На правах рукописи

Деменков Павел Сергеевич

МАТЕМАТИЧЕСКОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ДЛЯ РЕКОНСТРУКЦИИ АССОЦИАТИВНЫХ СЕТЕЙ МОЛЕКУЛЯРНО-ГЕНЕТИЧЕСКИХ ВЗАИМОДЕЙСТВИЙ

05.13.11 — Математическое и программное обеспечение
вычислительных машин, комплексов
и компьютерных сетей

Автореферат
диссертации на соискание учёной степени
кандидата технических наук

Томск — 2008

Работа выполнена в Институте математики им. С. Л. Соболева СО РАН.

Научный руководитель:	доктор физико-математических наук профессор, чл.-корр. РАН Гончаров Сергей Савостьянович
Официальные оппоненты:	доктор технических наук, доцент Тузовский Анатолий Фёдорович (Томский политехнический университет, г. Томск)
	кандидат физико-математических наук Мурзин Фёдор Александрович (Институт систем информатики СО РАН, г. Новосибирск)
Ведущая организация:	Иркутский государственный университет

Защита состоится " 10 " декабря 2008 г. в 14 ч. 30 мин. на заседании совета по защите докторских и кандидатских диссертаций Д 212.269.06 при Томском политехническом университете по адресу: 634034, г. Томск, ул. Советская, 84, институт "Кибернетический центр" ТПУ

С диссертацией можно ознакомиться в Научно-технической библиотеке Томского политехнического университета по адресу: 634034, г. Томск, ул. Белинского, 55.

Автореферат разослан "____"_____ 2008.

Учёный секретарь Совета
кандидат технических наук, доцент Сонькин М. А.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Объект исследования и актуальность темы. Активное применение современных информационных технологий, средств вычислительной техники и методов прикладной математики в области молекулярно-биологичеких и биомедицинских исследований заложили фундаментальную основу развития такого направления как биоинформатика.

Широкомасштабное секвенирование геномов, экспериментальные методы протеомики, геномики и транскриптомики обеспечивают колоссальный рост молекулярно-биологической информации, которую принципиально невозможно осмыслить и переработать без использования специальных программно-информационных средств. Во всем мире интенсивно ведутся исследования в области организации биологических систем и технологий, в частности: высокопроизводительных биочиповых (ДНК-микрочипы, белковые, клеточные и тканевые микрочипы, микрочипы на основе малых молекул); протеомных и метаболомных экспериментальных технологий, широко используемых в биомедицине, фармакологии, биотехнологии, агробиологии и других областях. При этом следует отметить, что в настоящее время активное развитие экспериментальных методов идентификации молекулярных взаимодействий на самых разных уровнях организации биологических систем значительно опережает развитие биоинформатических средств поддержки, анализа и интерпретации результатов экспериментов. Всё большую актуальность приобретают вопросы интеграции результатов анализа и интерпретации молекулярно-генетических данных, состоящие в выяснении связи генов, белков и метаболитов с функционированием молекулярно-генетических систем, с молекулярно-биологическими информационными ресурсами при формировании новых знаний в рассматриваемой области. Следует отметить, что знания о молекулярно-генетических взаимодействиях в клетке необходимы для решения широкого круга практически важных задач в области биотехнологии и агробиологии, биомедицины и фармакологии, в частности:

- поиск мишеней для создания лекарственных препаратов;

- оценка потенциальной эффективности и токсичности новых препаратов в доклинических испытаниях;
- идентификация биомаркерных молекул для создания эффективных диагностических систем;
- идентификация важных для продуктивности сельскохозяйственных культур генов;
- выбор генов-кандидатов для генотипирования.

Создание новых и идентификация существующих знаний, их применение на практике для диагностики, предупреждения и лечения различных заболеваний — одна из целей молекулярно-биологических и биомедицинских исследований, а разработка эффективных систем поддержки этих процессов на основе современных информационных технологий и концепции систем управления знаниями — одна из приоритетных задач биоинформатики.

На современном этапе из-за высоких темпов роста публикаций и электронных баз данных (БД) в области исследований биологических систем и разработки технологий особую актуальность приобретают вопросы создания адекватного инструментария для систематизации проблемной информации и решения задач идентификации существующих знаний. В частности, в условиях большого потока информации становится все сложнее восстанавливать недостающие связи между молекулярно-генетическими объектами, которые могут приводить к практическому использованию накопленных знаний.

Например, БД данных рефератов научных статей по современным исследованиям в области генетики, молекулярной биологии и биомедицины Pubmed содержит около 15 миллионов публикаций на конец 2006 года и их объем увеличивается в среднем на 500 тысяч статей в год. Созданные в мире тысячи фактографических медико-биологических БД содержат разнообразную информацию о биологических объектах и их взаимодействиях на уровне геномов, клеток и организмов. Объёмы этих БД чрезвычайно велики. Так, БД NCBI Gene содержит 1933023 записей (2006 год), количество которых постоянно увеличивается. Существуют базы данных содержащие информацию о полиморфизмах, связанных с заболеваниями человека, животных и рас-

тений (например, база данных OMIM содержит информацию о 17212 генах, связанных с патологиями человека). В базе данных Gene Ontology представлено формализованное описание молекулярных функций белков и генов, процессов, в которых они участвуют (130696 биологических процессов и 128548 молекулярных функций для 107701 клеточных компонент). В базах данных KEGG, EcoCyc, MetaCyc, GeneNet и др. представлены миллионы фактов о биомедицински и биотехнологически значимых молекулярно-генетических взаимодействиях, генных сетях, метаболических путях, путях передачи сигналов и др. Если учесть, что заметная часть информации в БД по данному направлению слабо структурирована и представлена в текстовом виде, то становится ещё более очевидной актуальность соответствующего математического и программного инструментария.

Создание новых и идентификация существующих знаний как базовые виды деятельности в жизненном цикле знаний рассматриваются в качестве системообразущих объектов в системах управления знаниями (СУЗ). Активные исследования в области создания СУЗ начались с 90-х годов прошлого столетия. Среди авторов публикаций следует выделить исследования O. Bodenreider, K.M. Wiig, T.H. Davenport, L. Prusak, S.B. Martins, H. Takeuchi, J.M. Firestone, I. Nonaka, С.М. Климова, Т.А. Гавриловой, А.Ф. Тузовского и В.З. Ямпольского.

Анализ работ этих авторов показывает, что одним из основных подходов к созданию СУЗ и его компонент является семантический подход, который основан на использовании методов и технологий по работе со смыслом, семантикой данных, информации и знаниями, таких как онтологии предметных областей, технологии их построения и сопровождения, семантические метаданные, семантический поиск, системы логического вывода, семантическое профилирование знаний экспертов, семантические порталы и сети и т.п. И все это с соответствующей технологической поддержкой в части языков описания, моделей, программных инструментов и систем.

Существуют различные методы представления накопленных знаний, в число которых входят продукционные модели, семантические сети, фреймы

и онтологии. Из перечисленных наиболее часто для описания биологических систем применяются семантические сети и онтологии.

Цель работы: разработка комплекса методов, моделей и алгоритмов для создания информационно-программной системы обеспечения поиска новых и идентификации существующих знаний в области молекулярно-биологических исследований на основе автоматизации процесса реконструкции сетей ассоциативных взаимосвязей между молекулярно-генетическими объектами из научных текстов и фактографических баз данных.

Для достижения поставленной цели исследования были поставлены и решены следующие **задачи**:

1. Выявление состава и структуры знаний с созданием онтологической модели их представления для исследований в области молекулярно-генетических взаимодействий.
2. Разработка подходов и методов извлечения знаний из текстовых источников информации для заданной предметной области.
3. Разработка средств интеграции информации, накопленной в существующих открытых фактографических базах данных.
4. Разработка архитектуры программно-информационной системы для автоматизации реконструкции сетей ассоциативных связей на основе созданной онтологической модели, реализация её в виде программно-информационного комплекса с графическим пользовательским интерфейсом.
5. Апробация технологии применения разработанной программно-информационной системы на примере решения задачи анализа особенностей ассоциативных белковых сетей человека.

Методы исследования. Для решения поставленных задач в работе используются методы системного анализа, теории графов, теории создания систем управления знаниями, объектно-ориентированного проектирования и программирования.

Научная новизна. В диссертационной работе предложен подход к решению задач по обеспечению одного из базовых видов деятельности в жиз-

ненном цикле знаний: поиска новых и идентификации существующих знаний в области молекулярно-биологических исследований, который реализован в виде проблемно-ориентированной информационно-программной системы — одной из основных подсистем системы управления знаниями в рассматриваемой области.

Получены следующие основные результаты, обладающие научной новизной:

1. Предложена онтологическая модель для описания молекулярно-генетических объектов, процессов, заболеваний и взаимоотношений между ними.
2. Разработан новый метод извлечения информации о молекулярно-генетических взаимодействиях из текстов рефератов научных статей и общедоступных фактографических баз данных, ориентированных на фармакологию, биотехнологию и биомедицину.
3. На основе предложенных онтологической модели, методов и алгоритмов разработана первая отечественная информационная система Associative Network Discovery (ИС AND), которая по полноте представления типов взаимодействий и извлечённых фактов превосходит аналогичные зарубежные разработки.
4. Разработан метод машинного обучения на основе известного алгоритма КРАБ, адаптированный для предсказания изменения термодинамической стабильности белка при одиночной аминокислотной замене.
5. С использованием созданной ИС и адаптированного метода КРАБ проведён анализ человеческого протеома на предмет влияния аминокислотных замен на термодинамическую стабильность белков.

Научная и практическая ценность. Разработанная на основе предложенных методов, моделей и алгоритмов ИС AND обеспечивает компьютерную поддержку исследований в таких областях современной науки как молекулярная биология, генетика, биотехнологии, биомедицина, фармакология, агробиология и др. Система позволяет проблемным специалистам легко ориентироваться в огромных гетерогенных хранилищах знаний в области

биологии и медицины, быстро извлекать необходимую информацию с достаточно высокой точностью и осуществлять своевременный мониторинг вновь появляющихся фактов. Она может быть полезна для студентов, аспирантов и молодых учёных для быстрого погружения в предметную область и ознакомления с новейшими открытиями, связанными с интересующими исследователя биологическими объектами. ИС AND закладывает базу для создания СУЗ в области молекулярно-биологических исследований.

Реализация и внедрение результатов работы. Система AND внедрена в Институте цитологии и генетики СО РАН (г. Новосибирск) с целью получения новых знаний, проведения прикладных исследований и опытно-конструкторских разработок в таких областях, как системная биология, структурная и функциональная геномика, транскриптомика, протеомика, метаболомика и др.

Апробация работы. Результаты работы докладывались и обсуждались на следующих конференциях:

- международная конференция «The Sixth International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2008)» (Новосибирск, Россия, 2008 г.);
- международная конференция «3-rd Moscow Conference on Computational Molecular Biology» (Москва, Россия, 2007 г.);
- международная конференция «The fourth Moscow International Congress Biotechnology: State of the Art and Prospects of Development»(Москва, Россия, 2007 г.);
- международная конференция «8th Meeting German / Russian Virtual Network on Computational Systems Biology» (Билефельд, Германия, 2007 г.);
- международная конференция «3-rd International Conference: Basic Science for Medicine» (Новосибирск, Россия, 2006 г.);
- международная конференция «The Fifth International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2006)» (Новосибирск, Россия, 2006 г.);

- российская конференция «VI Всероссийской научно — практической конференции AS'2007 (СИСТЕМЫ АВТОМАТИЗАЦИИ в образовании, науке и производстве)» (Новокузнецк, Россия, 2007 г.).

Публикации. По теме диссертационной работы опубликовано 12 печатных работ , из которых 2 работы [2,8] опубликовано в журналах из списка ВАК РФ.

Структура и объем работы. Диссертационная работа состоит из введения, четырёх глав, заключения, списка использованной литературы из 77 наименований. Основная часть работы изложена на 122 страницах машинописного текста, содержит 34 рисунка, 2 таблицы и 8 приложений.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность исследуемой проблемы, сформулирована цель и задачи диссертационной работы, перечислены полученные в диссертации новые результаты, их практическая ценность, представлены положения, выносимые на защиту и описана структура диссертации.

В первой главе представлен аналитический обзор способов представления знаний, методов извлечения знаний о молекулярных взаимодействиях из фактографических баз данных и электронных текстов научных публикаций, а также основные понятия технологии Text-mining. Рассматривается только часть задач, которые решаются в рамках технологии Text-mining. Это связано с тем, что главные цели настоящей работы связаны с извлечением знаний о молекулярных взаимодействиях из текстовых данных.

На основе анализа существующих методов извлечения информации о молекулярных взаимодействиях из фактографических баз данных и электронных текстов научных публикаций выявлено, что для экстракции информации из фактографических баз данных необходима комбинация экспертного анализа структуры и формата базы данных и автоматического извлечения этих данных с помощью программы-конвертора. Работа с каждой базой данных подразумевает индивидуальный подход в связи с существенными различиями между способами хранения фактов в базах данных, форматах представления данных, а также режимах доступа к ним.

Задача извлечения информации о взаимодействиях из текстов существенно сложнее предыдущей. При решении этой задачи существующие подходы проходят следующие этапы: создание выборки текстов, заведомо содержащих информацию о молекулярных взаимодействиях; создание словарей и онтологий, описывающих объекты предметной области, то есть белки, гены, метаболиты и другие объекты; предварительная подготовка текста с разметкой на нем слов и выражений, соответствующих названиям объектов, нормализацией текста, синтаксическим разбором предложений; обработка подготовленных текстов, с целью извлечения фактов о взаимосвязях между объектами; верификация полученных результатов.

На основе анализа эффективности существующих систем для обработки подготовленного текста был выбран метод шаблонов, обладающий с одной стороны достаточной точностью распознавания, а с другой стороны не требующий больших затрат времени на разработку и применение.

Данные, полученные из фактографических и текстовых баз данных необходимо интегрировать, с учётом дублирования, то есть наличия одинаковых взаимодействий, выявленных из различных источников.

Во второй главе описывается онтологическая модель представления знаний о взаимосвязях между молекулярно-генетическими объектами, заболеваниями и процессами.

Под онтологией в данной работе понимается набор $O =< C, R, F >$, где

$C = C_t \cup C_o \cup C_{sp} \cup C_{ti} \cup C_i \cup C_r$ — множество понятий предметной области. C_t — это множество типов объектов, C_o — множество молекулярно-генетических объектов, заболеваний, процессов и клеточных компонент. C_{sp} — множество организмов. C_{ti} — множество типов взаимосвязей между объектами. C_i — множество взаимосвязей между объектами. C_r — множество ролей объектов во взаимосвязях;

$R = \{is\_a,\ role,\ present,\ exists\_in\} \cup R_1$ — множество отношений между понятиями заданной предметной области. R_1 — множество отношений определяющих взаимосвязи различных типов между объектами. *present*

— отношение определяющее допустимые роли объектов во взаимосвязях конкретного типа. $role$ — отношение указывающее на роль объекта во взаимосвязи. $exists\_in$ — отношение связывающее молекулярно-генетические объекты с организмами, в которых они встречаются.

$F = \{f : R_1 \to C_i\}$ — множество функций интерпретации, состоящее из взаимно однозначного отображения множества отношений R_1 на множество C_i.

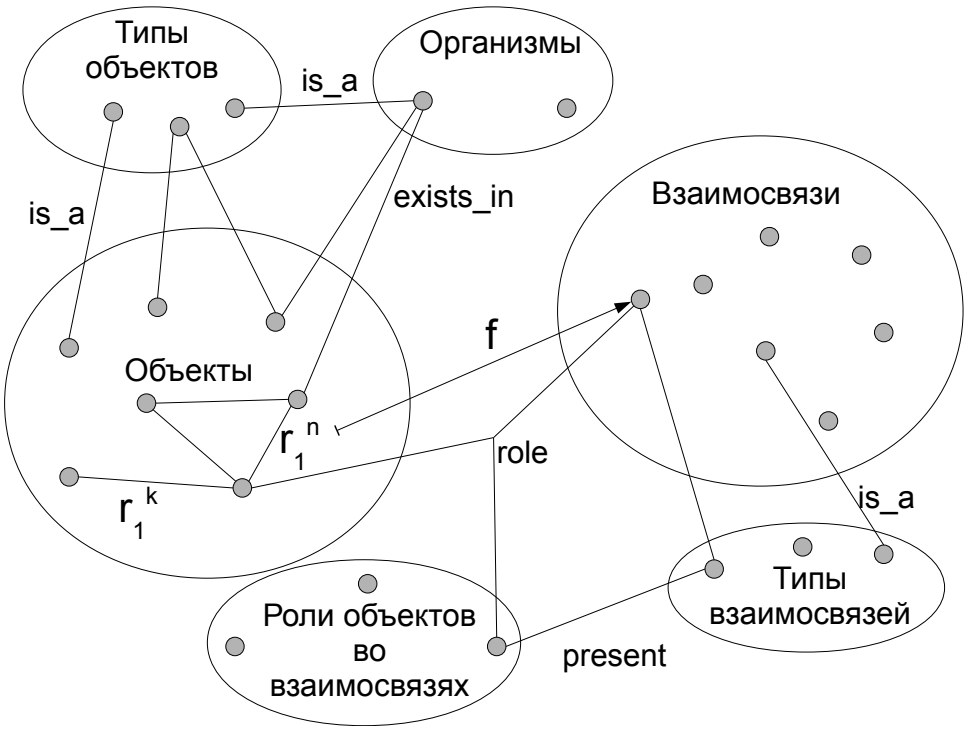

Рис. 1. Структура онтологической модели знаний о молекулярно-генетических взаимодействиях

Ассоциативной семантической сетью(ассоциативной сетью) будем называть двудольный граф, построенный по онтологии, описанной выше. Вершинами такого графа являются молекулярно-генетические объекты: (белки, гены, низкомолекулярные вещества, микроРНК), заболевания, клеточные компоненты, биопроцессы (реакции, регуляторные события), регуляторные, транспортные и метаболические пути (обозначим множество таких вершин V), а также особый тип объекта — «взаимодействие» (I). Объекты характеризуются типом, именем, списком синонимов, ссылками на базы данных,

в которых имеется информация об объекте, а также организмом, в котором встречается объект. Для объектов различного типа задаются ограничения на связи с другими объектами. Рёбрами графа являются отношения взаимодействия или ассоциации между объектами. Каждое взаимодействие и ассоциация характеризуется набором атрибутов, которые могут включать, в частности: список участвующих в нем объектов, роли участников (регулятор, объект подверженный регуляции, катализатор, субстрат, продукт и др.), тип взаимодействия и т.д.

Описываются алгоритмы составления словарей названий молекулярно-генетических объектов, химических веществ, биопроцессов и др. Методы анализа текстовых источников информации с целью составления словарей для описания знаний о молекулярно-генетических объектах и системах основываются на следующих подходах:

- анализ баз данных, в которых встречаются названия молекулярно-генетических объектов и отношений, применяемых для описания молекулярных взаимодействий;
- анализ текстов научных публикаций с целью выявления имён молекулярно-генетических объектов и отношений.

В главе представлен алгоритм для извлечения знаний о взаимосвязях между молекулярно-генетическими объектами, заболеваниями и процессами из текстов рефератов научных статей, основанный на использовании шаблонов. Также описаны алгоритмы для извлечения информации о взаимодействиях из фактографических баз данных MINT, IntAct, TRRD и GeneNet.

Функциональная схема системы извлечения знаний о молекулярных взаимодействиях в клетке представлена на рис. 2.

Исходными данными являются внешние источники данных, включающие:

- фактографические базы данных, которые используются для составления словарей;
- фактографические базы данных, которые используются для извлечения знаний о молекулярно-генетических объектах;

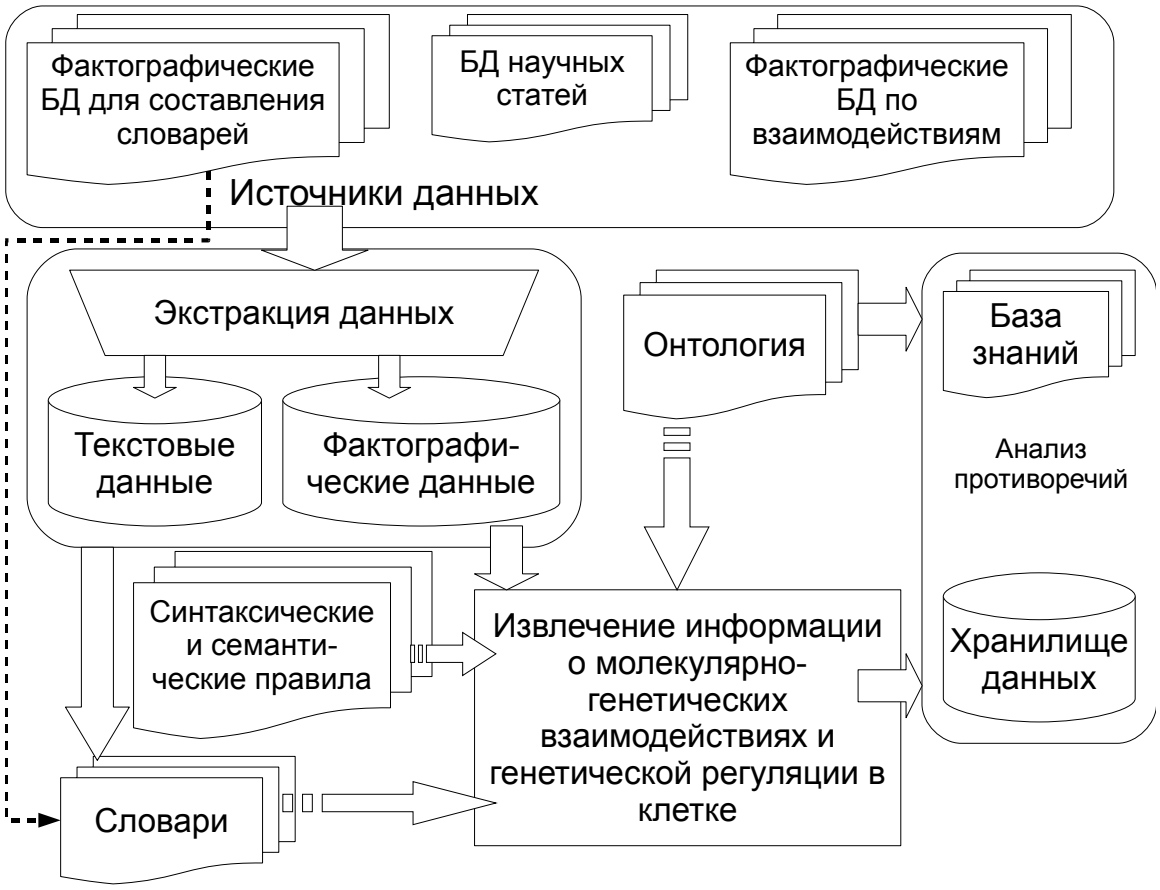

Рис. 2. Функциональная схема системы извлечения и интеграции знаний о молекулярных взаимодействиях в клетке.

- фактографические базы данных, которые используются для извлечения знаний о молекулярных взаимодействиях в клетке и генных сетях;
- базы библиографических данных (PubMed), которые используются для извлечения знаний о молекулярных взаимодействиях в клетке.

Экстрагированная информация в виде фрагментов текстовых данных накапливаются в системе для дальнейшего анализа и извлечения из неё знаний о молекулярно-генетических объектах и молекулярных взаимодействиях в клетке.

Алгоритмы извлечения знаний из текстовых данных используют словари, синтаксические и семантические правила, а также шаблоны, которые являются частью базы знаний системы Associative Network Discovery (AND).

В третьей главе представлено описание клиент-серверной архитектуры

информационной системы AND.

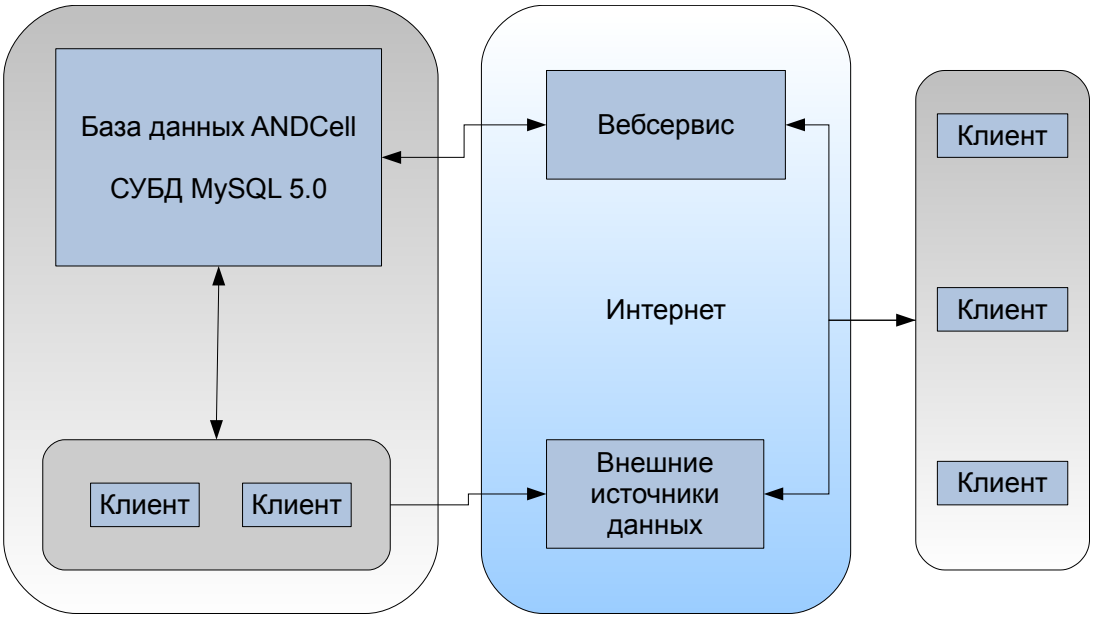

Рис. 3. Схема программного комплекса

В разработанной системе предусмотрено два типа доступа к БД: прямой доступ с компьютеров локальной сети, где расположен сервер БД системы; доступ с клиентских рабочих мест через сеть Интернет. Предложенная схема обеспечивает необходимый уровень безопасности.

Прямой доступ из локальной сети используется для администрирования системы и проведения работ по её развитию. Доступ через Интернет реализован через Web-сервис со специализированной системой авторизации и аутентификации пользователей. Одновременно, Web-сервис осуществляет контроль за количеством одновременных соединений для каждого пользователя, а также за количеством выполненных запросов для построения сетей. Эти возможности сервиса используются для ограничений демонстрационного режима использования системы.

Клиентское приложение ANDVisio для графического представления ассоциативных сетей написано на языке программирования ObjectPascal с использованием библиотеки LCL (Lazarus Component Library) для создания кроссплатформенного графического пользовательского интерфейса. Для работы с базой данных использовался набор компонент ZeosDBO, поз-

воляющий организовать унифицированный интерфейс доступа к различным СУБД. Отображение графа производится средствами библиотеки OpenGL, использование которой дало возможность работать с графами с огромным количеством вершин не испытывая при этом неудобств, связанных со скоростью визуализации.

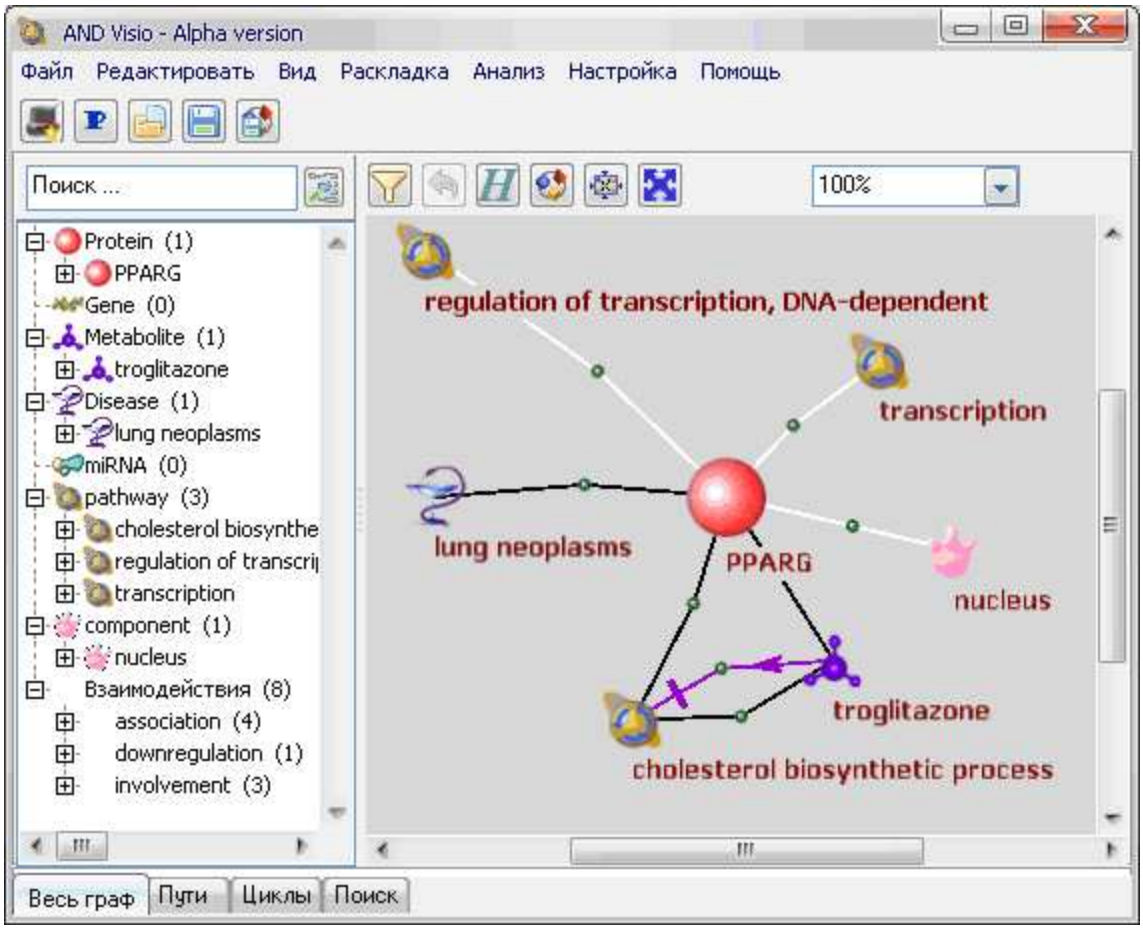

Рис. 4. Главное окно программы ANDVisio

На рисунке 4 приведено главное окно приложения. В левой части окна находится древовидное представление объектов текущего графа. Все объекты разбиты на группы по типам. В скобках приведено количество объектов каждого типа. Объекты характеризуются названием, списком синонимов и организмом. «Взаимодействия» разбиты по типам, в скобках указано количество «взаимодействий» каждого типа. Для каждого взаимодействия можно посмотреть список всех его участников. Программа позволяет строить

сети по списку синонимов интересующих объектов и/или ссылок на базы данных, с указанием для каждого объекта уровня сети. Под уровнем сети подразумевается максимальное удаление объектов графа от заданных в запросе объектов. Также возможно настроить фильтрацию сети: указать типы интересующих объектов, список организмов, типы «взаимодействий», множество баз данных, из которых извлечены факты взаимодействия. С загруженным в приложение графом можно осуществлять следующие действия: отфильтровать объекты по указанным критериям, удалить выделенные объекты, расширить граф от выделенных объектов. В программе реализованы классические алгоритмы поиска кратчайших путей в графе между двумя объектами, минимальных циклов, содержащих указанный объект, а также алгоритм построения множества фундаментальных циклов. Программа позволяет искать кратчайшие пути между объектами, минимальные циклы, в которые входит объект, строить множество фундаментальных циклов. Реализован функционал по сохранению сетей в различные форматы (XML, GenNet, бинарный файл). Возможно сохранить изображение сети в различных графических форматах. Реализована загрузка графа из файлов в бинарном и XML форматах. Также в этой главе описывается реализованный вебсервис для доступа к данным. Использование сервиса позволяет более безопасно предоставить доступ к данным через сеть Интернет. В вебсервисе реализованы средства авторизации, аутентификации и учёта активности пользователей. Один раздел в главе посвящён описанию разработанных алгоритмов раскладки графов в пространстве. Идея обоих алгоритмов состоит в построении модели, описывающей взаимосвязи между вершинами графа. Основное свойство этих систем – саморегуляция, т.е. способность системы самостоятельно находить оптимальное состояние. Первая модель использует принципы взаимодействия заряженных частиц, а во второй все вершины в графе связаны между собой пружинками с разными длинами в состоянии покоя. Обе модели описываются системами уравнений, которые решаются методом простой релаксации.

Пусть все вершины в графе пронумерованы от 1 до N, I_V — множество

индексов вершин из V, I_I — множество индексов вершин из I, а I_k — множество индексов вершин, инцидентных вершине с индексом k. Тогда система уравнений, описывающая взаимодействие вершин графа в первой модели, будет выглядит следующим образом:

$$F(X) = \begin{cases} \sum\limits_{\substack{k \neq m \\ k \in I_V}} F_q(\vec{x}_k, \vec{x}_m) + \sum\limits_{k \neq m} F_i(\vec{x}_k, \vec{x}_m) + \\ \quad + \sum\limits_{k \in I_m} F_p(\vec{x}_k, \vec{x}_m) + F_r(\vec{x}_m) = 0, & \forall m \in I_V \\ \sum\limits_{k \neq m} F_i(\vec{x}_k, \vec{x}_m) + \sum\limits_{k \in I_m} F_p(\vec{x}_k, \vec{x}_m) + \\ \quad + F_r(\vec{x}_m) = 0, & \forall m \in I_I \end{cases}$$

где силы действующие между вершинами определены так:

$$\begin{aligned} F_q(\vec{x}_1, \vec{x}_2) &= \frac{K}{\|\vec{x}_2 - \vec{x}_1\|^2} \frac{\vec{x}_2 - \vec{x}_1}{\|\vec{x}_2 - \vec{x}_1\|}, \text{где } K \geq 0 \\ F_p(\vec{x}_1, \vec{x}_2) &= P \left(\|\vec{x}_2 - \vec{x}_1\| + \frac{1}{\|\vec{x}_2 - \vec{x}_1\|} - 1 \right) \frac{\vec{x}_2 - \vec{x}_1}{\|\vec{x}_2 - \vec{x}_1\|}, \text{где } P \leq 0 \\ F_i(\vec{x}_1, \vec{x}_2) &= \frac{L}{\|\vec{x}_2 - \vec{x}_1\|} \frac{\vec{x}_2 - \vec{x}_1}{\|\vec{x}_2 - \vec{x}_1\|}, \text{где } L \geq 0 \\ F_r(\vec{x}) &= M \tan \left(\frac{\pi \|\vec{x}\|}{2R} \right) \frac{\vec{x}}{\|\vec{x}\|}, \text{где } M \leq 0. \end{aligned}$$

Система уравнений второй модели раскладки:

$$F(X) = \begin{cases} \sum\limits_{\substack{k \neq m \\ k \in I_V}} F_{ot}(\vec{x}_k, \vec{x}_m) + \sum\limits_{k \in I_m} F_{pr}(\vec{x}_k, \vec{x}_m) + \\ \quad + F_r(\vec{x}_m) = 0, & \forall m \in I_V \end{cases}$$

где силы действующие между вершинами заданы следующими выражениями:

$$\begin{aligned} F_{ot}(\vec{x}_1, \vec{x}_2) &= K(L - \|\vec{x}_2 - \vec{x}_1\|) \frac{\vec{x}_2 - \vec{x}_1}{\|\vec{x}_2 - \vec{x}_1\|}, \text{где } K, L \geq 0 \\ F_{pr}(\vec{x}_1, \vec{x}_2) &= P(D - \|\vec{x}_2 - \vec{x}_1\|) \frac{\vec{x}_2 - \vec{x}_1}{\|\vec{x}_2 - \vec{x}_1\|}, \text{где } P, D \geq 0. \end{aligned}$$

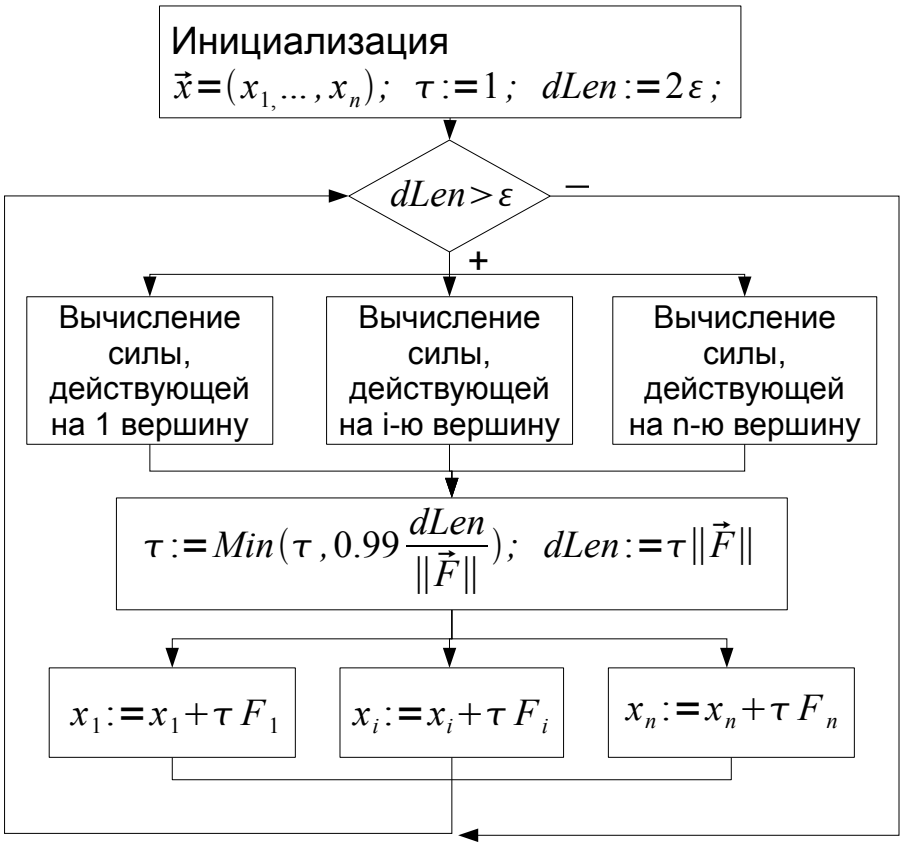

Рис. 5. Схема работы параллельного алгоритма решения системы нелинейных уравнений методом релаксаций.

Расчеты по каждому алгоритму разбиваются на N параллельных потоков, схема работы алгоритмов представлена на рис. 5.

Четвёртая глава посвящена описанию применения разработанной информационной системы AND для анализа человеческого протеома. В главе описывается метод машинного обучения на основе алгоритма КРАБ. Алгоритм адаптирован и используется для оценки изменения стабильности белков при одиночных заменах аминокислот в его последовательности. Проводится оценка чувствительности белков человека к мутациям. Результаты оценки накладываются на ассоциативную сеть взаимосвязи белков человека между собой. Кластеризация графа методом спектрального анализа показала наличие групп белков, тесно связанных между собой. Выявлено, что белки чувствительные к мутациям находящиеся в одном кластере отвечают за схожие

процессы жизнедеятельности клетки.

ОСНОВНЫЕ ВЫВОДЫ И РЕЗУЛЬТАТЫ РАБОТЫ

В ходе выполнения работы были получены следующие результаты:

1. Впервые в России разработана информационная система обеспечивающая экстракцию и интеграцию знаний о молекулярных взаимодействиях из большинства доступных гетерогенных источников информации: научных публикаций, разнородных экспериментальных данных, представленных в фактографических базах данных. Система по многим параметрам (полноте представления типов взаимодействий, количеству извлечённых фактов и др.) превосходит зарубежные аналоги.
2. Разработана онтологическая модель представления знаний о взаимосвязях между молекулярно-генетическими объектами, заболеваниями, процессами и клеточными компонентами.
3. Разработан метод и основные технологии извлечения знаний о молекулярно-генетических взаимодействиях на основе технологии Text-mining.
4. Разработана клиент-серверная архитектура программно-информационной системы AND для автоматизации процессов реконструкции сетей ассоциативных связей на основе созданной онтологической модели. Предложенная архитектура реализована в виде программно-информационного комплекса с платформо-независимым графическим пользовательским интерфейсом для представления ассоциативных сетей молекулярно-генетических взаимодействий.
5. Разработаны алгоритмы раскладки графа ассоциативных сетей на плоскости. Реализация алгоритмов позволяет использовать вычислительные возможности современных многопроцессорных систем и/или выполнять вычисления с использованием графических ускорителей.
6. Разработан метод предсказания изменения термодинамической стабильности белков при одиночных аминокислотных мутациях на основе адаптированного метода КРАБ. Применение созданной ИС и предложенного метода позволило выявить, что белки чувствительные к му-

тациям, находящиеся в одном кластере, отвечают за схожие процессы жизнедеятельности клетки.

Результаты применения созданной на основе разработанных моделей, методов и алгоритмов ИС AND показали их эффективность, в части обеспечения поддержки процессов поиска новых и идентификации существующих знаний в области молекулярно-биологических исследований. По результатам апробации ИС AND можно сделать вывод о перспективности её применения для:

- реконструкции и анализа сетевых моделей сложных молекулярно-генетических взаимодействий (генные сети), которые, как показывает опыт работы с российскими и зарубежными коллегами, востребованы в области биоинформационных, биотехнологических и биомедицинских исследований;
- проведения прикладных исследований и опытно-конструкторских разработок в таких областях, как системная биология, структурная и функциональная геномика, транскрип-томика, протеомика, метаболомика;
- решения прикладных задач в области фармакологии, биомедицины и биотехнологии для поиска новых лекарственных средств и оценки их действия на организм: реконструкция генных и метаболических сетей, описывающих на молекулярно-генетическом уровне функционирование клеток нормального и больного организма, пути проникновения патогена в клетку и взаимодействия с клеткой хозяина, пути и варианты воздействия лекарственных средств;
- построение ассоциативных семантических сетей, связывающих симптомы и наблюдения за пациентом с заболеваниями и методами лечения (в медицине).

Автор выражает искреннюю благодарность своим научным руководителям С. С. Гончарову и В. А. Иванисенко за постановку задач, всестороннюю поддержку и внимание к работе.

Список работ автора по теме диссертации

[1] *Деменков П. С., Аман Е. Э., Иванисенко В. А.* Associative network discovery (and) - компьютерная система для автоматической реконструкции ассоциативных сетей молекулярно-генетических взаимодействий // Труды VI Всероссийской научно - практической конференции AS'2007 (СИСТЕМЫ АВТОМАТИЗАЦИИ в образовании, науке и производстве). — 2007. — С. 51.

[2] *Деменков П. С., Аман Е. Э., Иванисенко В. А.* Associative network discovery (and) - компьютерная система для автоматической реконструкции сетей ассоциативных знаний о молекулярно-генетических взаимодействиях // *Вычислительные технологии.* — 2008. — Т. 13, № 2. — С. 15–19.

[3] *Деменков П. С., Яркова Е. Э., Иванисенко В. А., Колчанов Н. А., Гончаров С. С.* Предсказание изменения термодинамической стабильности белков при одиночных аминокислотных заменах // Системная компьютерная биология / Под ред. Н. А. Колчанова, С. С. Гончарова. — Новосибирск: Издательство СО РАН, 2008. — С. 269–275.

[4] *Aman E. E., Demenkov P. S., Ivanisenko V. A.* Analysis of the tertiary structure of the ppar and rxr transcriptional factors and their mutant variants // Proceedings of the BGRS-2006. — Vol. 1. — 2006. — Pp. 227–230.

[4] *Аман Е. Э., Деменков П. С., Иванисенко В. А.* Анализ третичной структуры транскрипционных факторов ppar и rxr и их мутантных вариантов // Труды БГРС-2006. — Т. 1. — 2006. — С. 227–230.

[5] *Aman E. E., Demenkov P. S., Ivanisenko V. A.* Textomics: the instrument for biological knowledge discovery // The fourth Moscow International Congress Biotechnology: State of the Art and Prospects of Development. — Vol. 2. — 2007. — P. 391.

[5] *Аман Е. Э., Деменков П. С., Иванисенко В. А.* Текстомика: инструмент поиска биологических знаний // Пятый московский международный биотехнологический конгресс. — Т. 2. — 2007. — С. 391.

[6] *Ivanisenko V. A., Demenkov P. S., Aman E. E., Pintus S. S., Kolchanov N. A.* Associative network and protein structure discovery: a soft-

ware complex for facilitating search of targets for drugs, drug design, and evaluation of molecular toxicity // 3rd International conference "Basic science for medicine". — 2007. — P. 92.

[6] *Иванисенко В. А., Деменков П. С., Аман Е. Э., Пинтус С. С., Колчанов Н. А.* Associative network and protein structure discovery: программный комплекс для облегчения поиска лекарственных целей, создания лекарств и оценки токсичности молекул // 3-й международная конференция "Основы науки для медицины". — 2007. — С. 92.

[7] *Aman E. E., Demenkov P. S., Nemiatov A. I., Ivanisenko V. A.* Associative network discovery (and) - software package for automated reconstruction of molecular-genetic association networks // Proceedings of the 3-rd Moscow Conference on Computational Molecular Biology. — 2007. — Pp. 33–34.

[7] *Аман Е. Э., Деменков П. С., Немятов А. И., Иванисенко В. А.* Associative network discovery (and) - программный пакет для автоматической реконструкции молекулярно-генетических ассоциативных сетей // Труды 3-й московской конференции по вычислительной молекулярной биологии. — 2007. — С. 33–34.

[8] *Demenkov P. S., Aman E. E., Ivanisenko V. A.* Prediction of the changes in thermodynamic stability of proteins caused by single amino acid substitutions // *Biophysics*. — 2006. — Vol. 51, no. Suppl. 1. — P. 49.

[8] *Деменков П. С., Аман Е. Э., Иванисенко В. А.* Предсказание изменения термодинамической стабильности белков при одиночной аминокислотной замене // *Биофизика*. — 2006. — Т. 51. — С. 49.

[9] *Demenkov P. S., Ivanisenko V. A.* Prediction in changes of protein thermodynamic stability upon single mutations // Proceedings of the BGRS-2006. — Vol. 1. — 2006. — Pp. 256–259.

[9] *Деменков П. С., Иванисенко В. А.* Предсказание изменения термодинамической стабильности белка при одиночных мутациях // Труды БГРС-2006. — Т. 1. — 2006. — С. 256–259.

[10] *Aman E. E., Demenkov P. S., Pintus S. S., Nemiatov A. I., Apasieva N. V., Korotkov R. O., Ignatieva E. V., Podkolodny N. L., Ivanisenko V. A.* Devel-

opment of a computer system for the automated reconstruction of molecular-genetic interaction networks // Proceedings BGRS-2006. — Vol. 3. — 2006. — Pp. 15–18.

[10] *Аман Е. Э., Деменков П. С., Пинтус С. С., Немятов А. И., Апасьева Н. В., Коротков Р. О., Игнатьева Е. В., Подколодный Н. Л., Иванисенко В. А.* Разработка компьютерной системы для автоматической реконструкции сетей молекулярно-генетических взаимодействий // Труды БГРС-2006. — Т. 3. — 2006. — С. 15–18.

[11] *Ivanisenko V. A., Pintus S. S., Demenkov P. S., Krestyanova M. A., Litvenko E. K., Grigorovich D. A., Debelov V. A.* Fastprot: a computational workbench for recognition of the structural and functional determinants in protein tertiary structures // Bioinformatics of Genome Regulation and Structure II / Ed. by N.Kolchanov, R. Hofestaedt. — Springer Science+Business Media, Inc, 2006. — Pp. 305–316.

[11] *Иванисенко В. А., Пинтус С. С., Деменков П. С., Крестьянова М. А., Литвенко Е. К., Григорович Д. А., Дебелов В. А.* Fastprot: автоматизированное рабочее место для распознавания структурных и функциональных детерминант в третичной структуре белка // Биоинформатика регуляции генома и структуры II / под. ред. Н. Колчанова, Р. Хофештадта. — Springer Science+Business Media, Inc, 2006. — С. 305–316.

[12] *Ivanisenko V. A., Demenkov P. S., Aman E. E., Pintus S. S., Fomin E. S.* Structure discovery - computer tools for protein analysis and search of drug target // The fourth Moscow International Congress Biotechnology: State of the Art and Prospects of Development. — Vol. 2. — 2007. — P. 395.

[12] *Иванисенко В. А., Деменков П. С., Аман Е. Э., Пинтус С. С., Фомин Э. С.* Structure discovery - компьютерные утилиты для анализа белков и поиска лекарственных целей // Пятый московский международный биотехнологический конгресс. — Т. 2. — 2007. — С. 395.

Деменков Павел Сергеевич

Математическое и программное обеспечение для реконструкции ассоциативных сетей молекулярно-генетических взаимодействий

Автореферат
диссертации на соискание учёной степени
кандидата технических наук

Подписано в печать 15.10.08. Формат 60x84 1/16.
Усл. печ. л. 1,5. Уч.-изд. л. 1,5. Тираж 100 экз. Заказ №187.

Отпечатано в ООО "Омега Принт"
630090, Новосибирск, пр. Лаврентьева, 6