданного набора первичных структур белков. Метод основан на представлении множественного выравнивания в виде консенсусной последовательности с последующей статистической оценкой её информационного содержания.

4.2. Для надсемейства цитохромов P450 показано, что выявленные участки локального сходства соответствуют структурно-функциональным мотивам. В консенсусе надсемейства мотивы общности определяют фолд-детерминирующую основу белка, обеспечивают фиксацию гема, молекулярного кислорода и формирование канала доступа лиганда. В консенсусе семейства стероловых деметилаз мотивы частного отвечают участкам специфичного узнавания субстрата.

4.3. Применение разработанного алгоритма позволяет определять уровень отсека при проведении кластеризации последовательностей надсемейства. Повышение уровня соответствия между составом кластеров и семействами, сформированными согласно общепринятой номенклатуре, свидетельствует о значимости выявляемых мотивов для задачи определения функциональной специфичности цитохромов P450.

#### **5. Список опубликованных работ по теме диссертации**

Lisitsa A.V., Gusev S.A., Miroshnichenko Y.V., Archakov A.I. “Bioinformatic insight into the unity and diversity of cytochromes P450” (2003) Proceedings of the 13-th Conference on Cytochromes P450, pp.7-13.

Лисица А.В., Мирошниченко Ю.В., Иванов Н.А., Арчаков А.И. Общее и частное в структурной организации белков надсемейства P450. Аллергия, астма и клиническая иммунология, 2003, т.7, №8, с.14-19.

Пономаренко Е.А., Лисица А.В., Карузина И.И., Мирошниченко Ю.В. Автоматизированное аннотирование функциональных свойств белков надсемейства цитохромов P450. Аллергия, астма и клиническая иммунология, 2003, т.7, №8, с.95-99.

Лисица А.В., Гусев С.А., Мирошниченко Ю.В., Кузнецова Г.П., Лазарев В.Н., Скворцов В.С., Карузина И.И., Говорун В.М., Арчаков А.И. Структурно-функциональные

мотивы стероловых 14-альфа-деметираз (CYP51). Биомедицинская химия, 2004, 50(6): 554-565.

Лисица А.В., Мирошниченко Ю.В., Пономаренко Е.А. (2003) База знаний по цитохромам P450. Симпозиум «Биоинформатика и компьютерное моделирование лекарств» в рамках X Российского национального конгресса «Человек и лекарство».