На правах рукописи

ЛИСИЦА Андрей Валерьевич

БАЗА ЗНАНИЙ ПО ЦИТОХРОМАМ P450: РАЗРАБОТКА И ПРИМЕНЕНИЕ

03.00.28 - биоинформатика

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
доктора биологических наук

Москва - 2007

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

Научный консультант:	доктор биологических наук, профессор, академик РАМН Арчаков Александр Иванович
Официальные оппоненты:	доктор биологических наук Гельфанд Михаил Сергеевич
	доктор физико-математических наук, профессор Шайтан Константин Вольдемарович
	доктор физико-математических наук Туманян Владимир Гаевич
Ведущая организация	Федеральное государственное учреждение Научно-исследовательский институт физико-химической медицины Росздрава

Защита состоится «26» апреля 2007 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: 119121, Москва, ул. Погодинская, 10.

С диссертацией можно ознакомиться в библиотеке ГУ НИИ биомедицинской химии имени В.Н. Ореховича РАМН по адресу: 119121, Москва, ул. Погодинская, 10.

Автореферат разослан «____» _____ 2007 года.

Ученый секретарь Диссертационного совета
кандидат биологических наук В.С. Былинкина

1. ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

1.1 Актуальность проблемы

Концепция замещения ранее созданных баз данных – статических информационных ресурсов – динамически развивающимися *базами знаний* обусловлена необходимостью своевременной разноплановой обработки накапливающегося экспериментального материала. Отличительным признаком базы знаний является гибкая структура данных, способная эффективно адаптироваться к быстро меняющимся условиям поставленной задачи. Особенно актуальной такая способность становится в контексте приложений к задачам молекулярной биологии, характеризующимся противоречием между неполнотой отобранных экспериментальных данных и системной целостностью объекта исследования – живого организма.

Анализ молекулярных процессов является наиболее детальным способом изучения живых систем, который доступен современным исследователям. Технологически проведение широкомасштабных исследований молекулярных систем стало возможным благодаря созданию высокоэффективных технологий. Одновременно, достижения геномных и постгеномных технологий вывели на первый план задачи, связанные с хранением и обработкой получаемой информации. Качественный скачок в развитии молекулярной биологии, обусловленный внедрением новых технологий и накоплением множества разрозненных, но взаимодополняющих экспериментальных данных, ознаменовался появлением новой научной дисциплины – системной биологии. В рамках системной биологии концепция базы знаний получила признание как формализованный подход для выявления скрытых закономерностей в накопленных данных.

Системные подходы в молекулярной биологии находят свое отражение в развитии баз знаний на основе исторически сложившихся глобальных информационных ресурсов: на смену базе данных по первичной структуре белковых последовательностей SwissProt [http://au.expasy.org/sprot/] приходит ресурс нового поколения – база знаний UniProt [http://www.expasy.uniprot.org/]; база данных геномных последовательностей GenBank [http://www.ncbi.nlm.nih.gov/Genbank/index.html] в настоящее время тоже рассматривается в контексте целого арсенала вспомогательных алгоритмических средств работы с данными, т.е. представляет собой базу знаний, объединяемую техническим понятием «комплекс ресурсов NCBI» [http://www.ncbi.nlm.nih.gov]. Вышеперечисленные системы являются проблемно-ориентированными, что подразумевает хранение информации обо всем разнообразии генов и белков.

С существенным отставанием от глобальных систем хранения информации развиваются объектно-ориентированные базы знаний, к которым относится база знаний по цитохромам Р450. Задачей объектно-ориентированных ресурсов является сбор всех имеющихся данных об одном классе белков.

Объектная ориентированность базы знаний по цитохромам Р450 определяет ее уникальность и очерчивает область ее потенциального применения в научных исследованиях. Разработанная структура базы знаний позволяет в рамках одной

системы параллельно накапливать информацию о структурном и функциональном разнообразии цитохромов Р450. Структурный и функциональный потоки аккумулируют данные независимо друг от друга, формируя объективные предпосылки для развития гипотез о наличии структурно-функциональных взаимосвязей. Развитие гипотезы происходит в контексте определенной статистической или алгоритмической модели, которая строится исходя из одного типа данных, а проверяется на другом. Так, рассматриваемая в работе «островная гипотеза» [Nishikawa, 1993] строения белковых молекул, в рамках базы знаний формализуется в виде статистического метода выявления структурных мотивов двух типов путем сравнительного анализа последовательностей аминокислотных остатков [Lisitsa et al., 2003]. Найденные структурные мотивы затем используются для корректировки существующей классификации. Вносимые корректировки, т.е. отличия традиционной классификации от варианта, построенного на основе «островной гипотезы», составляют сущность нового знания о структурно-функциональных взаимосвязях в надсемействе цитохромов Р450.

Важным фактором, определяющим значимость представляемой работы, является функциональная роль ферментов надсемейства цитохромов Р450. Реализуемая цитохромами Р450 реакция монооксигеназного катализа является необходимым звеном в обеспечении жизнедеятельности организмов, начиная с простейших и заканчивая многоклеточными эукариотами. Многогранность каталитических особенностей, структурное разнообразие известных генетических форм, широкий арсенал методов экспериментальных исследований делают цитохромы Р450 актуальным объектом для апробации технологии создания объектно-ориентированной базы знаний.

Цель работы – создать информационно-вычислительный ресурс (базу знаний), позволяющий проводить систематизированный анализ общедоступных данных о структурных и функциональных особенностях белков надсемейства цитохромов Р450. База знаний должна предоставлять интегрированную платформу для проведения исследований надсемейства биоинформационными методами. В рамках достижения поставленной цели сформулированы следующие **задачи**:

1. Разработать структуру данных и способы автоматизации процедуры пополнения информационного массива; реализовать контекстно-зависимые схемы адаптации структуры данных; обеспечить автоматические средства сопряжения базы знаний с другими информационными системами.
2. Внести в базу знаний сведения о структуре и функции цитохромов Р450, в полном объеме отражающие современный уровень исследований в данной области.
3. Интегрировать в базу знаний базовые алгоритмы биоинформатики, предназначенные для сравнительного анализа последовательностей аминокислотных остатков; разработать интерактивные средства работы с этими алгоритмами.
4. С использованием базы знаний выполнить комплекс работ по анализу структурно-функциональных особенностей цитохромов Р450 и предложить объективные подходы к классификации белков надсемейства.

1.2 Научная новизна и практическая значимость

Впервые показана возможность создания базы знаний и её последующего применения для решения научно-исследовательских задач, связанных с анализом надсемейства цитохромов Р450.

Разработан способ формализации эмпирических знаний, накопленных в результате экспериментов по изучению структуры и функции цитохромов Р450, и предоставлен доступ к широкому спектру биоинформационных алгоритмов, таких, как алгоритмы выравнивания последовательностей, кластерного анализа, методы построения консенсусных последовательностей и выявления структурно-функциональных мотивов.

Проведены исследования подходов к созданию объективной классификации надсемейства цитохромов Р450 с использованием комбинации хорошо изученных методов биоинформатики. Для этого разработан инструментарий оригинальных методов, включающий:

- метод иерархического выравнивания, позволяющий осуществлять выравнивание консенсусных последовательностей;
- метод структурно-функционального картирования, предназначенный для обозначения на аминокислотной последовательности элементов вторичной структуры белка, субстрат-узнающих участков, точечных мутаций, структурно-функциональных мотивов и др.
- метод инвентаризации, позволяющий распределить белки надсемейства по кластерам и реконструировать последовательность-предшественник для каждого кластера;
- метод индексирования белков, позволяющий сгенерировать целостную модель эволюционирования белков анализируемой группы от гипотетического белка-прародителя;
- метод выявления структурно-функциональных мотивов, используемый для обозначения в составе консенсусной последовательности статистически-значимых локальных участков консервативности.

Разработан комплекс подходов и методических приемов, который может быть использован для прогнозирования функциональной специфичности новых форм цитохромов Р450. Информация, содержащаяся в базе знаний, может быть применена при моделировании пространственных структур цитохромов Р450 и при создании структурно-функциональных моделей. Выявленные структурно-функциональные мотивы могут быть использованы при планировании генно-инженерных экспериментов по созданию искусственных форм цитохромов Р450 с новыми функциями. Практическая роль разработанной базы знаний также важна в качестве интерактивного справочного и обучающего пособия.

Впервые для отдельного надсемейства белков представлены результаты масштабного технологического программирования, ориентированного на организацию взаимосвязанных сценариев работы пользователя с данными. Разработанные сценарии включают в себя до 8 этапов, на каждом из которых пользователь получает дополнительную информацию об объекте исследования. Эта информация потенциально является основой для построения научных *гипотез* и

дальнейшего рационального планирования научно-исследовательской работы. Апробированные технологические приемы могут быть перенесены на другие группы белков, кроме цитохромов Р450, и представляют практическую значимость с точки зрения развития современных подходов к обработке молекулярно-биологических данных.

1.3 Основные положения, выносимые на защиту

1. База знаний обеспечивает интегрированную платформу для хранения и анализа информации о структурно-функциональных особенностях белков надсемейства цитохромов Р450.

2. База знаний поддерживает основные методы обработки информационного массива и позволяет применять эти методы для выполнения научно-исследовательской работы.

3. Применение базы знаний позволяет систематизировать методы кластерного анализа первичных структур цитохромов Р450, установить наличие мотивов общего и частного характера и применить найденные мотивы для реализации нового способа классификации белков надсемейства цитохромов Р450.

1.4 Апробация работы

Основные положения диссертационной работы были доложены и обсуждены на симпозиумах и конференциях:

- 7-th International Conference “Biochemistry & Biophysics of cytochrome P450: Structure & Function, Biotehnology & Ecological Aspekts (INCO-TNC Joint Stock Company, 1992);
- 9-th International Conference “Cytochrome P450: Biochemistry, Biophysics and Molecular Biology” (Zarich, 1995);
- 3-th IUBMB Company Molecular Recognition (Singapore, 1995);
- 12-th International symposium on microsomes and drug oxidations (Montpellier France Le Corum, 1998);
- International workshop “From Sequence to function: Experimental and Bioinformatic Studies of Cytochrome P450 Superfamily” (Moscow, 2000);
- 13 International Symposium on Microsomes and Drug Oxidation.-Stresa-Italy.-Satellite Symposium of the VII World Conference on Clinical Pharmacology and Therapeutics (Florence, 2000);
- 4-th International Conference on Molecular Structural Biology (Vienna, 2001);
- 12-th International Conference on Cytochrome P450. Biochemistry, Biophysics and Molecular Biology (France, 2001);
- International Meeting on Proteome Analysis (Munchen, 2001);

- International Conference Genomics and Bioinformatics for Medicine (St.Peterburg-Moscow, 2002);
- 14th International Symposium in Microsomes and Drug Oxidation (Sapporo Japan, 2002);
- 13-th International Conference on Cytochromes P450 (Prague, 2003);
- 5th International Conference on Molecular Structural Biology (Vienna, 2003);
- Сессия ИВТН (Москва, 2003);
- X Российский национальный конгресс «Человек и лекарство» (Москва, 2003);
- 2nd International conference "Genomics, Proteomics and Bioinformatics for Medicine" (Moscow-Ples-Moscow, 2004);
- 7th International symposium on Cytochrome P450. Biodiversity and biotechnology (Japan, 2004);
- XII Всероссийская научно-методическая конференция «Телематика'2005» (Санкт-Петербург, 2005);
- 14th International conference on Cytochromes P450: biophysics and bioinformatics (Dallas, USA, 2005);
- HUPO 4th annual world congress (Munich, Germany, 2005);
- Сессии ИВТН-2006 (Москва, 2006);
- 5rd International conference on bioinformatics of genome regulation and structure (Novosibirsk, 2006);
- 3rd International conference "Genomics, proteomics, bioinformatics and nanotechnologies for medicine" (Novosibirsk, 2006);
- HUPO 5rd annual world congress (Long Beach, California, 2006);

Статистика посещения Веб-сайта, на котором размещена база знаний (http://cpd.ibmh.msk.su/), фиксирует более 200 обращений в год, из них 80% - от иностранных коллег.

Результаты диссертации легли в основу работы «База знаний по цитохромам P450: медицинские и биологические аспекты», удостоенной Премии Правительства Российской Федерации в области науки и техники для молодых ученых (раздел «Медицина») за 2006 год. Получено 2 свидетельства о регистрации программных продуктов для ЭВМ (№2004620199, №2006611941).

1.5 Публикации

Материалы диссертационной работы отражены в 63 публикациях: в 25 статьях и 38 материалах российских и международных научных конференций.

1.6 Объем и структура диссертации

Диссертационная работа изложена на 256 страницах машинописного текста, включая 34 таблицы, 64 рисунка. Состоит из введения, обзора литературы, материалов и методов исследования, результатов и обсуждения, выводов и списка литературы, включающего 275 источников.

2. ОБЪЕКТ И МЕТОДЫ ИССЛЕДОВАНИЯ

2.1 Надсемейство цитохромов Р450

Цитохромы Р450 – надсемейство, насчитывающее более 3 тыс. белков. Ферменты данной группы выявлены во всех царствах живой природы. При этом, несмотря на эволюционную разобщенность видов организмов, цитохромы Р450 сохраняют общность черт первичной структуры и пространственной организации – в этом смысле цитохромы Р450 интересны как объект для изучения общих закономерностей молекулярной эволюции.

Выявление генов, кодирующих цитохромы Р450 в представителях различных царств, свидетельствует о функциональной важности этого фермента и его необходимости для биологических организмов. Предположить общность некоего гена-предшественника для всего надсемейства, который затем наследовался видами по мере развития биосферы, подвергался дупликациям, дивергенции, латеральной диффузии и т.д., в свете современных воззрений на общие тенденции развития живой природы достаточно сложно. С другой стороны, при рассмотрении надсемейства следует особо учитывать доминирование требований к ферментативной функции цитохромов Р450, заключающейся в повышении растворимости веществ в воде. Данное общее начало, частично ограничивая процесс дивергенции дуплицированных генов, закрепляя гены, приобретенные в результате трансфера, и также способствуя конвергенции, определило, по-видимому, современный структурно-функциональный «ландшафт» надсемейства, и, одновременно, сделало его перспективным объектом для исследования в рамках базы знаний.

В эволюционном плане наряду с многообразием форм следует отметить особенности распространения цитохромов Р450 среди филогенетических царств. Если в составе полностью прочитанных геномов бактерий и простейших присутствуют в большинстве случаев 1-5 формы цитохрома Р450 и имеется много примеров отсутствия фермента вообще, то в царстве растений отмечается исключительное многообразие. Так, в геноме двудольного растения *A.thaliana* –187 форм цитохромов Р450, у однодольных *O. sativa* – их более 100. Начиная с насекомых наличие цитохромов Р450 в геноме становится обязательным условием существования организма, при этом многообразие форм сокращается по сравнению с растениями – 80 форм у дрозофиллы, 50 форм у человека [Nelson, 1998].

С точки зрения каталитической функции цитохромы Р450 участвуют в реакции монооксигеназного катализа, играя в ней ведущую роль за счет способности избирательно связываться с субстратом и ориентировать его в активном центре. Стехиометрические характеристики монооксигеназной реакции, причины избирательного катализа по строго определенным положениям молекулы-лиганда являются предметом тщательного изучения и позволяют объяснить фундаментальные механизмы выполнения белками ферментативной функции.

Функционирование цитохромов Р450 определяется взаимодействием с белками-партнерам. Цитохром Р450 замыкает собой цепочку переноса электронов и использует полученные редокс-эквиваленты для окисления субстрата. В качестве партнеров могут фигурировать как несколько белков (например, НАДФН-цитохром

P450 редуктаза и цитохром b5, адренодоксин редуктаза и адренодоксин) так и один белок – редуктаза.

Со структурной точки зрения цитохромы P450 характеризуются общностью пространственного фолда, определяемого взаимным расположением высококонсервативных участков. Последние преимущественно располагаются на С-конце последовательности аминокислотных остатков. К ним причисляют: гем пептид, альфа-спирали I и K – элементы, которые можно без труда различить как при сопоставлении пространственных структур, так и при множественном выравнивании гомологичных последовательностей.

Номенклатура надсемейства цитохромов P450 поддерживается путем проведения экспертной оценки их структурных особенностей. Основным критерием является сходство первичных структур: в семейство объединяются последовательности с гомологией более 40%, в подсемейство – последовательности, гомологичные на 46% и более [Nelson et al., 1996]. Наряду с постулированными принципами структурного сходства белков, номенклатура цитохромов P450 несет черты объединения ферментов по критерию функциональной близости (не имеющему строгой формализации) и на основании сходства структур генов – количество экзонов, сдвиги рамки считывания (на настоящий момент четкой концепции сходства структур генов цитохромов P450 также не сформулировано).

В организмах животных цитохромы P450 представляют интерес как ферменты первой фазы трансформации ксенобиотиков, в частности, лекарств и техногенных соединений – прокарциногенов. В связи с этим значительный объем исследовательских работ проводится для определения значимости отдельных форм цитохромов P450 для метаболизма лекарств и при мониторинге уровня загрязнения окружающей среды. Уникальность функции монооксигеназного катализа привлекает к ферментам надсемейства внимание биотехнологов, а участие этих белков в метаболизме гербицидов и пестицидов позволяет использовать результаты исследований для нужд сельского хозяйства.

2.2 Используемые информационные ресурсы

База знаний совместима с двумя категориями общедоступных информационных ресурсов, применяемых в молекулярной биологии. К первой категории принадлежат проблемно-ориентированные глобальные банки данных, такие как GenBank, SwissProt и PDB. Во вторую категорию входят менее известные широкой научной общественности базы данных, поддерживаемые усилиями узкоспециализированных научных коллективов. Ко второй категории, в частности, следует отнести Веб-сайт, на котором размещается информация об официальной номенклатуре белков надсемейства цитохромов P450 [http://drnelson.utmem.edu/CytochromeP450.html].

Табл. 1. Глобальные информационные ресурсы, используемые при разработке базы знаний по цитохромам Р450.

Глобальные ресурсы	
SwissProt [http://au.expasy.org/sprot/]	Первичные структуры белков
GenBank [http://www.ncbi.nlm.nih.gov/Genbank/index.html]	Структура генов, последовательности кДНК
Protein Data Bank [http://www.rcsb.org/pdb/home/home.do]	Пространственные структуры белков
KEGG [http://www.genome.ad.jp/kegg/]	Метаболические пути
HapMap [http://www.hapmap.org]	Локализация генов на хромосоме и средства визуализации структуры генов
PubMed [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed]	Литературные ссылки
ChemIDplus [http://chem.sis.nlm.nih.gov/chemidplus/]	Структурные формулы химических соединений

Кроме рассмотренных выше основных категорий информационных ресурсов в базе знаний используются вспомогательные источники данных. Примером такого источника может служить база данных структур низкомолекулярных химических соединений ChemIDplus. В табл. 1 систематизированы наиболее важные информационные ресурсы, с которыми сопряжена база знаний по цитохромам Р450.

Узкоспециализированные информационные ресурсы, как правило, являются источником данных о функциональных особенностях отдельных форм цитохромов Р450. С другой стороны, глобальные банки данных играют ведущую роль для получения информации о структуре макромолекул – генов, кодирующих цитохромы Р450, и об их продуктах – белках. Особое место среди источников информации, пополняющих базу данных, занимает Веб-страница номенклатурного комитета [http://drnelson.utmem.edu/CytochromeP450.html].

По характеру работы, проводимой с внешним информационным ресурсом в рамках базы знаний, следует выделять автоматический и контролируемый режимы работы. При автоматическом режиме в рамках заданного набора правил база знаний сортирует поступающую информацию и вносит ее в соответствующие поля данных. Контролируемый режим подразумевает интеграцию данных в базу знаний с последующей экспертной доработкой (подтверждением). Предусмотрено итеративное чередование автоматизированного и контролируемого режимов с целью обучения системы способам распознавания новой информации.

Совокупность информационных ресурсов, приведенная в таблице 1, не является статической. Структура и алгоритмическое обеспечение базы знаний по

цитохромам P450 обеспечивают механизмы привлечения новых категорий информации, используя для этого динамическую модель описания составляющих элементов.

2.3 Алгоритмы работы с первичной структурой белка

В базу знаний интегрирован базовый набор алгоритмов для обработки больших массивов данных по последовательностям аминокислотных остатков. В состав базового набора входят программы локального [Altshul et al., 1990], парного и множественного выравнивания последовательностей [Gotoh, 2000; Gotoh, 1999], алгоритмы построения иерархической кластеризации на основе матрицы парного сходства набора последовательностей [Sneath & Sokal, 1973], методы сегментирования кластеров в составе группы белков [Davies & Bouldin, 1979; Halkidi et al., 2001] и способы построения консенсусных последовательностей по результатам множественного выравнивания [Taylor, 1990].

При работе с парным и множественным выравниванием используется метод рандомизированных запусков для уточнения таких параметров, как штраф за открытие и продление вставки. Оптимальные значения штрафа при парном выравнивании соответствуют максимальному различию в значениях идентичности (или гомологии) между родственными последовательностями (т.е. между последовательностями цитохромов P450) и случайно сгенерированными символьными строками сравнимой длины. При оптимизации результатов множественного выравнивания рандомизации подвергаются не только значения штрафов за вставку, но и порядок вводимых для выравнивания последовательностей. Эвристический характер применяемого для множественного выравнивания алгоритма PRRP [Gotoh, 1999] обуславливает необходимость рандомизации для достижения оптимального результата – консенсуса с наибольшим количеством консервативных остатков.

Для выявления в составе консенсусной последовательности структурно-функциональных мотивов использовался статистический критерий Шермана [Sherman, 1957; Sneath, 1998]. Критерий позволяет выявить в составе консенсуса множественного выравнивания статистически неслучайные компактные кластеры консервативных остатков. На основе статистического критерия рассчитывалось информационное содержание консенсусной последовательности, которое затем применялось для определения границ кластеров в составе надсемейства.

В работе методы анализа первичной структуры цитохромов P450 используются в комбинации как друг с другом, так и с другими методами, например с алгоритмом кластерного анализа и алгоритмом выявления структурно-функциональных мотивов. Так, метод инвентаризации основан на комбинации парного выравнивания, кластерного анализа и множественного выравнивания. Метод индексирования использует результаты инвентаризации для расчета парным выравниванием расстояний между белком и консенсусом группы белков. Метод выявления структурно-функциональных мотивов обрабатывает консенсусные

последовательности, генерируемые в результате множественного выравнивания, и позволяет скорректировать результаты кластерного анализа.

В работе применяются следующие методы определения уровня отсечения для процедуры кластерного анализа:

- L-метод, основанный на анализе динамики агломерации [Lewi et al., 1992];
- индекс Джаккарда [Halkidi et al., 2001], позволяющий сравнить два варианта кластеризации между собой по составу кластеров;
- индекс Дэвиса-Болдина [Davies & Bouldin, 1979], основанный на анализе распределения длин ветвей дендрограммы;
- критерий максимума информационного содержания консенсусной последовательности, основанный на анализе структурно-функциональных мотивов.

Общая схема использования алгоритмов анализа первичной структуры белка приведена на схеме 1. Пользователь, принимая за основу некоторую определенную группу белков (в частных случаях, в качестве такого рода группы может фигурировать либо все надсемейство, либо отдельный его представитель), экспортирует ее в модуль аналитической обработки. В рамках этого модуля пользователь осуществляет поиск гомологичных последовательностей в глобальных банках данных, применяя программу BLAST [Altschul et al., 1990]. В случае если гомологичные белки обнаруживаются и ранее не были аннотированы и помещены в базу знаний, проводится индексация новых поступлений. Метод кластерного анализа позволяет оценить правомочность отнесения нового белка к анализируемой группе; углубленный анализ проводится на уровне структурно-функциональных взаимосвязей.

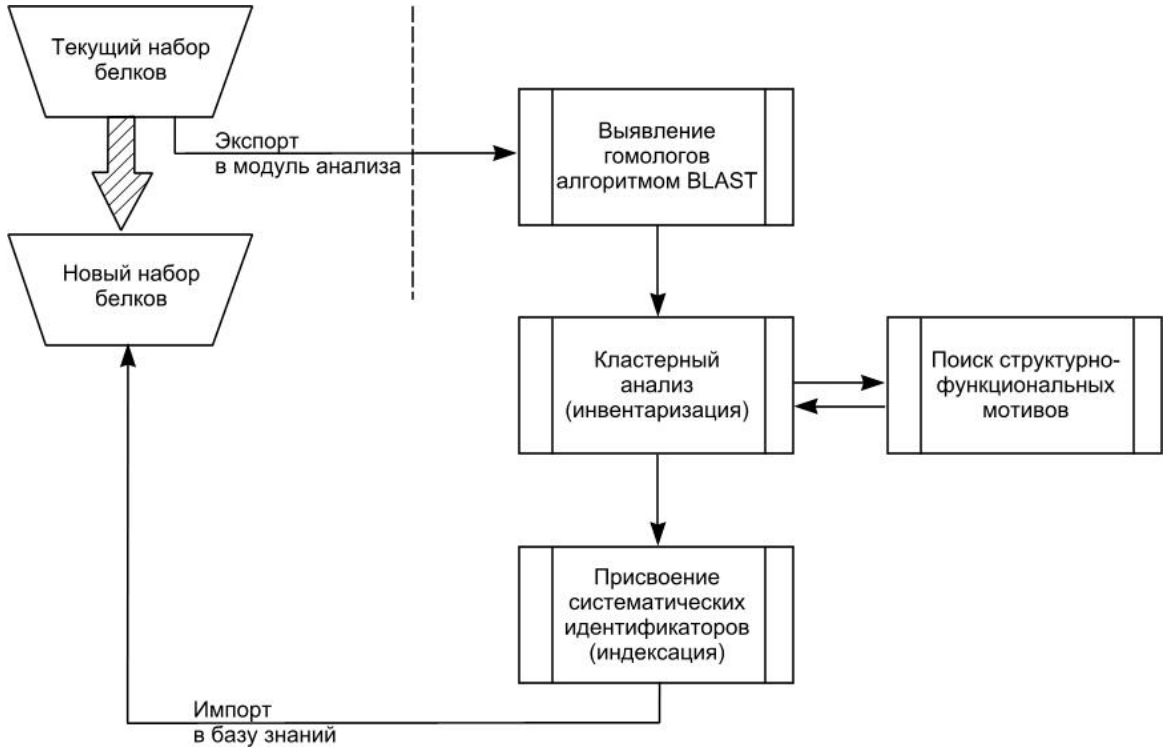

Схема 1. Обобщенная схема, иллюстрирующая механизмы обработки первичных структур белков в рамках базы знаний по цитохромам Р450.

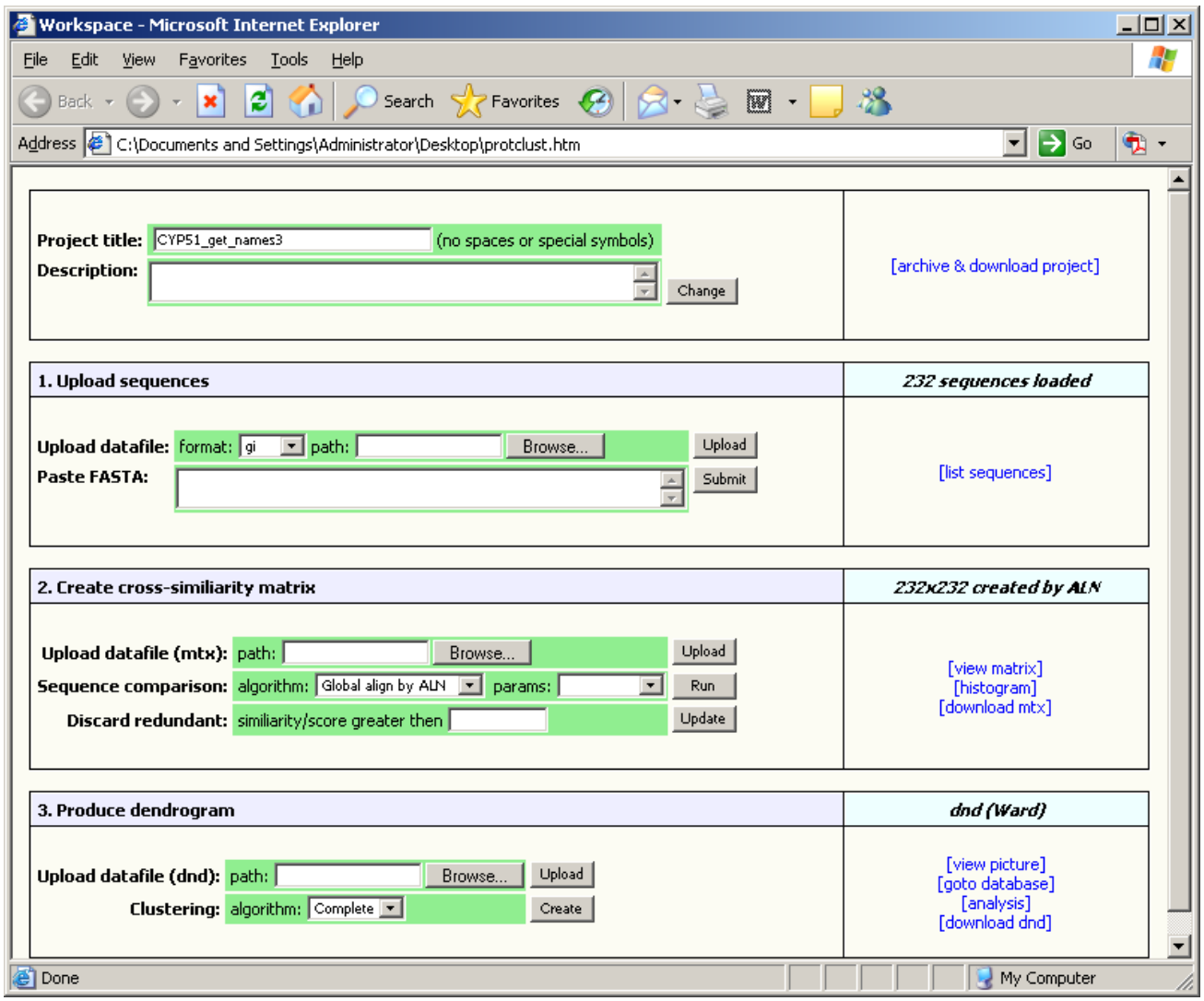

Рис. 1. Рабочий экран модуля анализа первичных структур белка, предназначенный для пополнения базы знаний новой информацией.

В результате применения вышеуказанных процедур пользователю предлагается присвоить новому белку систематический идентификатор. При этом наряду с формальным индикатором, вычисленным базой знаний, в ряде случаев предлагается идентификатор, рекомендованный номенклатурным комитетом. Анализируя и подтверждая решения системы на каждом этапе, куратор базы знаний подготавливает информацию к импорту, после чего новая аннотированная последовательность становится доступной широкому кругу пользователей.

На рис. 1 представлен рабочий экран модуля, предназначенного для работы с набором первичных структур белков. Модуль включает в себя четыре основных этапа работы: (1) загрузка исходного набора последовательностей в базу знаний (осуществляется автоматически); (2) построение матрицы парных сравнений; (3) проведение кластерного анализа; (4) анализ данных – инвентаризация, выявление структурно-функциональных мотивов и индексация.

2.4 Методы автоматического (текстомного) анализа документов

Автоматический анализ документов является источником информации о функциональных свойствах белков надсемейства цитохромов Р450. Аннотирование функциональных свойств осуществляется путем анализа текстов документов, написанных на естественном языке (английском) и размещенных в системе MedLine. В базу знаний по цитохромам Р450 включены два метода: метод оценки релевантности публикации к тематике информационной системы и метод смыслового анализа текста.

Для оценки релевантности документа используется вычислительный алгоритм, предложенный в работе [Mosteller & Wallace, 1984]. Алгоритм анализирует обучающую выборку текстов, сформированную экспертом, и рассчитывает частоту встречаемости каждого термина. Частота встречаемости в обучающей выборке сравнивается с фоновой частотой, оцениваемой по случайно сформированной выборке текстов. Понятие релевантности документа вводится как вероятность встретить термин в документе с заданной тематической направленностью с учетом фоновой частоты встречаемости данного термина в научных статьях. Термины, используемые для определения релевантности документа, являются дискриминаторными.

Алгоритм текстомного анализа, заложенный в базе знаний, кроме определения релевантности документа, осуществляет семантический анализ, как описано в [Muller, 2004]. Для этого определяются маркерные термины или тэги, характеризующиеся высокой частотой встречаемости в обучающей выборке. Специфическая лексика вводится в виде контролируемых словарей терминов. Маркерные и контролируемые термины используются для конструирования семантических шаблонов, позволяющих конвертировать текст в содержимое полей базы знаний.

2.5 Способы оценки взаимоотношений структура-активность

В базе знаний по цитохромам Р450 реализовано два метода оценки взаимоотношений структура-активность. Первый метод основан на использовании системы прогнозирования спектра активности биологически активного вещества PASS [Poroikov et al., 2003], второй – на применении методов распознавания образов к набору молекулярных дескрипторов химического соединения [Korolev et al., 2003].

Методика прогнозирования спектра активности химического соединения базируется на сопоставлении структурных формул [Borodina et al., 2003]. Модифицированный коэффициент Танимото используется в качестве меры сходства структур, исходя из предположения, что сходство химических структур определяет сходство биологической функции [Васильев и Спасов, 2006]. В приложении к базе знаний по цитохромам Р450 в качестве биологической функции рассматривается взаимодействие определенной формы цитохрома Р450 с заданным химическим соединением.

В качестве входной информации модуль базы знаний получает структурную формулу химического соединения, подготовленную согласно определенному

формату. В качестве выходной информации выдается список форм цитохромов P450, взаимодействующих с данным соединением. Для каждой формы фермента вычисляется вероятностная оценка гипотезы о взаимодействии (рис. 2).

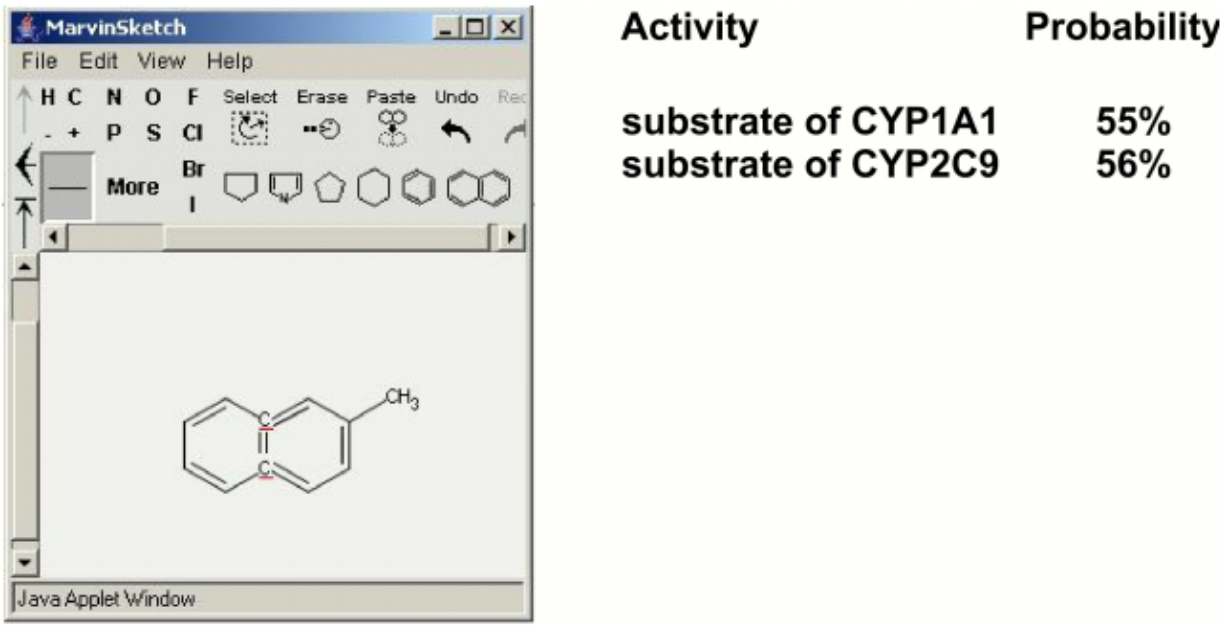

Рис. 2. Прогнозирование профиля взаимодействия химического соединения с ферментами надсемейства цитохромов P450.

В качестве альтернативного метода прогнозирования используется стандартный подход, основанный на молекулярных дескрипторах. Каждое химическое соединение конвертируется в набор числовых физико-химических характеристик: молекулярный вес, количество заместителей, коэффициент распределения октанол-вода и т.д. База знаний позволяет сформировать статистические списки химических соединений, взаимодействующих с определенными формами цитохромов P450. После векторизации сформированные списки выдаются в виде таблицы. Формат таблицы совместим со стандартными наборами методов распознавания образов, входящими в состав программного пакета MatLab и языка математического программирования R.

В качестве интерактивного средства обработки данных о физико-химических дескрипторах пользователю предлагается использовать пакет GEPAS [Vaquerizas et al., 2005] доступный в сети Интернет [http://www.gepas.org]. После загрузки данных в систему GEPAS возможно применение к ним широкого спектра методов распознавания образов, включая кластерный анализ, метод опорных векторов, метод упругих карт и т.д.

Использование кластерного анализа молекулярных дескрипторов позволяет оценить гетерогенность химических структур субстратов (лигандов) различных форм цитохромов P450 и изучить явления перекрестной субстратной специфичности.

3. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Разработанная информационная система является основным результатом выполненных работ. Возможность пользования созданной информационной системой в качестве базы знаний доказывается в диссертационной работе путем исследования фундаментальных проблем, связанных с систематикой, эволюцией и функциональной диверсификацией ферментов надсемейства цитохромов Р450. Основное внимание уделяется применению базы знаний для решения поисковых задач; рассматривается общая архитектура информационной системы, и описываются оригинальные методические приемы, позволяющие осуществлять аналитическую обработку данных.

3.1 Описание базы знаний

3.1.1 Общая характеристика

Разработанная база знаний реализует функции сбора и анализа данных. К средствам сбора данных относится комплекс интерактивных подпрограмм, использующих методы обработки первичных структур белков и текстов публикаций на естественных языках. Методы инвентаризации и индексации, лежащие в основе общего алгоритма обработки первичных структур, используются для выявления и систематизации новых форм цитохромов Р450. Средства анализа текстов резюме научных публикаций применяются для пополнения сведений о функциональных особенностях различных форм ферментов.

База знаний реализована в качестве интерактивной информационной системы, совместимой с сетью Интернет. Совместимость обеспечивает, с одной стороны, взаимодействие с основными общедоступными источниками исходных данных, а с другой – служит удобным средством для использования созданного ресурса.

База знаний осуществляет взаимодействие с пользователями в рамках набора сопрягаемых сценариев. Сценарии обеспечивают целенаправленность работы пользователя от момента первичного ознакомления с информационным массивом до генерации обобщающих статистических гипотез. В основу реализованного механизма генерации гипотез положены представления о структурно-функциональном единстве надсемейства. В рамках этой концепции были созданы оригинальные средства для выявления и оценки структурно-функциональных мотивов и для формирования оптимальной картины распределения белков по кластерам; были разработаны средства интеграции структурно-функциональных особенностей с тенденциями молекулярной эволюции и методами статистического анализа функционального разнообразия ферментов надсемейства цитохромов Р450.

Основным инструментом при работе пользователя с базой знаний является механизм формирования выборок. Выборка объединяет в себя ферменты, связанные определенной родственностью, например, сходством структур лигандов и спецификой катализа, особенностям внутриклеточной локализации, видовой специфичностью и др. Критерии создания выборок могут быть различны; вне зависимости от характера критерия, база знаний предоставляет аналитический

аппарат для выявления структурно-функциональных основ предполагаемого пользователем сходства.

Наряду с возможностью применения базы знаний в целях научного поиска, следует отметить ее информационно-справочное назначение. Динамическое пополнение информационного массива обеспечивает системный анализ сведений о белках надсемейства цитохромов Р450. Функциональные возможности базы знаний, такие, как возможность предоставления пользователю статистических справок различного формата, позволяют четко определять тенденции развития современных представлений о надсемействе цитохромов Р450, использовать систему в качестве интерактивного справочника при подготовке научных публикаций и обеспечивать методические средства для повышения квалификации кадров.

3.1.2 Архитектура базы знаний

Архитектура разработанной базы знаний по цитохромам Р450 представлена на рис.3.

Рис. 3. Архитектура базы знаний по цитохромам Р450.

Загрузка данных в систему осуществляется за счет организованного взаимодействия между автоматическими средствами обработки информации и экспертными оценками. Информация о первичных структурах обрабатывается согласно процедуре инвентаризации (см. далее), в ходе которой проводится кластерный анализ, построение консенсусных последовательностей и выявление структурно-функциональных мотивов. На основании выявленных мотивов присваиваются инвентаризационные идентификаторы каждой форме цитохромов Р450. Инвентаризационные идентификаторы сопоставляются номенклатурным названиям ферментных форм, что позволяет сохранять совместимость вводимых данных с общепринятой системой классификации надсемейства.

Источником информации о первичных структурах цитохромов Р450 являются глобальные банки данных, предоставляющие информацию о геномах. Программа локального выравнивания применяется для поиска гомологичных структур; за счет этого происходит пополнение состава семейств и подсемейств. При выявлении новой формы цитохрома Р450 база знаний обращается к сайту номенклатурного комитета с целью присвоения систематического идентификатора. Одновременно, новая последовательность аминокислотных остатков включается в процедуры инвентаризации и индексации, описанные далее. Интерактивные средства для поиска новых форм цитохромов Р450 и интеграции их в состав надсемейства реализованы в отдельных программных компонентах базы знаний. Подробное описание функциональных возможностей этих компонент приводится в диссертационной работе.

Разработанные программные средства представляют также возможность обновления данных о функциональной активности ферментов надсемейства путем внесения информации о характерных субстратах, ингибиторах и индукторах. Для этого используются средства автоматизации смыслового анализа резюме научных публикаций.

Структура программной части базы знаний создавалась с использованием максимально открытой архитектуры, с учетом возможности дополнения и развития системы в будущем. Можно выделить три основных логических модуля базы знаний (два из них реализованы в виде отдельных, обособленных программ, последний представляет собой комплекс программ): а) модуль визуализации текущей выборки позволяет эффективно указать конкретный объект в составе выборки (например, форму цитохрома Р450, семейство или подсемейство), б) модуль представления общей информации по выбранному объекту (статистическая справка), в) средства представления специализированной информации (ДНК, аминокислотная последовательность, список субстратов и многое другое).

Визуализация текущей выборки служит для отображения номенклатуры цитохромов Р450 различными способами, а также для инициирования операций над текущей выборкой (поиск по запросу, подготовка отчетов, выравнивание, кластерный анализ). Различные способы сортировки данных в выборке позволяют пользователю быстро перемещаться по информационному массиву и легко обнаружить интересующий его объект. После выбора объекта, активируется модуль вывода общей информации.

Модуль вывода общей информации работает по-разному, в зависимости от типа выбранного объекта. Если объектом является группа цитохромов Р450 (семейство, подсемейство, белки одного вида и т.п.), то выводится статистическая справка по количеству генов и белков в данной группе. Если же объект является формой цитохрома Р450, то модуль информации сканирует все файлы базы знаний и выводит на экран доступные для данной формы дополнительные сведения по трем категориям: а) структура б) функция и в) сгенерированные данные. Важным свойством модуля вывода информации является возможность скопировать текущий объект (т.е. или форму цитохрома Р450 или их группу) в отдельную выборку.

Каждой из перечисленных в предыдущем абзаце категорий данных сопоставлены специфические средства отображения информации: при выборе

первичной структуры выводится последовательность аминокислотных остатков в однобуквенной кодировке с подсветкой функционально важных областей; ДНК выводится с разметкой кодирующих участков; при выборе функциональной активности выводится список субстратов, индукторов, ингибиторов и характеристических реакций.

Указанная структура построения базы знаний в полной мере отвечает современным возможностям Веб-технологий, основанным на концепции максимально свободной, легко расширяемой архитектуры. Имея достаточную гибкость на уровне программирования, система высоко упорядочена с точки зрения пользователя: пользователь имеет ограниченное количество возможностей навигации, т.е. он фиксируется в рамках заранее продуманных сценариев.

3.1.3 Структурно-функциональные карты

Структурно-функциональное картирование – методический прием, применяющийся в рамках базы знаний по цитохромам Р450 для аннотирования первичных структур надсемейства. Аннотирование включает разметку на последовательности аминокислотных остатков участков, несущих структурно-функциональную нагрузку. В качестве таких участков рассматриваются элементы вторичной структуры (выявленные методом рентгено-структурного анализа либо по результатам моделирования), участки узнавания субстрата и консервативные мотивы (см. рис. 4).

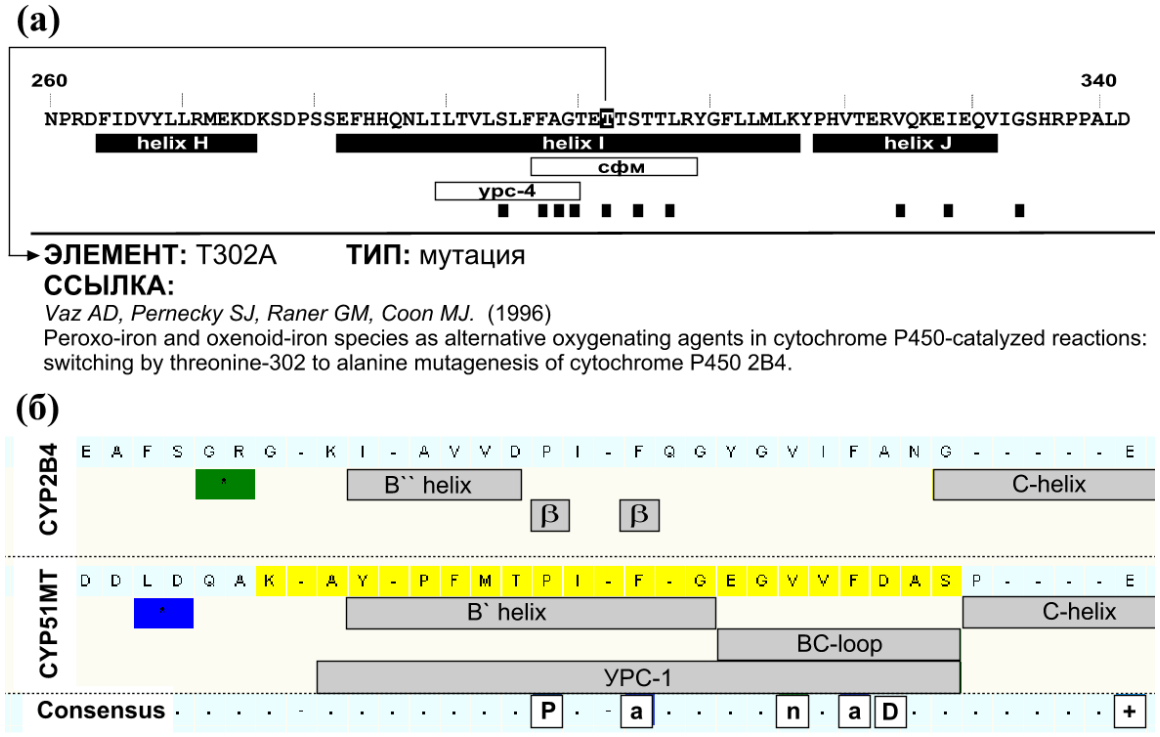

Рис. 4. (а) Создание структурно-функциональной карты. (б) Фрагмент сравнения структурно-функциональной карты цитохромов Р450 семейства CYP51 и подсемейства CYP2B. Сравнение проведено средствами программного компонента *Mal* базы знаний. «сфм» – структурно-фукнциональный мотив, «урс» - участок распознавания субстрата. Обозначения групп аминокислотных остатков: a= [FYW]; n=[LIMV]; +=[KRH].

В состав структурно-функциональных карт включается также экспериментальная информация о точечных мутациях. Указывается позиция аминокислотного остатка и соответствующая ссылка на опубликованные данные.

В ходе выполнения работы структурно-функциональные карты были созданы для 15 форм цитохромов P450, для которых известны пространственные структуры (рис. 4а). В настоящий момент информация, подготовленная в виде структурно-функциональных карт, доступна для подсемейств CYP1B, CYP2C, CYP3A (эти семейства содержат белки животных) и для 12 семейств цитохромов P450 бактериального происхождения.

Высокая гомология членов подсемейств CYP1B, CYP2C, CYP3A позволяет переносить структурно-функциональную информацию между сходными структурами. Перенос осуществляется на основании результатов парного или множественного выравнивания (рис. 4б).

На рисунке 4б представлен фрагмент выравнивания цитохрома CYP2B4 (сверху) с цитохромом семейства CYP51 (снизу). Фрагмент соответствует участку BC-петли. Видно, что участки, соответствующие спирали C, практически совпадают. В то же время, в области BC-петли наблюдаются существенные различия. Данный участок у белка CYP51 представлен достаточно протяженной спиралью B-prim, тогда как у цитохрома CYP2B4 в данном месте расположены бета-структуры (обозначены «β» на рис. 4б). Различие объяснимо с учетом того, что данная область (см. элемент “урс” на рис. 4) является участком распознавания субстрата, а субстратная специфичность сопоставляемых белков различна.

С использованием вышеописанного подхода выравнивания структурно-функциональных карт в ходе выполнения работы была дополнительно проведена аннотация 53-х последовательностей цитохромов P450, для которых нет экспериментальных данных о пространственной структуре.

3.1.4 Средства текстомного анализа

Для поиска и внесения в базу знаний информации о субстратах, ингибиторах и индукторах цитохромов P450 был разработан алгоритм анализа резюме научных публикаций.

В качестве обучающих данных алгоритму предоставляется выборка текстов резюме публикаций, сформированная путем экспертной оценки. При создании базы знаний по цитохромам P450 в основу обучающей выборки были положены литературные ссылки, размещенные в системе Human P450 Metabolism [Rendic & Di Carlo, 1997]. Размещенные на сайте литературные ссылки были выгружены в автоматическом режиме. Далее, с использованием средств поиска системы PubMed, каждая ссылка транслировалась в уникальный идентификатор статьи (PMID). Из ресурса PubMed резюме публикаций загружались по сформированному списку уникальных идентификаторов. В итоговой обучающей выборке содержалось 750 текстов резюме.

Сравнительная частота встречаемости терминов в обучающей и фоновой выборках использовалась для оценки релевантности документов. В качестве фоновой выборки использовалась коллекция из 36 тыс. резюме, в которых встречается упоминание цитохромов P450. Частота оценивалась согласно распределению

Пуассона как вероятность встретить термин в документе с заданной тематической направленностью, с учетом фоновой частоты встречаемости данного термина в научных статьях по тематике цитохромов Р450. Термины, удовлетворяющие статистическим условиям определения тематики резюме, отбирались в качестве *дискриминаторных* для дальнейшего использования в ходе автоматической обработки новых публикаций (терминология согласно [Mosteller F. & Wallace D., 1984]).

При анализе частот встречаемости наряду с дискриминаторными отбирались так называемые *маркерные* термины, характеризующиеся высокой частотой встречаемости в обучающей выборке. В категорию маркеров вошла общеупотребительная лексика, применяющаяся при описании экспериментов с ферментами надсемейства цитохромов Р450. В дальнейшем, маркерные термины были использованы для семантического анализа содержания резюме научных публикаций.

Частотный анализ был также применен для формирования контролируемых словарей терминов. В данном случае критерием отбора являлась низкая частота встречаемости термина. Подавляющее большинство отобранных таким образом терминов представляют собой номенклатурные названия химических соединений – субстратов, ингибиторов и индукторов цитохромов Р450. Контролируемые словари дополнялись также названиями видов экспериментальных животных, названиями тканей и клеточных компартментов, наименованиями широко используемых клеточных линий, номенклатурными и тривиальными названиями форм цитохромов Р450.

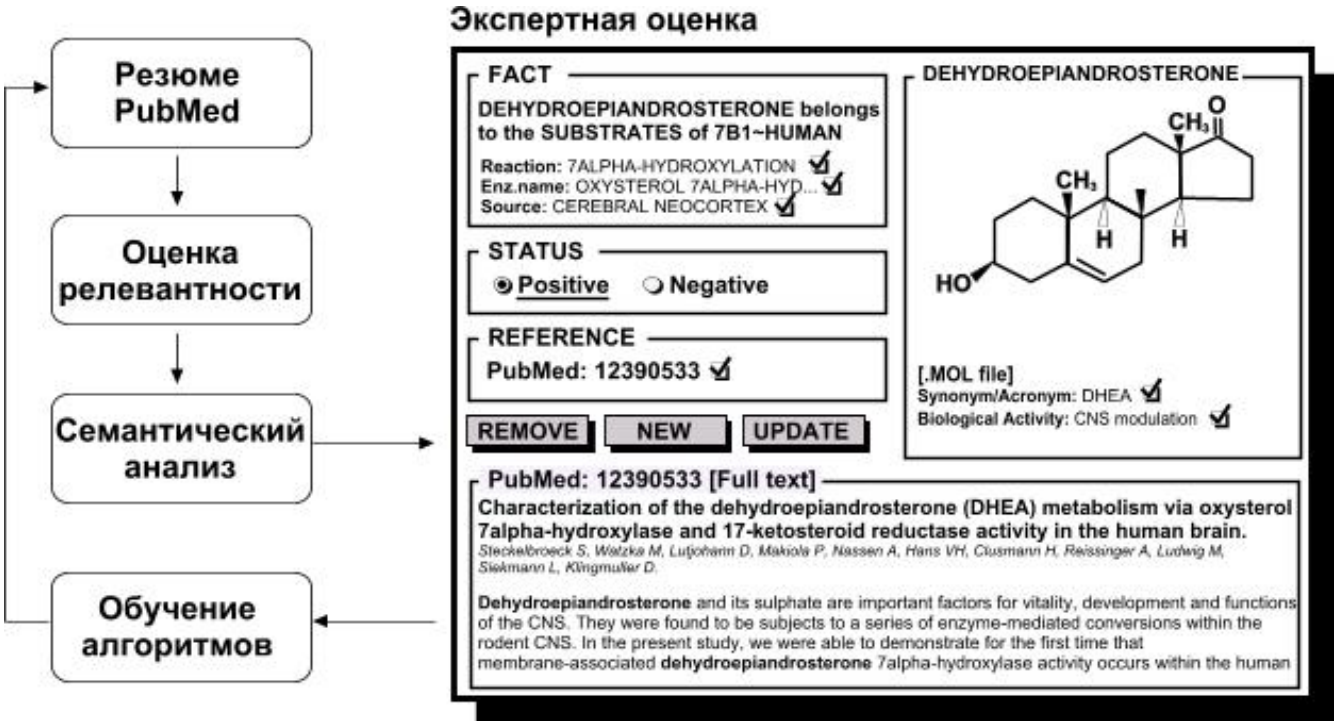

Схема 2. Применение методов семантического анализа научных публикаций в базе знаний по цитохромам Р450.

Процедура автоматического (текстомного) разбора текста резюме релевантной публикации была реализована следующим образом. Из каждого предложения, встречающегося в резюме обучающей выборки, были элиминированы все термины, кроме маркерных. Сформированные указанным образом шаблоны предложений сравнивались друг с другом с целью выявления наиболее часто встречающихся конструкций. Выявленные конструкции анализировались экспертами и оформлялись в виде типовых шаблонов. Указывались контексты, в рамках которых термин контролируемого словаря может быть классифицирован либо как субстрат, либо как индуктор, либо как ингибитор фермента надсемейства цитохромов Р450.

Схема 2 иллюстрирует общие принципы применения методов автоматизированного анализа публикаций в рамках созданной базы знаний. На первом этапе оценивается релевантность публикации. Далее, с использованием семантических шаблонов текст на естественном языке разбирается и транслируется в экспертную анкету. В экспертной анкете экстрагированная информация разнесена по формальным полям базы знаний. Эксперт проводит анализ анкеты, подтверждает, дополняет либо отклоняет решения автоматизированной системы. Экспертная оценка позволяет внести новую информацию в базу знаний и одновременно пополнить объем обучающих примеров. Циклический характер процедуры способствует повышению качества автоматизированного семантического анализа.

В работе показано, что использование методов текстомного анализа значительно повышает эффективность внесения данных в информационную систему. Уровень ошибок при определении релевантности не превышал 30%. При семантическом разборе уровень ошибок поднимался до 45-50%, однако, большинство ошибок было связано с внесением лишней информации, а не с упущением важных данных. Таким образом, большинство действий эксперта было связано с отклонением недостоверно внесенных данных, а не с вводом новой информации.

3.2 Объем собранных данных

Объем данных, введенный в базу знаний, постоянно обновляется за счет работы группы экспертов. В соответствии с общей схемой организации информационной системы вводятся данные о новых структурах цитохромов Р450, а также об их функциональной активности. По состоянию на начало 2006 г. общее количество форм цитохромов Р450 составляет более 2 тыс. (см. табл. 2)

Наибольшее количество генов, отнесенных к надсемейству цитохромов Р450, выявлено в геномах животных. Треть от общего количества цитохромов Р450 содержится в геномах растений, при этом доля растительных форм постоянно возрастает.

В табл. 3 сведена информация о функциональных свойствах цитохромов Р450.

Из табл. 2 и 3 видно, что экспериментальные данные о функциональных свойствах имеются только в отношении незначительного количества (13%) генов, кодирующих цитохромы Р450. При этом большинство фактов фигурируют как положительные утверждения, констатирующие наличие взаимосвязи между

ферментом и лигандом; фактов, базирующихся на отрицательных утверждениях, т.е. указывающих на отсутствие взаимосвязи, на порядок меньше.

При размещении информации о функциональных свойствах формы фермента в качестве основания указывается литературная ссылка и загружается соответствующее резюме публикации. Одновременно ведется работа по дополнению ссылок полнотекстовыми публикациями, источником которых является либо система PubMedCentral, либо частная переписка с авторами. К сожалению, как будет показано далее, разрыв между структурными и функциональными данными растет.

Табл. 2. Данные о первичных структурах цитохромов Р450, размещенных в базе знаний на январь 2006 г.

Название выборки	**Описание**	**Число генов\белков**
COMPLETE	Полная выборка, состоит из всех известных форм цитохромов Р450	2239
ANIMAL	Цитохромы Р450 в геномах животных и человека	1174
PLANT	Цитохромы Р450 в геномах растений	650
FUNGI	Цитохромы Р450 в геномах дрожжей	200
BACTERIA/ PROTISTA	Цитохромы Р450 в геномах бактерий и простейших	207
CYP51	Семейство стероловых деметилаз, которое предположительно являются родоначальником всех современных форм цитохрома Р450	42
PDB	Цитохромы Р450, для которых известны трехмерные структуры	17

Табл. 3 Информация о функциональных свойствах цитохромов Р450 в базе данных на январь 2006 г.

Описание	**Число записей**
Количество форм цитохромов Р450 с аннотированными функциональными свойствами	276
Количество химических соединений, взаимодействующих с ферментами надсемейства	1708
из них субстратов	1223
индукторов	115
ингибиторов	484
Количество резюме тематических публикаций	2044

3.3 Верификация виртуального эксперимента

Применение методов биоинформатики требует постановки контрольных экспериментов, направленных на оценку степени достоверности получаемого результата и его объективности.

В работе используются два способа независимой проверки объективности получаемого результата. Первый опирается на понятие согласованности результатов, получаемых одним и тем же методом, запускаемым с различными параметрами. В случае, если в рамках используемой алгоритмической схемы варьирование входных параметров после определенного количества итераций стабилизирует конечный результат (т.е. изменение результирующей картины при дальнейшем варьировании параметров пренебрежимо мало), данный алгоритм считается объективным. Примеры применения такого рода стабилизирующих схем приведены на рис. 5(а, б).

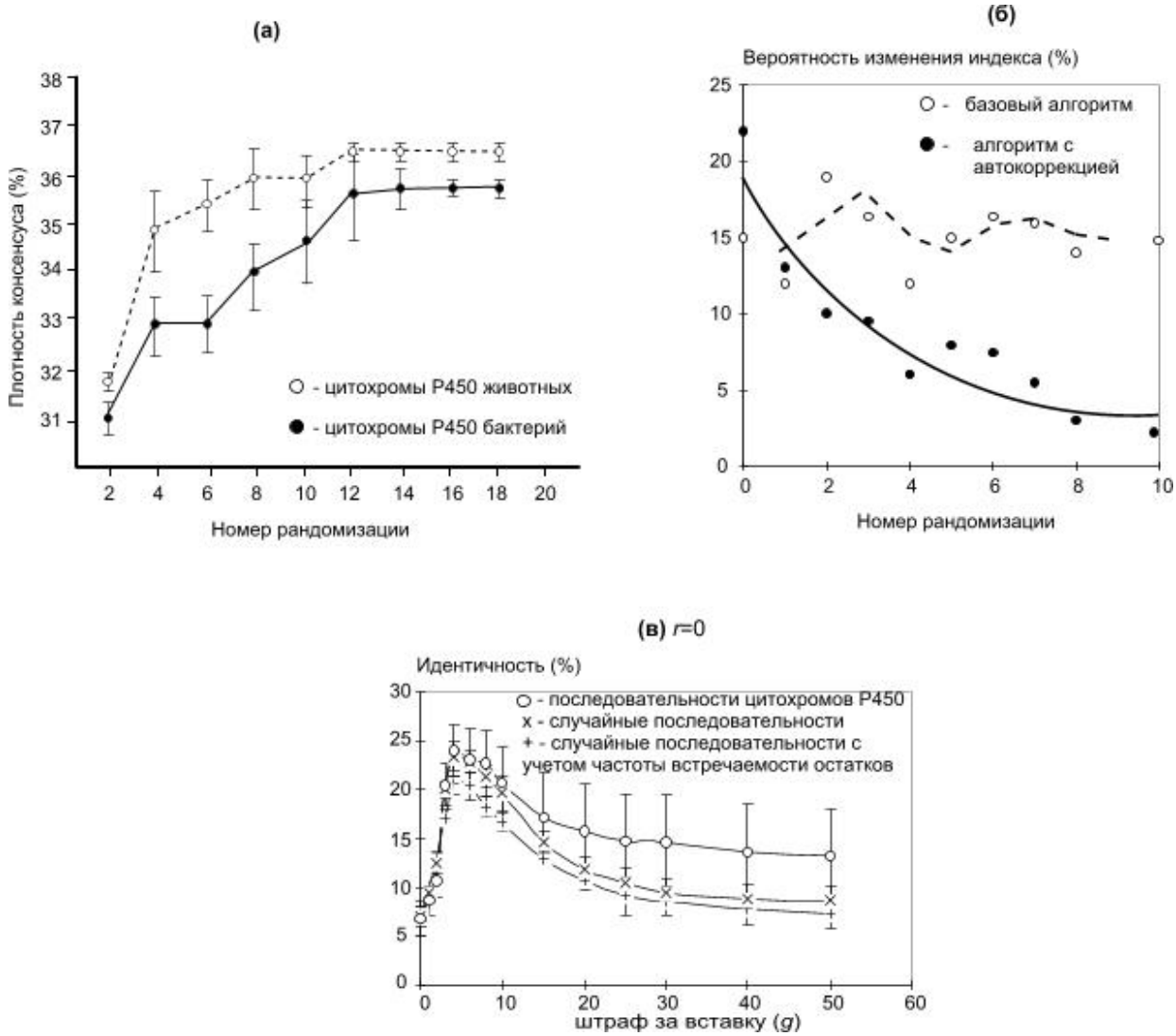

Рис. 5. Способы верификации результатов виртуального эксперимента: (а) стабилизация плотности консенсусной последовательности множественного выравнивания, построенного для различных выборок цитохромов P450; (б) стабилизация порядка следования белков в протеомном индексе надсемейства; (в) подбор оптимального штрафа за вставку.

На рис. 5а показано, как после 12-18 итераций плотность консенсуса – численный показатель эффективности множественного выравнивания – выходит на плато. При каждой итерации изменяется порядок следования записей во входном файле, содержащем последовательности и значения штрафов за открытие и продление вставки. Видно, что плотность консенсусной последовательности увеличивается с 31% до 36%, что существенно, учитывая, что общая длина выравнивания – 600-700 позиций.

Эффект, отображенный на рис. 5б определяется субоптимальным характером парного выравнивания: счет динамического программирования достигает максимального значения только при определенных значениях штрафа за вставку. Поскольку протеомный индекс надсемейства строится на основании счетов динамического программирования, то порядок следования объектов (консенсусов семейств в данном случае) изменяется с вероятностью около 15%, что отражено пунктиром на рисунке. После ввода в алгоритм построения индекса дополнений, позволяющих отобрать максимально возможный счет динамического программирования, достигается стабилизация индекса, и вероятность изменения порядка следования снижается до уровня 1,5%.

Второй принцип верификации основан на анализе поведения аналитических алгоритмов на рандомизированых наборах исходных данных. Критерием объективности является способность вычислительного метода различать стохастическую информацию – шум от информационно нагруженного сигнала. Так, например, случайно сгенерированные последовательности символов, кодирующих аминокислотные остатки, позволяют найти оптимальные значения параметров парного выравнивания (рис. 5в), оценить границы применимости метода кластерного анализа и подтвердить эффективность предлагаемого статистического критерия.

Для контроля достоверности результатов парного выравнивания, значение идентичности между родственными последовательностями цитохромов Р450 сравнивается со значением, полученным для случайных наборов букв соответствующей длины. Из рис. 5в видно, что хотя наибольшее значение идентичности достигается на уровне штрафа за открытие вставки 5-6, предпочтительней является область больших значений штрафа 20-40. На этом интервале различие между случайными и родственными последовательностями максимально.

При сравнении консенсусных строк, полученных множественным выравниванием цитохромов Р450 одного семейства и цитохромов Р450 разных семейств, установлено следующее. При одинаковой плотности консенсуса, т.е. когда в консенсусе членов одного семейства и консенсусе членов разных семейств присутствует одинаковое количество консервативных остатков на единицу длины, значения информационного содержания различны. Это означает, что величина информационного содержания чувствительна к характеру распределения консервативных остатков, а не просто к их количеству. Это свойство информационного содержания позволяет выявить семейства цитохромов Р450 за счет выраженного смещения максимума значений вправо по оси абсцисс.

3.4 Применение базы знаний в научной работе

3.4.1. Инвентаризация и индексация надсемейства

В диссертационной работе предлагается объективный подход к созданию целостной систематики белкового надсемейства. Подход основан на выполнении комплекса вычислительных процедур обработки данных по первичным структурам цитохромов Р450. На этапе *инвентаризации* осуществляется распределение белков по группам [Lisitsa & Archakov, 2003]. Основанием для отнесения белков к одной группе является сходство первичных структур. Для формирования групп используется метод кластерного анализа.

На последующем этапе анализа надсемейства – этапе *индексации* – осуществляется присвоение белкам уникальных идентификаторов (кодов). Код вычисляется путем численной оценки эволюционного расстояния, «пройденного» каждым белком от гипотетического белка-прародителя. Реконструкция белка-прародителя осуществляется путем проведения множественного выравнивания в группах гомологичных белков с последующим построением консенсусной последовательности.

В ходе инвентаризации были показаны границы статистически значимого уровня сходства последовательностей аминокислотных остатков. Анализ распределения счетов локального выравнивания позволил выявить пороговые значения, определяющие принадлежность аминокислотной последовательности к надсемейству. В ходе контрольного эксперимента использовалось распределение счетов локального выравнивания, полученное для случайных последовательностей. Из рис. 6 видно, что наиболее выраженный максимум гистограммы распределения счетов локального выравнивания приходится на значение 100 битовых единиц. Следовательно, при проведении аннотации новых геномов, последовательности, показывающие большие значения, могут быть отнесены к надсемейству Р450 автоматически. Значения меньше 100 битовых единиц, но превышающие пороговое значение 50 битовых единиц, установленное для случайно сгененированных последовательностей, принадлежат к так называемой «сумеречной зоне» (twilight zone). В том случае, если значение битового счета попадает в интервал «сумеречной зоны», автоматическое отнесение его в состав надсемейства не производится. В большинстве случаев значения битового счета в диапазоне от 50 до 100 единиц указывают на то, что анализируемая последовательность является фрагментом структуры цитохрома Р450. Указанные формальные критерии, основанные на интервальных оценках счета локального выравнивания, используются в базе знаний для автоматизированной актуализации данных о разнообразии первичных структур цитохромов Р450.

Общие и частные структурно-функциональные мотивы в цитохромах Р450. Возможность проведения автоматического аннотирования с использованием программы локального выравнивания позволяет предложить гипотезу о наличии элементов общего и частного в структуре белков надсемейства цитохромов Р450. Гипотеза основывается на анализе распределений счетов локального выравнивания (рис. 6а). Кроме определения формальных границ надсемейства, распределение

счетов обладает еще одной особенностью, заключающейся в наличии трех минорных пиков (обозначены «*» на гистограмме).

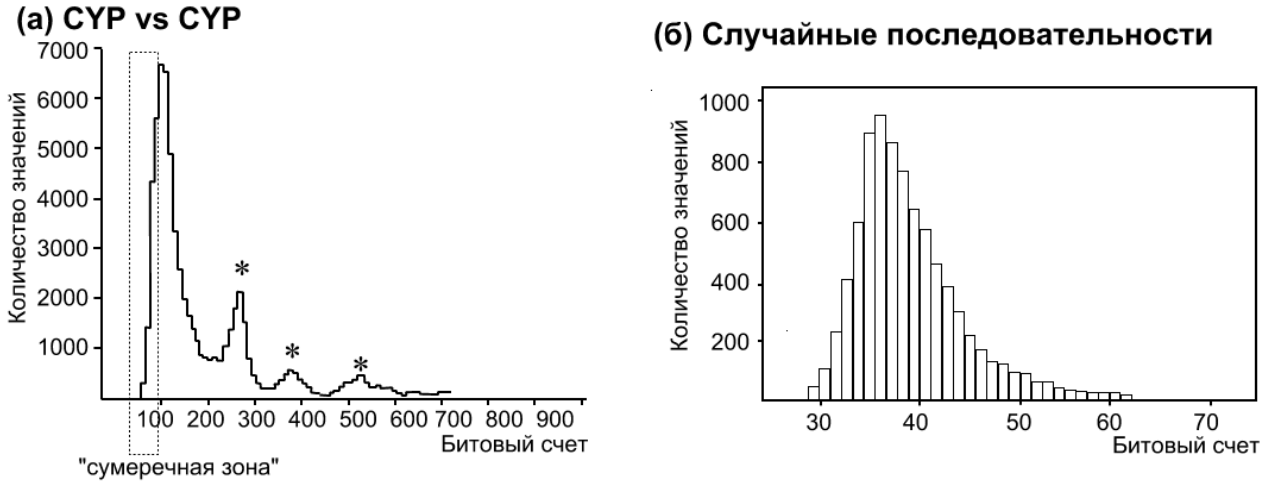

Рис. 6. Распределение счетов локального выравнивания: (а) для надсемейства цитохромов Р450 и (б) для случайно сгенерированных последовательностей сравнимой длины и с сохранением характерных для цитохромов Р450 особенностей аминокислотного состава. Знаком «*» обозначены минорные пики.

Анализ результатов локального выравнивания показывает, что максимальный первый пик формируется, главным образом, за счет совпадения в ультраконсервативных участках спирали I и гем-пептида, присущих все членам надсемейства. С другой стороны, минорные пики формируются за счет локальных участков сходства, характерных для отдельных семейств и подсемейств.

Таким образом, в работе *мотивы общности* были определены как элементы, обеспечивающие структурное единство надсемейства и позволяющие определять цитохромы Р450 среди белков других надсемейств, а *мотивы частного* – как элементы, отвечающие за проявления специфической функциональной активности конкретных форм фермента. Наличие мотивов общего наблюдается в форме наиболее выраженного, первого пика на гистограмме распределения счетов локального выравнивания, а наличие мотивов частного – в форме минорных пиков (рис. 6а).

В основе реализованной в базе знаний процедуры инвентаризации лежит итеративный подход, сходный с подходом, применяемым в программе PSI-BLAST [Altschul et al., 1997]. На каждой последующей итерации выборка, состоящая из аннотированных последовательностей цитохромов Р450, используется в качестве запроса для выявления новых гомологов в глобальных банках данных. Выявленные гомологичные последовательности присоединяются к выборке. Далее проводится кластерный анализ, множественное выравнивание белков в составе кластеров и распознавание структурно-функциональных мотивов (см. далее). Итогом анализа является подтверждение гипотезы о принадлежности белка к надсемейству и размещение его в составе соответствующей классификационной категории.

В работе детальному анализу подвергаются субъективные факторы, которые могут повлиять на отнесение того или иного белка как к самому надсемейству

цитохромов P450, так и к отдельным семействам и подсемействам в его составе. Применение формальных критериев позволяет уточнить классификационный профиль надсемейства цитохромов P450. Так, например, было выявлено, что оптимальное соответствие между составом автоматически сформированных кластеров и номенклатурных категорий – семейств – достигается на уровне 35% идентичности, что несколько ниже, чем общепринятый порог 40%.

В ходе изучения особенностей различных алгоритмических подходов к объективизации результатов кластерного анализа были получены результаты, свидетельствующие о неприменимости ряда устоявшихся концепций молекулярной эволюции к данному надсемейству белков. Кластерный анализ, проведенный в соответствии со степенью сходства первичных структур, не позволяет восстановить филогенетическую историю развития видовой специфичности. Несмотря на наличие узлов бифуркации, соответствующих событиям дивергенции классов, отрядов и т.д., эффекты конвергенции и латеральной диффузии генов создают фоновый «эволюционный» шум, что выражается в необратимых искажениях при проекции дендрограммы кластеризации на временную шкалу. В текущей номенклатуре надсемейства цитохромов P450 обнаруженный эффект, по-видимому, выражается в наличии существенного количества отклонений от постулированных формальных правил классификации.

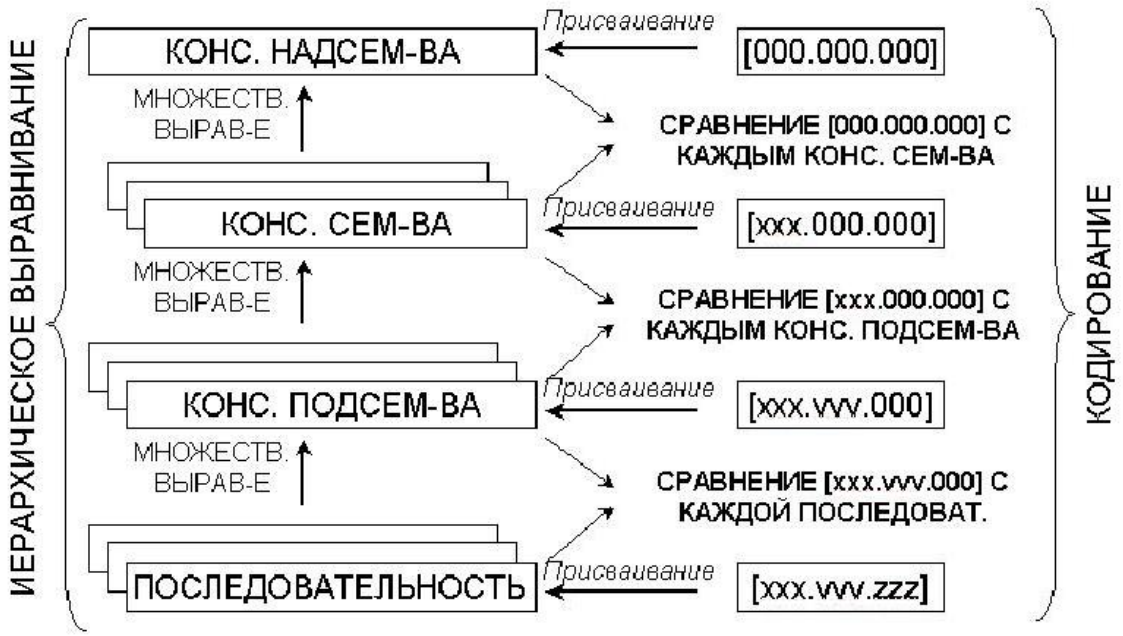

Рис. 7. Алгоритм индексации белкового надсемейства (*см. описание в тексте).*

Подход, развиваемый в рамках работы и получивший название *индексации* надсемейства, позволяет преодолеть ограничения молекулярно-эволюционных позиций, уравнивающих дендрограмму кластерного анализа с эволюционным древом. На рис. 7 отображена общая стратегия индексации, включающая применение иерархического выравнивания к результатам кластерного анализа.

Процедура индексации подразумевает последовательное замещение первичных структур, входящих в кластеры, реконструированными предковыми последовательностями, полученными путем построения консенсуса множественного выравнивания. При таком подходе удается минимизировать искажения, обусловленные не только уже упоминавшимися факторами (дрейф и конвергенция генов), но и различным числом последовательностей в группах цитохромов Р450.

При замещении группы родственных структур, кодирующих ферменты, одной консенсусной последовательностью, предполагается, что последняя несет в себе структурные компоненты, необходимые для реализации ферментной функции, обеспечивающей родство белков группы. Следовательно, можно допускать наличие сходных черт (структурных элементов) у гипотетической последовательности-прародителя группы.

В работе доказывается правомочность вышеизложенной трактовки понятия консенсусной последовательности в свете современных сведений о молекулярных механизмах монооксигеназного катализа. При этом обосновывается (для случая надсемейства цитохромов Р450) отказ от молекулярно-эволюционного подхода, оперирующего понятиями дупликации и последующей дивергенции генов, в пользу структурно-функциональной классификации кодируемых продуктов - белков. Указание на структурные и функциональные особенности белков в качестве основы для систематизации надсемейства нашло свое отражение в термине *протеомный индекс* надсемейства.

Каждому объекту индексации присваивается трехпозиционный код (индекс). В индексации принимают участие объекты четырех типов – консенсус надсемейства (является началом координат и имеет индекс 00.00.00), консенсусы семейств, консенсусы подсемейств и последовательности белков. Консенсусы генерируются при помощи иерархического множественного выравнивания. Затем, парное выравнивание применяется для вычисления расстояния между соподчиненными уровнями иерархии надсемейства. Для консенсусов семейств оценка сходства с консенсусом надсемейства позволяет присвоить первую позицию индекса. Чем большее значение индекса – тем более удален консенсус семейства от консенсуса надсемейства. Степень удаленности консенсусов подсемейств от консенсусов семейств в свою очередь используется для определения второй позиции индекса, а степень удаленность белков от консенсусов подсемейств – для определения третьей позиции индекса.

На рис. 8а видно, что при использовании подхода инвентаризации в консенсусных последовательностях сохраняются основные структурно-функциональные элементы, обеспечивающие целостность пространственной организации белка и специфичность каталитической активности. К таким участкам относятся: трансмембраный домен и пролиновый «якорь» (для белков, локализованных в мембране ЭПР либо в митохондриях), участок альфа-спирали С, обеспечивающий (предположительно) транспорт электронов в активный центр фермента, фиксирующий аппарат гема и др.

С другой стороны, упорядочивание белков надсемейства по удаленности их первичной структуры от реконструированного прародителя-консенсуса, отвечает, в определенной мере, современным представлениям об эволюционном развитии живой

природы: наибольшее число бактериальных цитохромов Р450 (семейства 1-50) оказываются наиболее близкими к консенсусной последовательности надсемейства. По мере уменьшения степени структурного родства консенсусов семейств с консенсусом надсемейства, доля бактериальных цитохромов уменьшается и растет доля консенсусов цитохромов Р450 животных (рис. 8б).

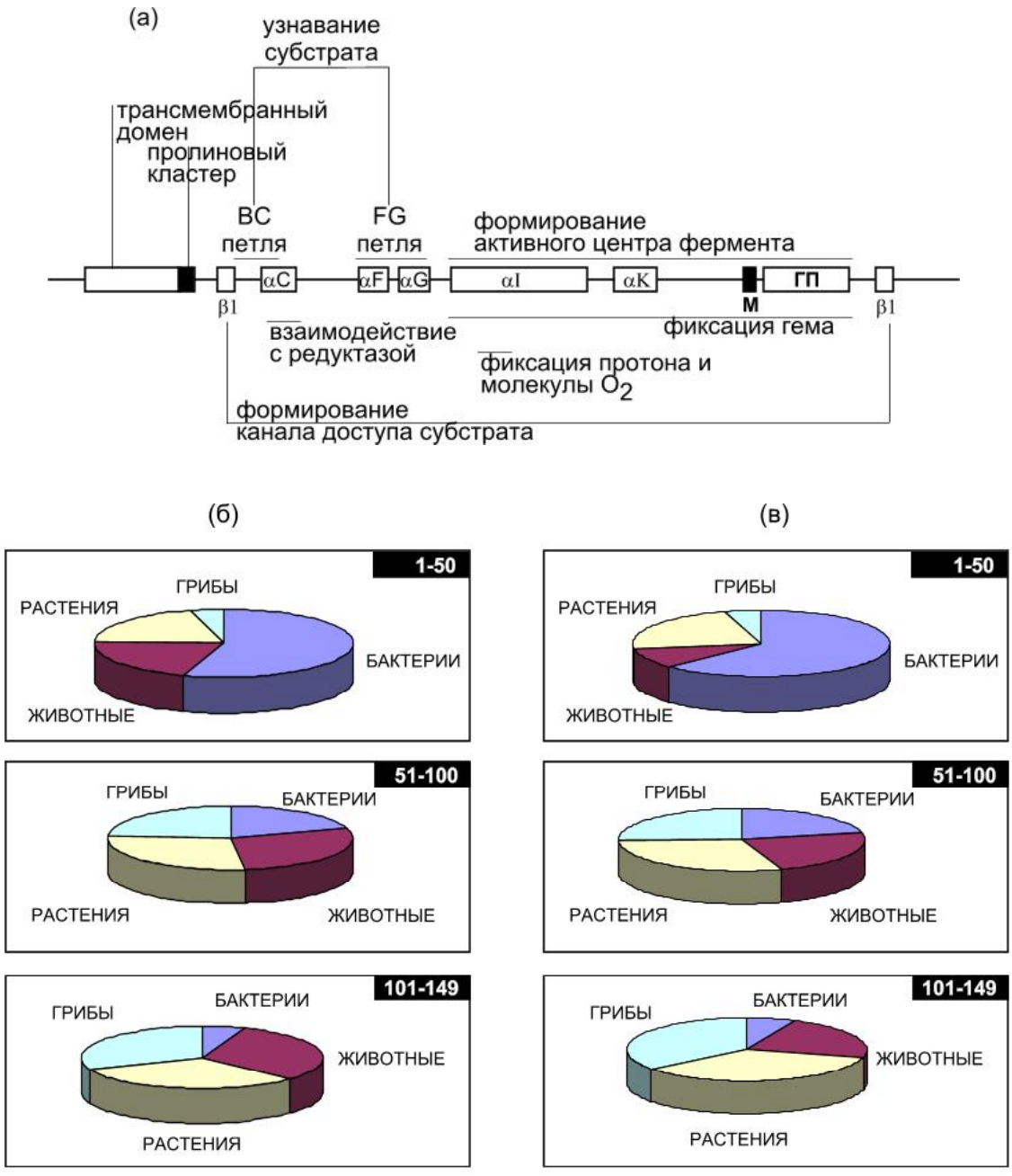

Рис. 8. Результаты индексации надсемейства цитохромов Р450: (а) структурные элементы, выявленные в консенсусной последовательности надсемейства с их функциональной аннотацией: М - меандр, ГП - гем-пептид, α – альфа-спираль, β – бета-структура; (б) реконструкция эволюционных взаимоотношений царств живой природы на основе протеомного индекса, (в) – то же, но без включения консенсусов семейств *C.elegans*.

На рис. 8б показано, как представители различных царств живой природы распределились по частям индекса. Бактериальные, животные и растительные цитохромы Р450 представлены одинаковым количеством кластеров в индексе (~29%), при этом для цитохромов дрожжевого происхождения имеется только 19 кластеров (~13%). Для достижения сбалансированной картины при подготовке диаграмм на рис. 8 каждый кластер дрожжей считался за два.

Таким образом, цитохромы Р450 бактериального происхождения расположены наиболее близко к консенсусу надсемейства. Возможно, что в ходе молекулярной эволюции в бактериях появились белки, которые соответствуют белкам, существовавшим на ранних стадиях эволюции эукариот. В этом смысле, бактериальные цитохромы Р450 можно рассматривать в качестве предшественников эукариотических форм. Безусловно, следует оговориться, что бактериальные цитохромы Р450 лишь выглядят как предшественники в силу специфических особенностей эволюционирования.

На другом конце шкалы в определенной степени превалируют цитохромы Р450 грибов и высших растений. Высшие растения по оценкам появились около 0,5 млрд. лет назад, приблизительно одновременно с ракообразными. В этом случае тоже прослеживается корреляция между молекулярной эволюцией и макроэволюцией.

Делать какие либо обобщения в отношении филогенетического расположения представителей царства грибов преждевременно, т.к., во-первых, количество известных форм цитохромов Р450 грибов невелико, а, во-вторых, имеющиеся формы получены исключительно из дрожжей.

Группа кластеров цитохромов Р450 животных не заняла доминирующего положения ни в одной из частей индекса. Это можно объяснить, если учесть тот факт, что 40% кластеров животных сформированы цитохромами Р450 *C.elegans*. Прародители современных нематод возникли более 1,5 млрд. лет назад, и, с учетом этого, интересно отметить, что 61% всех животных кластеров первой трети индекса составляют именно кластеры *C.elegans*. В остальных двух частях индекса доля кластеров *C.elegans* не превышает 20% (рис. 8в).

Полученные в ходе инвентаризации данные обозначили проблему фундаментального характера, заключающуюся в определении понятия кластера белков надсемейства цитохромов Р450. Граничные условия разделения кластеров, очевидно, влияют как непосредственно на состав консенсусных последовательностей, так и на общее распределение белков по отношению к гипотетическому прародителю. Далее рассматриваются результаты, полученные в ходе решения задачи нахождения оптимального разбиения надсемейства на кластеры с привлечением концепции структурно-функциональных мотивов.

3.4.2 Структурно-функциональные мотивы и их применение

Выявление структурно-функциональных мотивов. В работе проводится сравнительный анализ различных подходов к определению в наборе первичных структур белков статистически значимых участков локального сходства. Сравнительному анализу подвергаются: критерий серий, метод локального выравнивания, и оригинальный метод, основанный на статистическом критерии Шермана [Sneath, 1998].

Критерий серий [Айвазян с соавт., 1983] является упрощенным способом оценки характера распределения консервативных остатков в консенсусе, основанным на оценке наиболее протяженного непрерывного участка сходства между набором выровненных первичных структур. Результаты, полученные с применением критерия серий, показали, что, в случае наличия четкой гомологии, критерий пригоден для градации между группами родственных и неродственных белков (в качестве родственных использовались последовательности, относящиеся к одному семейству цитохромов Р450, в качестве неродственных – относящиеся к разным семействам). Однако недостатки критерия проявляются в том случае, если сходство между родственными последовательностями неочевидно.

Методы, основанные на оценке совокупности участков локального выравнивания, оказались более адекватны для обнаружения сходства последовательностей белков надсемейства цитохромов Р450. В основу правомочности этого утверждения легли данные о корреляции между численными оценками характера распределения консервативных остатков в консенсусной последовательности, полученными при помощи программы локального выравнивания BLAST, и аналогичными оценками, сделанными при помощи критерия Шермана (см. рис. 9б). Расчет критерия Шермана заключается в статистической оценке характера распределения консервативных остатков в консенсусной последовательности, т.е. в обобщенном представлении результатов множественного выравнивания, как совокупности участков локального сходства. При адаптации указанного критерия к задачам диссертационной работы были внесены алгоритмические изменения; правомочность внесенных изменений подтверждается полученными результатами [Lisitsa et al., 2003].

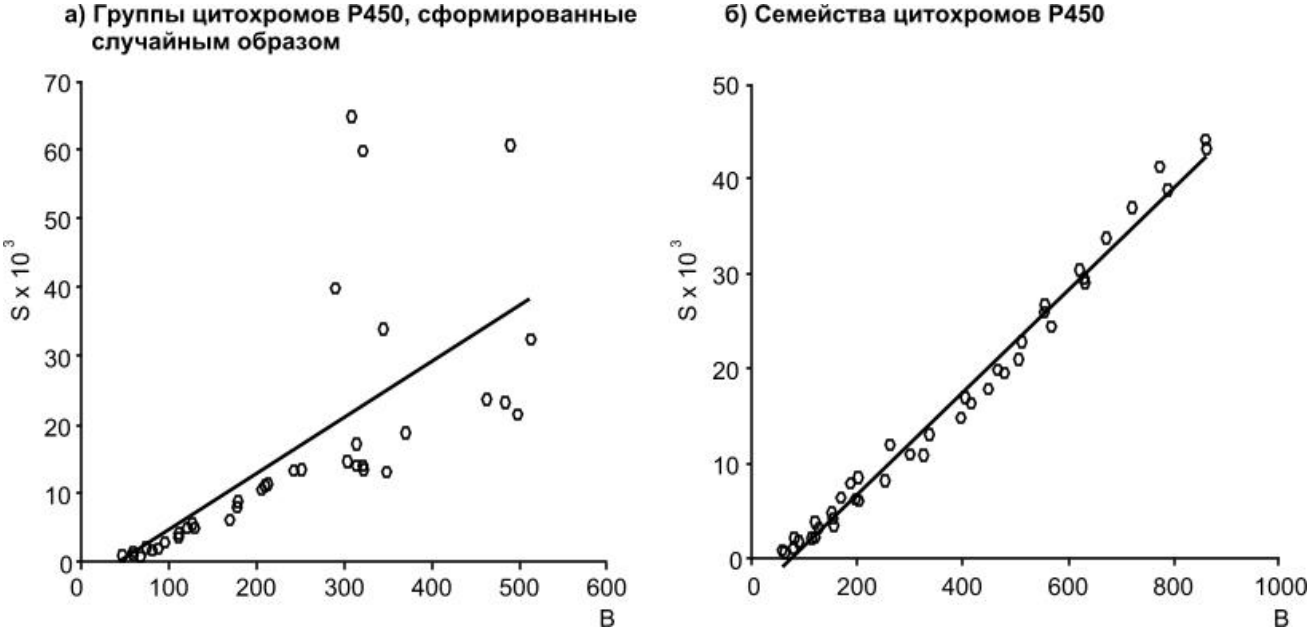

Рис. 9. Корреляция между счетом локального выравнивания (B) и оценкой, вычисляемой на основе критерия Шермана (S), для различных выборок: (а) выборка, содержащая представителей различных семейств цитохромов Р450 и (б) выборки цитохромов Р450, относящихся к одному семейству.

Для доказательства значимости гипотезы о наличии в последовательностях белков надсемейства цитохромов Р450 двух типов структурно-функциональных мотивов была разработана специальная методика. Основу разработанной методики составили методы корреляционного анализа. Корреляция между результатами, получаемыми двумя независимыми алгоритмическими методами (BLAST и критерий Шермана), рассматривалась в качестве подтверждения гипотезы о применимости обоих критериев для выявления структурно-функциональных мотивов.

На рис. 9а показано, как при формировании случайных выборок последовательностей нарушается корреляция между согласованными оценочными критериями. Включение в выборку представителей различных семейств приводит к нарушениям корреляционной зависимости между двумя алгоритмами (коэффициент линейной корреляции – 0,47). С другой стороны, представители одного семейства, демонстрируют хорошую согласованность на всей области определения (коэффициент линейной корреляции – 0,98). Эти данные указывают на применимость оценок, основанных на анализе локального сходства последовательностей, для определения границ семейств цитохромов Р450.

Критерий наличия двух типов структурно-функциональных мотивов в консенсусных последовательностях семейств и подсемейств позволяет провести уточнение границ групп, получаемых в результате кластерного анализа. В работе рассматривается несколько подходов, позволяющих определить границы кластеров. В качестве наиболее простого подхода употребляется L-метод (метод анализа динамики агломерации), основанный на анализе зависимости числа кластеров от выбранного уровня отсечения. При аппроксимации зависимости двумя прямыми в качестве оптимального уровня отсечения выбирается точка их пересечения. Эта точка соответствует выраженному изменению скорости слияния объектов в кластеры и может служить граничным условием при проведении иерархического кластерного анализа.

Табл. 4. Границы семейств и подсемейств в составе надсемейства цитохромов Р450, установленные с применением L-метода. N/A – не определен.

Уровень отсечения	**Над-семейство**	**Белки животных**	**Белки бактерий**	**Белки грибов**	**Белки растений**
Семейства	39%	30%	39%	31%	N/A
Подсемейства	62%	64%	N/A	N/A	

Применение L-метода к выборке структур цитохромов Р450 не привело к получению однозначно интерпретируемого результата [Lisitsa et al., 2003]. Скорость агломерации была монотонна на всем протяжении области определения; соответственно, результирующая аппроксимация прямыми характеризовалась низкой степенью достоверности. Изучение данного феномена показало, что причина неприменимости L -метода для задачи выявления групп в составе надсемейства

цитохромов Р450, заключается в неоднородности законов кластерообразования в пределах выборки. Так, для выборки, состоящей из цитохромов Р450 животных, граница отсечения семейств проходит на уровне 30%, тогда как для белков низших грибов эту границу следует установить на уровне 39% (см. табл. 4). В группе белков растений границы кластеров размыты и не выявляются L-методом. Значения для всего надсемейства – 39% для семейств и 62% для подсемейств, - являются суперпозицией уровней отсечения различных групп. Эти границы также нечетко определяются L-методом.

Табл. 5. Уровни отсечения кластеров, установленные для семейств цитохромов Р450 с использованием различных методов.

№	Метод	Граница семейства (% идентичности последовательностей)	Индекс соответствия с номенклатурой (индекс Джаккарда), %
1.	Индекс Дэвиса-Болдина	39	67
2.	L-метод (динамика агломерации)	39	68
3.	Наилучшее соответствие с номенклатурой при фиксированном уровне отсечения	35	80
4.	Критерий структурно-функциональных мотивов (адаптируемый уровень отсечения)	15-43	84

С учетом полученных данных, дальнейшая работа заключалась в нахождении более приемлемого способа установления границ кластеров, чем L-метод. Была апробирована группа подходов, основывающихся на анализе статистических свойств дендрограммы кластерного анализа. В качестве показательного примера такого рода подхода в работе рассматривается индекс Дэвиса-Болдина (метод №1 в табл. 5). Индекс основан на сравнении статистических характеристик распределения расстояний между объектами, входящих в состав одного кластера, с характеристиками распределения расстояний до объектов вне кластера.

Применение индекса Дэвиса-Болдина хотя и не принесло положительных результатов, но еще более четко обозначило основную особенность кластеризационного профиля надсемейства. Минимум значений индекса, отвечающий оптимальной границе кластера, располагался на различных уровнях отсечения, в зависимости от конкретного семейства. Таким образом, было установлено, что граница отсечения индивидуальна для каждого семейства цитохромов Р450; следовательно, оптимальный критерий отсечения должен быть взаимосвязан со структурно-функциональными свойствами ферментов.

Алгоритмизировать выявленную взаимосвязь удается в рамках развиваемой в работе концепции структурно-функциональных мотивов. Заключение о правомочности выделения кластеров делается на основании наличия мотивов в

консенсусной последовательности множественного выравнивания. Мотивы выявляются путем применения критерия Шермана, на основании которого рассчитывается информационное содержание, как отношение значений критерия, полученных для консенсусной последовательности, и значений, полученных для строки, сгенерированной путем инвертирования консенсуса. (Под инвертированием следует понимать формальную замену вариабельных позиций консервативными и наоборот.)

Критерием качества кластеризации служит степень соответствия между составом номенклатурных групп (семейств и подсемейств цитохромов Р450) и составом кластеров, полученных в автоматическом режиме. Из табл. 5 видно, что применение критерия структурно-функциональных мотивов позволяет добиться наилучшего соответствия.

Для интерпретации данных, приведенных в табл. 5, следует указать, что методы нахождения уровня отсечения №1, №2 и №4 работают в режиме «без обучения», т.е. алгоритмы работают без привязки к существующей номенклатуре. Алгоритм №3 наоборот, ориентирован на существующую номенклатуру, однако, даже в этом случае его применение показывает худший результат (80%), чем предлагаемый нами метод №4 (84%). Принципиальная новизна метода №4 заключается в том, что анализ структурно-функциональных мотивов позволяет привлечь дополнительную информацию для уточнения границ кластеров; в результате линия отсечения варьирует в диапазоне от 15 до 43% идентичности (см. табл. 5).

В работе понятие структурно-функциональных мотивов вводится путем развития сложившихся представлений о способах оценки степени сложности последовательностей, кодирующих биологические макромолекулы. Рассматриваются современные подходы к оценке информационного содержания структур белков и ДНК; отмечается неприменимость стандартных подходов, основанных на расчетах в рамках шенноновского определения количества информации. Выдвигается гипотеза о наличии локальных элементов сходства (мотивов или «островов»), определяющих информационное содержание последовательности аминокислотных остатков. Предлагается способ выявления мотивов путем статистического анализа характера распределения консервативных остатков в составе консенсусной последовательности множественного выравнивания.

С точки зрения термодинамики, определяющей сборку пространственной структуры белка, концепция структурно-функциональных мотивов отвечает представлениям так называемой «островной» гипотезы [Nishikawa, 1993]. Согласно этой гипотезе разрешенные структуры белков, т.е. те первичные структуры, которые могут преодолеть термодинамический барьер фолдинга и принять определенную пространственную конформацию, являются относительно редким явлением, по сравнению со значительно большим числом возможных структур, не способных к фолдингу. Таким образом, в ходе молекулярной эволюции для обеспечения необходимых ферментативных функций белок должен консервативно сохранять некий остов – фолд-детерминирующую основу. Одновременно, специализация фермента в отношении новых функций приводит к возникновению мутаций, не затрагивающих фолд-детерминирующую основу, но обеспечивающую

специфичность взаимодействия с лигандом и избирательность каталитической активности.

Вышеизложенная общая концепция нашла свое подтверждение в рамках проделанной работы по изучению структурно-функциональных особенностей белков надсемейства цитохромов Р450. Выявленные в ходе инвентаризации надсемейства общие мотивы, по-видимому, и являются компонентами фолд-детерминирующей основы белков надсемейства. Мотивы общего обеспечивают такие базовые функции, как фиксация гема, закрепление в мембране эндоплазматического ретикулума (для микросомальных форм цитохромов Р450), фиксация молекулярного кислорода в каталитическом центре, взаимодействие с редокс-партнерами (см. рис. 8а).

С другой стороны, частные мотивы являются продуктом специализации представителей различных семейств цитохромов Р450 в отношении уникальных функций. В это утверждение доказывается: а) путем анализа мотивов семейства стероловых деметилаз; б) путем коррекции результатов кластерного анализа на основании критерия структурно-функциональных мотивов. В ходе изучения семейства стероловых деметилаз было показано, что частные мотивы расположены в элементах структуры белка, участвующих в узнавании субстрата [Лисица, 2004]. Распространение этого наблюдения на другие семейства цитохромов Р450 следует из повышения степени соответствия между составом номенклатурных групп и результатами кластерного анализа, достигнутого в результате применения критерия структурно-функциональных мотивов (см. табл. 5).

Анализ двух типов структурно-функциональных мотивов может быть предложен в качестве метода прогнозирования функциональной специфичности новых форм цитохромов Р450, выявляемых в геномах. Кроме функционального аннотирования, мотивы могут быть использованы в качестве элементов для конструирования химерных форм цитохромов Р450 с заданными функциями. Обе сферы применения мотивов представляют интерес с точки зрения современных задач биотехнологии.

3.5 Заключение: взаимосвязь между структурой белка и его субстратной специфичностью как основная проблема биоинформационных исследований надсемейства цитохромов Р450

База знаний по цитохромам Р450 позволяет выявить центральную научную проблему, связанную с данным надсемейством белков. Очевидно, проблема является общим отражением ситуации в молекулярной биологии, которая сложилась в связи с экспоненциальным ростом количества расшифрованных геномов. Как отображено на рис. 10, рост количества генов, кодирующих цитохромы Р450, является экспоненциальным. Так, в 1991 г. сиквенирование одного гена являлось трудоемким процессом, и каждый найденный ген, кодирующий новую форму цитохрома Р450, публиковался в виде отдельной статьи. Поиск и сиквенирование нового гена занимало количество времени, сравнимое с объемом исследования по выделению кодируемого геном белка и характеристики его специфических свойств. Как

следствие, количество функционально охарактеризованных цитохромов Р450 только в 2 раза уступало общему количеству известных генов.

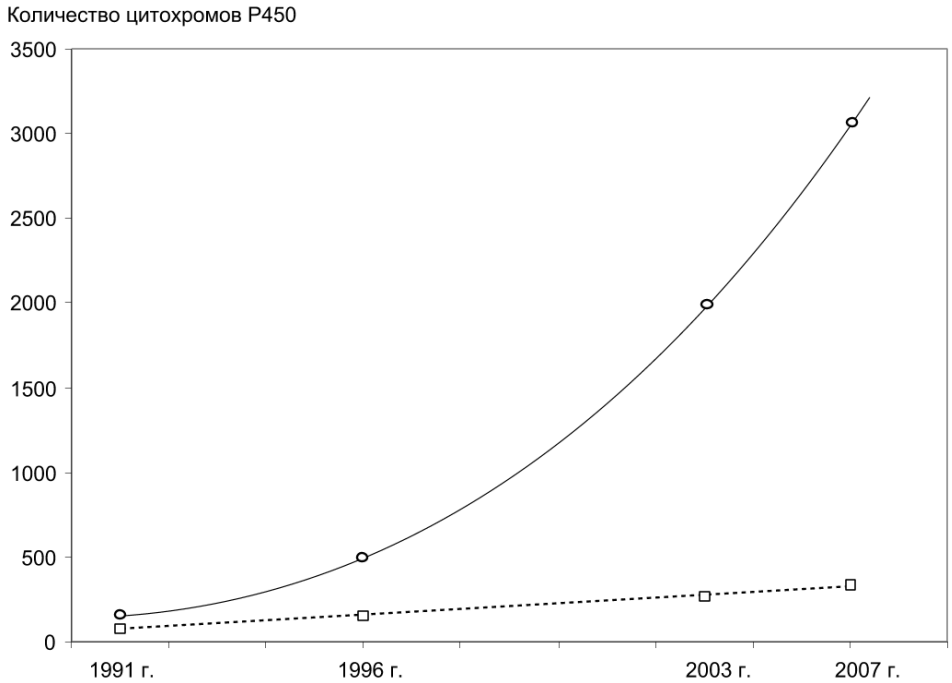

Рис. 10. Увеличение объемов информации о надсемействе цитохромов Р450:

O - количество расшифрованных первичных структур [Nelson et al., 1991; Nelson et al., 1996; Estabrook, 2003; Nelson, 2006];

□ - количество форм цитохромов Р450, для которых известна субстратная специфичность по отношению как минимум к одному химическому соединению.

Повышение эффективности методических подходов к сиквенированию геномов привело к тому, что, начиная с 1996 г., нарастает разрыв между количеством известных генов и охарактеризованными белками. Как следствие в глобальных банках данных быстро накапливаются сведения о новых формах цитохромов Р450 в составе цельных геномов. Информация о способах клонирования, выделения кодирующей ДНК и препарата белка, не говоря уже о функциональных характеристиках фермента, отсутствует.

В области получения препарата белка и изучении его функциональных свойств за прошедшие 15 лет методический прогресс был не так ощутим, как в технологиях расшифровки геномов [Cham, 2005; Bennet et al., 2005]. Как отображено на рис. 10, накопление данных о функциональных свойствах белков надсемейства происходит медленно, и за прошедшие 15 лет количество изученных с точки зрения субстратной специфичности форм увеличилось менее чем в 5 раз, в то время, как количество обнаруженных генов цитохромов Р450 возросло более чем в 100 раз.

В действительности ситуация представляется даже более сложной, поскольку существует множество форм цитохромов Р450, каждая из которых способна окислять несколько или даже десятки субстратов [Werk-reichhart, 2000]. Так, например, в настоящее время для каждой из 334 форм цитохромов Р450 известны 1 и более субстратов, но лишь для каждого из 79 цитохромов Р450 известно более 5 субстратов.

Ряд исследователей [Korolev et al., 2003; Lewis et al., 2006; Borodina et al., 2004] ведут работы по вычислительному прогнозированию каталитически активной формы фермента по заданной структурной формуле химического соединения. Для этого используются различные варианты методов анализа взаимоотношений структура-активность (QSAR, см. рис. 11а). Однако, как следует из вышесказанного, область такого рода исследования довольно ограниченная – статистически достоверные данные могут быть получены всего для 79 белков, при их общем количестве более 3 тыс. Уровень предсказательной достоверности указанной группы методов невысок, и колеблется в диапазоне 60-70%, что определяется перекрестной субстратной специфичностью, присущей многим формам цитохрома Р450.

Рис. 11. Традиционные подходы к установлению взаимосвязи структура-функция (а) и необходимый путь дальнейшего развития биоинформационных методов в области надсемейства цитохромов Р450 (б).

Не отрицая важность прогнозирования профиля взаимодействия химического соединения с цитохромами Р450 с точки зрения задач исследования фармакокинетики прототипов новых лекарств, хотелось бы, опираясь на данные рис. 10, привлечь внимание к наиболее важной, прямо противоположной проблеме, схематически изображенной на рис. 11б. В данном случае исследователь исходно располагает информацией только о первичной структуре цитохрома Р450, расшифрованной на основании анализа геномной информации. Вполне вероятно, что предсказанный белок является функционально активным, т.е. может быть использован для решения прикладных задач биотехнологии, создания биологически активных соединений, мониторинга окружающей среды и проч. Однако, на пути практического применения сведений о первичной структуре новой формы цитохрома Р450 стоит препятствие, заключающееся в большой трудоемкости экспериментальных методов исследования, направленных на выявление функции белка. Преодолеть этот барьер возможно только с использованием вычислительных методов прогнозирования, которые позволили бы за счет анализа особенностей последовательности аминокислотных остатков сделать

выводы о субстратной специфичности, предсказать наиболее вероятные каталитические реакции и их продукты.

В настоящее время однозначных подходов к решению указанной проблемы не существует. Использование базы знаний для проведения инвентаризации, индексации, поиска общих и частных мотивов позволило выявить ограничения методов, основанных на алгоритме выравнивания. Дальнейшее развитие исследований в направлении предсказания функций новых форм цитохромов Р450, по-видимому, должно быть связано с дополнительными аналитическими алгоритмами, такими как, молекулярное моделирование, анализ распределения аминокислотных остатков [Otaki et al., 2006], изучение закономерностей скрытой периодичности [Turutina et al., 2006], выявление минимальных модульных повторов [Barney, 2006] и проч.

4. ВЫВОДЫ

1. Разработана информационная система – база знаний, обеспечивающая интегрированную платформу для хранения и анализа информации о структурно-функциональных особенностях белков надсемейства цитохромов Р450.

2. В состав базы знаний включены методы обработки информационного массива и разработана технология применения этих методов в научно-исследовательской работе.

3. С использованием базы знаний получены следующие научные результаты:
а) разработаны подходы к кластерному анализу (инвентаризации) надсемейства цитохромов Р450;
б) метод индексации предложен в качестве объективного способа упорядочивания белков надсемейства по степени их родства с реконструируемыми консенсусами-предшественниками;
в) установлено, что последовательностям белков надсемейства цитохромов Р450 присущи мотивы двух типов: мотивы общего характера, обеспечивающие единство фолда и механизмов катализа, и мотивы частного характера, обеспечивающие специфичность функциональной активности.

4. Предложен новый подход к классификации белков надсемейства цитохромов Р450, основанный на анализе мотивов общего и частного характера в первичных структурах этих белков.

5. Показано, что общая тенденция накопления данных в отношении надсемейства характеризуется прогрессирующим отставанием объема сведений о функциональной активности от объема информации о расшифрованных генах, кодирующих цитохромы Р450. Задача прогнозирования функциональной активности новых форм определена как основная проблема в дальнейшем развитии биоинформационных исследований в области надсемейства цитохромов Р450.

5. СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Archakov A.I., Ivanov A.S., Lisitsa A.V., Rukavishnikov I.G. Leucine clusters (L) 3+N as sites of interaction of microsomal cytochromes P450 with the membrane phospholipids. //In: Proceedings of 7-th International Conference “Biochemistry & Biophysics of cytochrome P450: Structure & Function, Biotehnology & Ecological Aspekts. (Archakov A.I., Bachmanova G.I., eds).-INCO-TNC Joint Stock Company.-1992.-P.716-718.
2. Archakov A.I., Bachmanova G.I., Sandler M.K., Tutochkin I.Y., Lisitsa A.V. Cytochrome P450 database and its scientific application. //In: Proceedings of 7-th International Conference “Biochemistry & Biophysics of Cytochrome P450: Structure & Function, Biotehnology & Ecological Aspekts. (Archakov A.I., Bachmanova G.I., eds).-INCO-TNC Joint Stock Company.-1992.-P.-673-679.
3. Archakov A.I., Degtyarenko K.N., Lisitsa A.V. Common motifs in microsomal cytochrome P450 N-terminal membrane fragment of cytochrome P450 LM2. //J. Basic & Clinical Physiology & Pharmacology.-1992.-P.97-98.
4. Lisitsa A.V., Bachmanova G.I., Archakov A.I. Cytochrome P450 database for prediction of drugs fate in living systems. //In: Abstr. 9-th International Conference “Cytochrome P450: Biochemistry, Biophysics and Molecular Biology”.-Zarich.-1995.-P.173.
5. Archakov A.I., Bachmanova G.I., Lisitsa A.V., Sandler M.K. Prediction of drugs fate by using cytochrome P450 database. //In: Abstr. of 3-th IUBMB Company Molecular Recognition.-Singapore.-1995.-P.103.
6. Archakov A.I., Lyashenko A.A., Lisitsa A.V., Koymans L. Cytochrome P450 database and its usage for analysis of structural functional domains and substrate specificity. //Exp. Toxic. Pathol.-1996.-V.48(5).-P.329-330.
7. Archakov A.I., Lisitsa A.V., ZgodaV.G., Koymans L. The determination of the cytochrome P450 superfamily frontiers. //In: Abstr. 12-th International symposium on microsomes and drug oxidations.-Montpellier France Le Corum.-1998.-P.419.
8. Gusev S., Archakov A., Lisitsa A., Zgoda V., Koymans L. N-Tuple alignment of biological texts. //In: Abstr. 12-th International symposium on microsomes and drug oxidations.-Montpellier France Le Corum.-1998.-P.420.
9. Lisitsa A.V., Archakov A.I., Koymans L. Automatic procedure for the proteins clusterization applied to the cytochrome P450 superfamily. //In: Abstr. 12-th International symposium on microsomes and drug oxidations.-Montpellier France Le Corum.-1998.-P.421.
10. Archakov A.I., Lisitsa A.V., Zgoda V.G., Ivanov A.S., Koymans L. Clusterization of P450 superfamily using the objective pair alignment method and the UPGMA program. //J. Mol. Model.-1998.-V.4.-P.234-238.
11. Lisitsa A., Gusev S. Cytochrome P450 database and its scientific application. //In: Abstr. International workshop “From Sequence to function: Experimental and Bioinformatic Studies of Cytochrome P450 Superfamily”.-Moscow.-2000.-P.21.
12. Gusev S., Lisitsa A. Creation of structural functional map for P450 proteins. //In: Abstr. International workshop “From Sequence to function: Experimental and Bioinformatic Studies of Cytochrome P450 Superfamily”.-Moscow.-2000.-P.25.

13. Ivanov A., Dubanov A., Skvortsov V., Gusev S., Lisitsa A., Archakov A. Genome analysis and computer modelling of cytochromes P450 from *Mycobacterium tuberculosis*. //In: Abstr. International workshop "From Sequence to function: Experimental and Bioinformatic Studies of Cytochrome P450 Superfamily".-Moscow.-2000.-P.26.
14. Gusev S.A., Lisitsa A.V., Karuzina I.I., Archakov A.I. Cytochrome P450 Database. //In: Abstr. 13 International Symposium on Microsomes and Drug Oxidation.-Stresa-Italy.-Satellite Symposium of the VII World Conference on Clinical Pharmacology and Therapeutics.-Florence.-2000.-P.179.
15. Lisitsa A.V., Gusev S.A., Archakov A.I. Structural functional motifs in 14α-demethylase of *Mycobacterium tuberculosis*. //In: Abstr. 13 International Symposium on Microsomes and Drug Oxidation.-Stresa-Italy.-Satellite Symposium of the VII World Conference on Clinical Pharmacology and Therapeutics.-Florence.-2000.-P.180.
16. Лисица А.В., Гусев С.А. Биоинформатика первичной структуры белков. //Вопросы мед. химии.-2001.-Т.47.-С.659-663.
17. Archakov A., Lisitsa A., Gusev S., Koymans L., Janssen P. Inventory of the Cytochrome P450 superfamily. //J.Mol.Model.-2001.-V.7.-P.140-142.
18. Lisitsa A.V., Gusev S.A., Karuzina I.I., Archakov A.I., Koymans L. Cytochrome P450 database. //SAR QSAR Environ Res.-2001.-V.12(4).-P.359-66.
19. Lisitsa A.V., Gusev S.A., Archakov A.I. Application of protein indexing to Cytochrome P450 superfamily. //In: Abstr. 4-th International Conference on Molecular Structural Biology.-Vienna.-2001.-P.81.
20. Lisitsa A.V., Gusev S.A., Archakov A.I. Index of cytochrome P450 superfamily. //In: Abstr. 12-th International Conference on Cytochrome P450. Biochemistry, Biophysics and Molecular Biology.-France.-2001.-P.102.
21. Shumyantseva V.V., Bulko T.V., Petushkova N.A., Lisitsa A.V., Archakov A.I. Specific binding of riboflavin to the cytochrome P450 2B4: fluorometric and spectroscopic studies. //In: Abstr. 12-th International Conference on Cytochrome P450. Biochemistry, Biophysics and Molecular Biology.-France.-2001.-P.116.
22. Gusev S.A., Lisitsa A.V., Archakov A.I. Structural-functional motifs of cytochromes P450. //In: Abstr. 12-th International Conference on Cytochrome P450. Biochemistry, Biophysics and Molecular Biology.-France.-2001.-P.119.
23. Archakov A.I., Lisitsa A.V., Gusev S.A., Govorun V.M. Proteomic indexing of Cytochrome P450 Superfamily. //In: Abstr. International Meeting on Proteome Analysis.-Munchen.-2001.-P.133.
24. Lisitsa A.V. Cytochrome P450 Database: From data to knowledge. //In: Abstr. International Conference Genomics and Bioinformatics for Medicine.-St.Peterburg-Moscow.-2002.-P.52.
25. Borodina Ju.V., Lisitsa A.V., Poroikov V.V., Filimonov D.A, Sobolev B.N., Archakov A.I. If there exists correspondence between similarity of substrates and protein sequences in cytochrome P450 superfamily? //In: Abstr. International Conference Genomics and Bioinformatics for Medicine.-St.Peterburg-Moscow.-2002.-P.76.
26. Archakov A.I., Karuzina I.I., Petushkova N.A., Lisitsa A.V., Zgoda V.G. Production of carbon monoxide by cytochrome P450 during iron-dependent lipid peroxidation. //Toxicology in vitro.-2002.-V.16.-P.1-10.

27. Lisitsa A., Borodina Ju., Filimonov D., Archakov A. Cytochrome P450 Database: from data to knowledge. //In: Abstr. 14th International Symposiun in Microsomes and Drug Oxidation.-Sapporo Japan.-2002.-P.84.
28. Лисица А.В., Мирошниченко Ю.В., Иванов А.С., Арчаков А.И. Общее и частное в структурной организации белков надсемейства цитохромов Р450. //Аллергия, астма и клиническая иммунология.-2003.-№8.-С.14-19.
29. Пономаренко Е.А., Лисица А.В., Карузина И.И., Мирошниченко Ю.В. Автоматизированное аннотирование функциональных свойств белков надсемейства цитохромов Р450. //Аллергия, астма и клиническая иммунология.-2003.-№8.-С.95-99.
30. Арчаков А.И., Канаева И.П., Петушкова Н.А., Згода В.Г., Лисица А.В., Карузина И.И. Создание протеомных карт белков микросом клеток печени и лимфоцитов крови мышей с целью разработки новых диагностических тестов. //Аллергия, астма и клиническая иммунология.-2003.-№9.-С.179-181.
31. Иванов А.С., Скворцов В.С., Сеченых А.А., Дубанов А.В., Лисица А.В. Компьютерное моделирование трехмерной структуры цитохрома Р450. //Биомедицинская химия.-2003.-Т.49(3).-С.221-37.
32. Borodina Y.V., Lisitsa A.V., Poroikov V.V., Filimonov D.A., Sobolev B.N., Archakov A.I. If there Exists Correspondence between Similarity of Substrates and Protein Sequences in Cytochromes P450 Superfamily? //Nova Acta Heopoldina.-2003.-V.329.-P.47-55.
33. Archakov A.I., Lisitsa A.V., Gusev S.A., Miroshnichenko Yu.V. Bioinformatic Insight into the Structural Unity and Diversity of Cytochromes P450. //In: Proceedings 13-th International Conference on Cytochromes P450.-Prague.-2003.-P.7-13.
34. Ivanov A.S., Skvortsov V.S., Lisitsa A.V., Archakov A.I. General trends in 3D modelling of cytochromes P450. //In: Proceedings 13-th International Conference on Cytochromes P450.-Prague.-2003.-P.47-54.
35. Lisitsa A.V, Ponomarenko E.A., Karuzina I.I., Ivanov A.S., Archakov A.I. Balance Sheet for Cytochrome P450 Knowledgebase. //In: Proceedings 13-th International Conference on Cytochromes P450.-Prague.-2003.-P.67-73.
36. Lisitsa A.V., Gusev S.A., Archakov A.I. Motif-based criterion corrects the clustering of protein sequences. //In: Abstr. 5th International Conference on Molecular Structural Biology.-Vienna.-2003.-P.79.
37. Lisitsa A.V., Archakov A.I., Lewi P., Janssen P. Bioinformatic insight into the unity and diversity of cytochromes P450. //Methods and Findings in Experimental and Clinical Pharmacology.-2003.-V.25(9).-P.733-745.
38. Иванов А.С., Скворцов В.С., Сеченых А.А., Дубанов А.В., Лисица А.В. Проблемы и перспективы компьютерного моделирования трехмерной структуры цитохромов Р450. //Сборник материалов Сессии ИВТН-2003.-Москва.-2003.-С.6-7.
39. Иванов Н.А. , Лисица А.В., Пономаренко Е.А., Арчаков А.И. Тематический анализ резюме научных публикаций в области цитохромов Р450. //Сборник материалов Сессии ИВТН-2003.-Москва.-2003.-С.28-29.
40. Lisitsa A.V., Archakov A.I. Bioinformatics of protein primary structure. //In: Annals of the European Academy of Sciences. Khalatnikov I.M. (Ed). EAS Publishing House: Brussels.-2003.-P. 48-74.

41. Лисица А.В., Мирошниченко Ю.В., Пономаренко Е.А. База знаний по цитохромам Р450. //Сборник научных трудов X Российского национального конгресса «Человек и лекарство».-2003.-С.730.
42. Канаева И.П., Петушкова Н.А., Лохов П.Г., Згода В.Г., Карузина И.И., Лисица А.В., Арчаков А.А. Изучение микросом печени мышей с помощью методов протеомного анализа. //Биомедицинская химия.-2004.- Т.50(4).-С.367-75.
43. Lisitsa A.V. Bioinformatic means for the integration of heterogeneous data and methods. //In: Abstr. 2nd International conference "Genomics, Proteomics and Bioinformatics for Medicine".-Moscow-Ples-Moscow.-2004.-P.59.
44. Lisitsa A.V., Archakov A.I. Cytochrome P450 Knowledgebase (CPK). //In: Abstr. 7th International symposium on Cytochrome P450. Biodiversity and biotechnology.-Japan.-2004.-P.43.
45. 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-деметилаз (CYP51). //Биомедицинская химия.-2004.-Т.50(6).-С.555-65.
46. Арчаков А.И., Гусев С.А., Лисица А.В. База данных по цитохромам Р450. //Свидетельство об официальной регистрации базы данных №2004620199.-2004.
47. Kanaeva I.P., Petushkova N.A., Lisitsa A.V., Lokhov P.G., Zgoda V.G., Karuzina I.I., Archakov A.I. Proteomic and biochemical analysis of the mouse liver microsomes. //Toxicol In Vitro.-2005.-V.19(6).-P.805-12.
48. Арчаков А.И., Лисица А.В. Биоинформатика и биоинформационные технологии. //Труды XII Всероссийской научно-методической конференции «Телематика'2005».-Санкт-Петербург.-2005.-Т.1.-С.55-56.
49. Lisitsa A.V., Ponomarenko E.A., Gusev S.A., Kuznetsova G.P., Karuzina I.I., Lewi P., Archakov A.I. Cytochrome P450 knowledgebase: structure and functionality. //In: Proceedings 14th International conference on cytochromes P450: biophysics and bioinformatics.-Dallas, USA.-2005.-P.29-34.
50. Lisitsa A.V. Integrated management of dataflow within the proteomic projects. //In: Abstr. HUPO 4th annual world congress "From defining the proteome to understanding function".-Munich, Germany.-2005.-P.83.
51. Петушкова Н.А., Канаева И.П., Шереметьева Г.Ф., Згода В.Г., Лохов П.Г., Лисица А.В., Карузина И.И., Арчаков А.И. Использование протеомных технологий для выявления и идентификации цитохромов Р450 микросом клеток печени человека. Аллергия, астма и клиническая иммунология.-2005.-№5.-С.11-17.
52. Пономаренко Е.А., Лисица А.В., Карузина И.И., Гусев С.А. База знаний по цитохромам Р450. //Сборник материалов Сессии ИВТН-2006.-Москва.-2006.-С.32.
53. Шумянцева В.В., Булко Т.В., Рудаков Ю.О., Саменкова Н.Ф., Лисица А.В., Карузина И.И., Арчаков А.И. Наноэлектрохимия цитохромов Р450: прямой перенос электронов и электрокатализ. //Биомедицинская химия.-2006.-Т.52(5).-С.458-68.
54. Ivanov A.S., Gnedenko O.V., Molnar A.A., Mezentsev Yu.V., Lisitsa A.V., Archakov A.I. Protein-protein interactions as new targets for drug design: interactive links between virtual and experimental approaches. //In: Abstr. 5rd International conference on bioinformatics of genome regulation and structure.- Novosibirsk.-2006.-V.1.-P.277-281.

55. Petushkova N.A., Kanaeva I.P., Lisitsa A.V., Sheremetyeva G.F., Zgoda V.G., Samenkova N.F., Karuzina I.I., Archakov A.I. Characterization of human liver cytochromes P450 by combining the biochemical and proteomic approaches. //Toxicol In Vitro.-2006.-V.20(6).-P.966-74.
56. Zgoda V., Tikhonova O., Lisitsa A., Archakov A. Proteomic profiles of induced hepatotoxicity at the subcellular level. //In: Abstr. 3rd International conference "Genomics, proteomics, bioinformatics and nanotechnologies for medicine".-Novosibirsk.-2006.-P.65.
57. Archakov A., Lisitsa A. Platform from genomes to drugs – escorting the data-driven drug design. //In: Abstr. 3rd International conference "Genomics, proteomics, bioinformatics and nanotechnologies for medicine".-Novosibirsk.-2006.-P.72.
58. Ivanov A., Molnar A., Lisitsa A., Archakov A. Integration of computer and experimental approaches for discovery of inhibitors of protein interactions. //In: Abstr. 3rd International conference "Genomics, proteomics, bioinformatics and nanotechnologies for medicine".-Novosibirsk.-2006.-P.77.
59. Zgoda V., Tikhonova O., Lisitsa A., Archakov A. Proteomic profiles of induced hepatotoxicity at the subcellular level. //In: Abstr. HUPO 5rd annual world congress.-Long Beach, California.-2006.-P.145.
60. Lisitsa A., Nikitin I., Archakov A., Podoplelov A., Thiele H. Recognizing the proteomic patterns of induced toxicity with 1D-ZOOMER approach. //In: Abstr. HUPO 5rd annual world congress.-Long Beach, California.-2006.-P.146.
61. Арчаков А.И., Лисица А.В., Пятницкий М.А., Руденко В.А., Тихонова О.В. Протей. //Свидетельство об официальной регистрации программы для ЭВМ №2006611941.-2006.
62. Zgoda V., Tikhonova O., Viglinskaya A., Serebriakova M., Lisitsa A., Archakov A. Proteomic profiles of induced hepatotoxicity at the subcellular level. //Proteomics.-2006.-V.6(16).-P.4662-4670.
63. Archakov A.I., Ivanov Y.D., Lisitsa A.V., Zgoda V.G. AFM fishing nanotechnology is the way to reverse the Avogadro number in proteomics. //Proteomics.-2007.-V.7(1).-P.4-9.