На правах рукописи

МЕДВЕДЕВА ИРИНА ВАДИМОВНА

КОМПЬЮТЕРНЫЙ АНАЛИЗ ЗАКОНОМЕРНОСТЕЙ КОДИРОВАНИЯ ФУНКЦИОНАЛЬНЫХ САЙТОВ БЕЛКОВ В ГЕНАХ ПОЗВОНОЧНЫХ

03.01.09 Математическая биология, биоинформатика

Автореферат
диссертации на соискание ученой степени
кандидата биологических наук

Новосибирск
2014

Работа выполнена в лаборатории компьютерной протеомики Федерального государственного бюджетного учреждения науки Институт цитологии и генетики Сибирского отделения Российской академии наук, г. Новосибирск, Россия.

Научный руководитель: кандидат биологических наук, доцент
Иванисенко Владимир Александрович

Официальные оппоненты: **Омельянчук Леонид Владимирович**,
доктор биологических наук, заведующий лабораторией генетики клеточного цикла, Федеральное государственное бюджетное учреждение науки Институт молекулярной и клеточной биологии Сибирское отделение Российской академии наук, г. Новосибирск

Москалев Алексей Александрович,
Доктор биологических наук, доцент, заведующий лабораторией молекулярной радиобиологии и геронтологии Института биологии Коми НЦ УрО РАН, г. Сыктывкар

Ведущее учреждение: ФГУН Государственный научный центр вирусологии и биотехнологии «Вектор», п. Кольцово, Новосибирская область

Защита диссертации состоится «___» _____ 2014 г. на утреннем заседании диссертационного совета Д 003.011.01 по защите диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук в Институте Цитологии и Генетики СО РАН в конференц-зале Института по адресу: 630090, г. Новосибирск, проспект ак. Лаврентьева, 10.

Тел.: (383)363-49-06, факс: (383) 333-12-78, e-mail:dissov@bionet.nsc.ru.

С диссертацией можно ознакомиться в библиотеке ИЦиГ СО РАН и на сайте института www.bionet.nsc.ru.

Автореферат разослан «___» _____ 2014 г.

Ученый секретарь
диссертационного совета,
доктор биологических наук Хлебодарова Т. М.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность проблемы. Исследование механизмов, лежащих в основе эволюции структуры и функции белка, является одним из важнейших разделов современной биологии. В ходе дискуссии в 1978 году Уильям Гилберт выдвинул предположение, согласно которому один экзон кодирует один домен (Gilbert, 1978). Однако дальнейшие исследования показали, что корреляция между границами доменной и экзонной структур наблюдается не всегда (Kaessmann et al., 2002). Непосредственно в функциональных взаимодействиях белка или его домена задействовано небольшое количество аминокислотных остатков, образующих функциональный сайт. Функция и структурная организация функциональных сайтов напрямую связаны с молекулярной эволюцией соответствующих генов и белков, однако, эта взаимосвязь оставалась практически не изученной.

Исследование закономерностей и анализ структурно-функциональной организации генов с учетом информации о расположении границ экзонов, доменов и функциональных сайтов белков как на уровне аминокислотных последовательностей, так и нуклеотидных последовательностей ДНК невозможны без применения биоинформатических методов. До недавнего времени возможности применения этих методов были ограничены небольшим числом полностью секвенированных геномов секвенированных геномов и расшифрованных третичных структур белков. В настоящее время накоплены огромные массивы молекулярно-генетических данных, представленных в базах последовательностей генов (GeneBank, EMBL, Ensembl и др.), белковых последовательностей (SwissProt, Trembl и др.), пространственных структур белков (PDB) и их функциональных сайтов (PDBSite, SitesBase). Интеграция этих ресурсов позволяет получить новые знания о структурно-функциональной организации экзонов, доменов, функциональных сайтов, участков с повышенной консервативностью и других генетических кодах, представленных в геномных последовательностях и их роли в эволюции молекулярно-генетических систем живых организмов.

Цель настоящего исследования. Цель работы состояла в выявлении закономерностей кодирования функциональных сайтов белков с использованием проекций границ экзонов на первичные и пространственные структуры белков. В связи с этим решались следующие задачи:

1. Разработка компьютерной системы, предназначенной для анализа проекций на аминокислотную последовательность белков экзонной структуры кодирующих их генов, границ доменов и позиций функциональных сайтов. Создание базы данных, интегрирующей результаты проекции и существующие ресурсы по структурно-функциональной организации белков и генов.
2. Интеграция компьютерной системы с программой BLAST с целью поиска гомологичных экзонов и участков полипептидов, кодируемых одним экзоном, и программой 3DPDBScan для осуществления структурного выравнивания анализируемого белка с пространственными структурами фрагментов белков, кодируемых одним экзоном.
3. Анализ закономерностей распределения фрагментов ДНК, кодирующих функциональные сайты белков, в экзонной структуре гена

4. Исследование распределения кодонов в фрагментах ДНК, кодирующих функциональные сайты белков, на границах экзонов.

Научная новизна. Впервые установлено, что функциональные сайты белков преимущественно кодируются более длинными экзонами. При этом оказалось, что в случае разрывных функциональных сайтов, кодирующие их фрагменты ДНК преимущественно распределяются в пределах одного или нескольких сближенных в последовательности гена экзонов. Впервые выявлены статистически значимые отличия между частотами фаз кодонов, расположенных на 5`-конце экзонов, кодирующих и не кодирующих функциональные сайты белков. Согласно этим данным нулевая фаза кодонов встречается реже в случаях экзонов, кодирующих функциональные сайты. Впервые выдвинута гипотеза о том, что экзоны, кодирующие только фрагменты функциональных сайтов белков, меньше подвержены перетасовкам по сравнению с другими экзонами. Таким образом, возникновение функциональных сайтов в аминокислотных последовательностях белков может быть фактором, ограничивающим изменчивость экзонной структуры генов, в том числе в результате перетасовок экзонов.

Впервые создана программно-информационная система, интегрирующая различные структурные и функциональные данные о белках и кодирующих их генах, включая белковые и геномные последовательности, экзон-интронную структуру, домены и функциональные сайты. Система включает в себя базу данных SitEx, содержащую данные о функциональных сайтах белков, нуклеотидных и аминокислотных последовательностях экзонов и соответствующих им фрагментов пространственных структур полипептидной цепи белка, а также программы анализа. Новизной обладают предоставляемые в системе возможности поиска по базе данных ДНК последовательностей экзонов с помощью программы BLASTN, а также поиска по базе данных фрагментов белков, кодируемых отдельно взятыми экзонами, с помощью BLASTP и программы 3DPDBScan, осуществляющей структурное выравнивание 3D структур этих фрагментов.

Практическая ценность. Разработанная компьютерная система SitEx имеет свободный доступ через Интернет и может использоваться для решения широкого круга фундаментальных и прикладных задач, связанных с анализом соотношения экзон-интронной структуры генов и структурно-функциональной организации кодируемых ими белков. SitEx позволяет проводить поиск гомологий между белковыми последовательностями, а также осуществлять структурное сравнение белков с учетом информации об экзон-интронной структуре кодирующих их генов. Функциональные возможности созданной системы SitEx могут быть использованы при планировании генно-инженерных экспериментов.

Положения, выносимые на защиту.

- Функциональные сайты белков значимо чаще, чем ожидается по случайным причинам, кодируются одним или близко расположенными в последовательности гена экзонами;

- Длина экзонов, кодирующих участок белка, содержащий аминокислотные остатки функциональных сайтов, в среднем значимо превышает длину остальных экзонов;

- Частота представленности фазы 0 кодонов, располагающихся на 5’-конце экзонов в районах ДНК, кодирующих функциональные сайты, значимо меньше частоты представленности фазы 0 кодонов на 5’-конце экзонов в районах ДНК, не кодирующих функциональные сайты.

- Кодоны, содержащие аденозин и тимин в третьей позиции, используются чаще во фрагментах ДНК длиной до 15 нуклеотидов на 5`-конце экзонов, кодирующих функциональные сайты белков человека.

Апробация. Результаты работы были представлены на международных и российских конференциях: на международной конференции 19th Annual International Conference on Intelligent Systems for Molecular Biology and 10th European Conference on Computational Biology (Австрия, Вена, 2011); на Международной конференции по биоинформатике регуляции и структуры генома (BGRS) в 2008, 2010, 2012 гг. (Россия, Новосибирск); на летней школе 2011 International German/Russian Summer School on Integrative Biological Pathway Analysis and Simulation (Германия, Билефельд, 2011); на Школе Молодых Ученых в 2008 и 2010 гг. (Россия, Новосибирск); на международной конференции The 2007 International Conference on Bioinformatics & Computational Biology (2007, США, Лас-Вегас); на Международной научной конференции студентов, аспирантов и молодых учёных "Ломоносов-2007" (Москва, Россия) и на XLIV Международной научной студенческой конференции "Студент и научно-технический прогресс" (Новосибирск, Россия).

Публикации. В результате выполнения работы было опубликовано 3 статьи в рецензируемых журналах, рекомендованных ВАК, 6 тезисов к российским и международным конференциям, получено одно свидетельство о государственной регистрации базы данных.

Личный вклад автора. Основные результаты работы были получены и проанализированы автором самостоятельно, а именно: (1) разработана структура и интерфейс базы данных SitEx; (2) разработаны алгоритмы и программы, с использованием которых проведен анализ геномных данных и данных по функциональным сайтам белков и заполнение на этой основе базы данных SitEx; (3) осуществлена интеграция доступных внешних программ BLAST и PDB3DScan в систему SitEx; (4) проведен анализ данных из базы данных SitEx по установлению закономерностей кодирования функциональных сайтов белков в геномах позвоночных. Реализация веб-версии компьютерной системы была осуществлена совместно с Деменковым П. С.

Структура и объем работы. Работа состоит из оглавления, списка сокращений, введения, трех глав, заключения, выводов, списка литературы и пяти приложений. Материал изложен на 108 страницах (101 страница текста и 7 страниц приложений), содержит 28 рисунков, 11 таблиц, 2 формулы.

Благодарности. Автор выражает искреннюю благодарность руководителю диссертации к.б.н. Иванисенко В.А., соавторам и коллегам по работе – академику РАН Колчанову Н.А., к.б.н. Деменкову П.С., к.б.н. Орлову Ю.Л., д.б.н. Кочетову А.В. за консультации и плодотворные научные дискуссии. Автор особо благодарен к.б.н. Рогозину И.Б. за большой объём консультаций по биологическим вопросам и за помощь в биологической интерпретации результатов.

Список сокращений: БД – база данных, ФС – функциональный сайт белка; ЭКФС – экзоны, содержащие кодоны, кодирующие аминокислоты функционального сайта; ЭНФС – экзоны, не содержащие кодоны, кодирующие аминокислоты функционального сайта.

СОДЕРЖАНИЕ РАБОТЫ

Глава 1. Обзор литературы. Рассмотрены основные принципы структурно-функциональной организации белков с анализом существующих данных о классификации функциональных сайтов, в том числе по физико-химическим свойствам и типу связываемых ими лигандов. Проведен анализ существующих теорий по молекулярной эволюции генов с подробным обсуждением гипотезы о корреляции границ экзонной структуры гена и доменной структуры белка. Сделан обзор существующих баз данных и веб-ресурсов, содержащих информацию о структурной организации и функциональных свойствах белков, включая базы данных по функциональным сайтам, а также существующим ресурсам, позволяющим анализировать проекцию экзонной структуры гена на первичную и пространственную структуру белка. В заключение показаны недостатки существующих подходов к изучению закономерностей кодирования функциональных сайтов белков и актуальность проведения исследований в этой области. Сделан вывод об отсутствии компьютерно-информационных систем, интегрирующих распределенные данные по структурно-функциональной организации белков и генов, позволяющих проводить анализ взаимной проекции их структур.

Глава 2. Компьютерная система SitEx. Разработана компьютерная система SitEx, предназначенная для анализа соотношения между экзон-интронной структурой генов и особенностями структурно-функциональной организации белков, включая структуру и свойства их доменов и функциональных сайтов.

Система состоит из трех интегрированных между собой компонентов:

1) базы данных, содержащей информацию о проекциях на аминокислотную последовательность белков экзонной структуры кодирующих их генов, границ функциональных и структурных доменов белков, а также позиций функциональных сайтов белков;
2) программных средств BLAST и 3DPDBScan, предназначенных для поиска по базе данных SitEx на основе анализа сходства нуклеотидных последовательностей генов, а также первичных и третичных структур белков;
3) веб-интерфейса, обеспечивающего доступ к базе данных и программным средствам, а также предоставляющего графическую визуализацию результатов.

При создании базы данных SitEx использовались данные из таких ресурсов как Ensembl (хранение полной информации о последовательности гена), Protein Data Bank (БД PDB, содержащая информацию о пространственной структуре белков), SCOP (структурная классификация белков) (рис. 1). В разделе приводится описание форматов данных этих ресурсов.

На первом шаге создания базы данных SitEx из БД PDB отбирались записи, содержащие координаты атомов пространственных структур полипептидов, имеющих менее 90% сходства между собой по аминокислотной последовательности, при этом находящихся в комплексе с различными лигандами. Кроме того проводилась фильтрация по организмам, рассматривались только позвоночные. Таким образом, из БД PDB (версия 55) было отобрано около 12 000 записей. На втором шаге, устанавливалось соответствие между отобранными записями БД PDB и базой данных Ensembl. Критериями соответствия записей БД PDB и БД Ensembl являлись указание идентификатора соответствующей записи БД PDB в записи БД Ensembl, а также сходство аминокислотных последовательностей (не менее 90% идентичности по аминокислотной последовательности), приведенных в данных записях, рассчитываемое с помощью глобального парного выравнивания с применением программы CLUSTALW. На этом шаге была отобрана 2021 уникальная запись.

Из записи PDB извлекалась следующая информация. Описание белков и лигандов извлекалось из полей HEADER, TITLE, COMPND, SOURCE, KEYWDS, HETNAM. Описание сайтов и информация об их позициях в аминокислотной последовательности извлекалось из полей REMARK 800 и SITE. Из поля ATOM извлекались координаты атомов полипептидов, которые использовались при поиске по базе данных SitEx с помощью структурного выравнивания, осуществляемого программой 3DPDBScan.

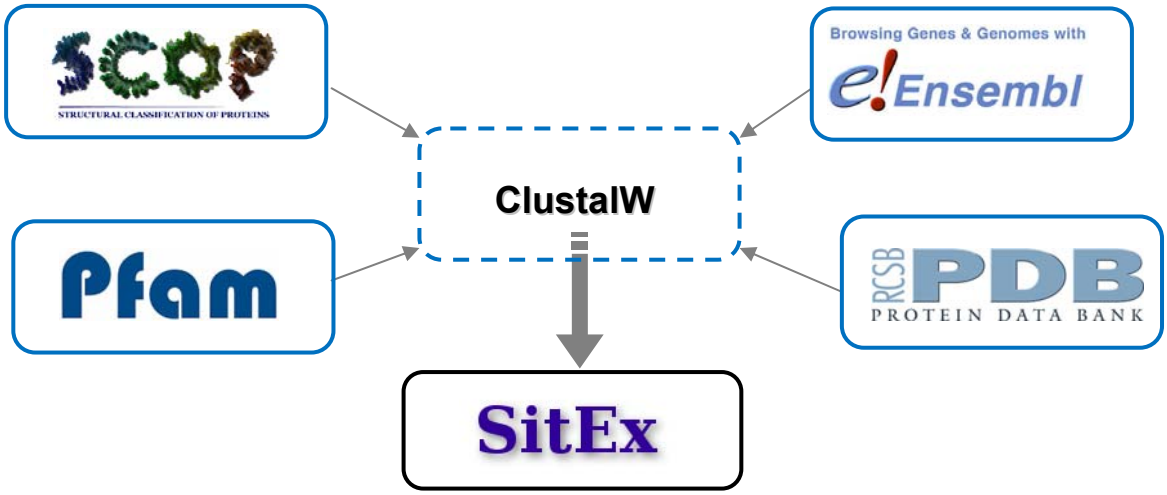

Рисунок 1. Схема интегрирования данных на основе компьютерных программ и баз данных, необходимых для разработки SitEx.

Из Ensembl для каждого белка извлекалось его наименование, кодирующая нуклеотидная последовательность, полная аминокислотная последовательность, а также информация о расположении границ экзонов в нуклеотидной последовательности и границ доменов Pfam в аминокислотной последовательности. Дополнительно, по заданному идентификатору записи PDB из базы данных SCOP извлекалась информация о границах структурных доменов белков. Работа с PDB велась на основе файлов в формате .pdb. Доступ к информации базы данных Ensembl осуществлялся с использованием функционала,

предоставляемого веб-интерфейсом, а также через открытый MySQL сервер (ensembldb.ensembl.org). Все программы для интеграции данных написаны на языке Perl и языка запросов MySQL.

Показатели разрывности функциональных сайтов белков

Все сайты, представленные в базе данных SitEx, характеризовались показателями разрывности сайта в последовательности белка и в экзонной структуре кодирующего гена. Коэффициент разрывности белковых функциональных сайтов в экзонной структуре кодирующих генов CoefE, вычислялся по формуле:

$$CoefE = 1 - \frac{N}{p_N^E - p_1^E + 1},$$

где p_1^E - порядковый номер первого экзона в последовательности гена, кодирующего функциональный сайт, p_N^E - порядковый номер последнего экзона в последовательности гена, кодирующего функциональный сайт, N – число экзонов последовательности гена, кодирующих функциональный сайт. Для расчета коэффициента разрывности функциональных сайтов в аминокислотных последовательностях CoefA применялась аналогичная формула:

$$CoefA = 1 - \frac{M}{p_M^A - p_1^A + 1},$$

где M – количество аминокислот функционального сайта в аминокислотной последовательности, p_M^A - позиция последнего аминокислотного остатка сайта, p_1^A - позиция первого остатка.

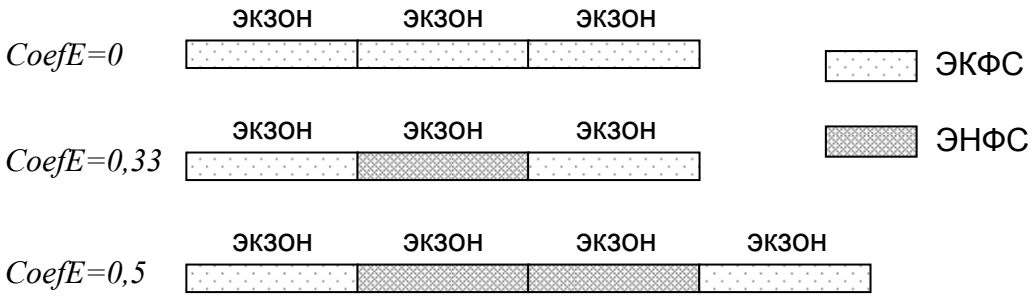

Рисунок 2. Пример значений коэффициента разрывности сайта по экзонам.

Как можно видеть из формул значения CoefE и CoefA лежат в интервале [0, 1). При этом их значения равны 0, если в пределах границ сайта, отмеченных на аминокислотной последовательности белка, в случае CoefA, или на экзонной структуре, в случае CoefE, располагаются только аминокислоты функционального сайта или экзоны, участвующие в кодировании функционального сайта, соответственно (рис. 2). В противном случае, значения этих коэффициентов стремятся к единице в зависимости от количества вставок в заданные границы функционального сайта аминокислот или экзонов, не связанных с данным сайтом.

Описание структуры базы данных SitEx

БД SitEx является реляционной базой данных, для создания которой использовалась система управления MySQL. Структура БД представлена на рис. 3.

Вся описательная информация о функциональных сайтах хранится в таблицах PDB_Site. Информация об идентификаторах Ensembl, кодирующих нуклеотидных и аминокислотных последовательностях, хранится в таблицах ENS_Chain. Описание белка, полученное из БД PDB, внесено в таблицу PDB_Chain. Последние две таблицы связаны между собой вспомогательной таблицей ENS_PDB_Chain.

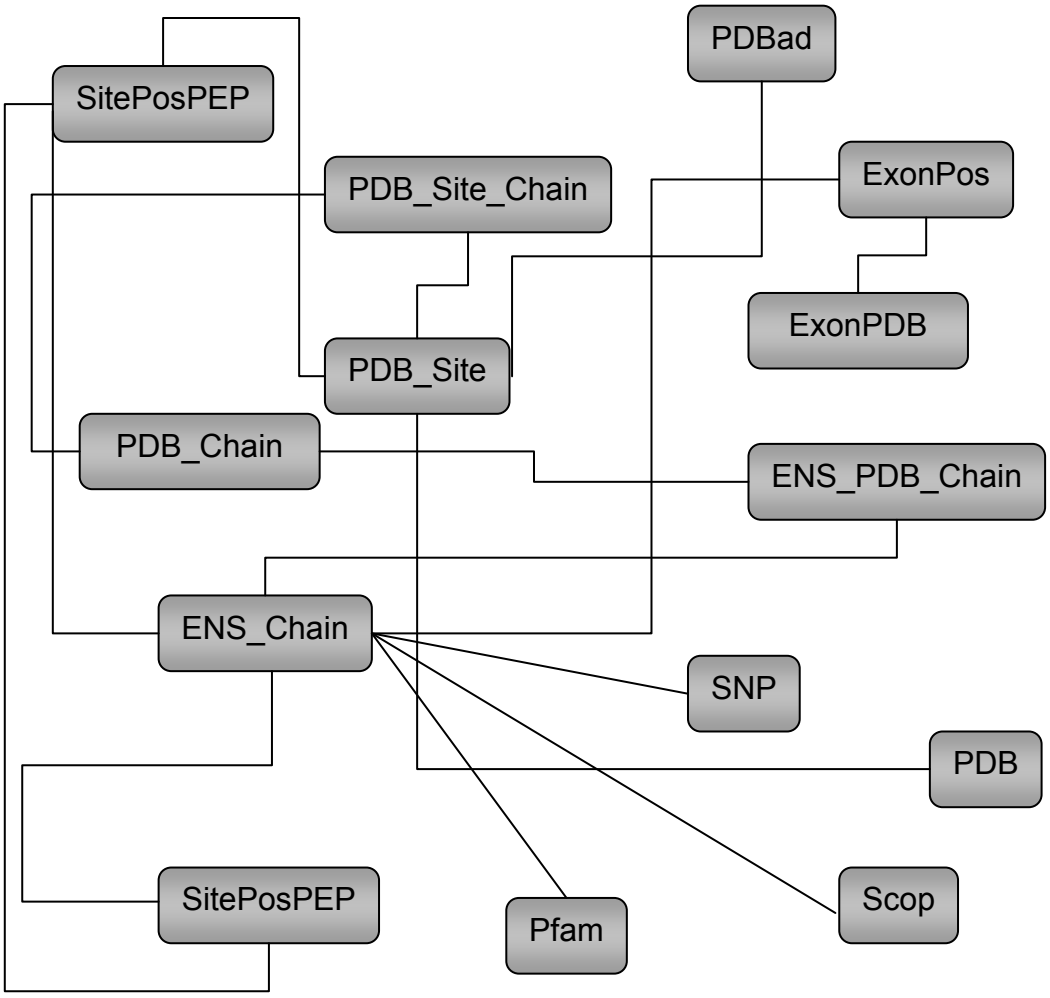

Рисунок 3. Схема структуры базы данных SitEx. Заголовки соответствуют названиям таблиц.

Поскольку функциональный сайт может быть распределен по нескольким белковым цепям, представленным в структуре PDB, то для описания связи между сайтом и цепями белка в БД создана таблица PDB_Site_Chain. Позиции аминокислот функционального сайта в последовательности белка из Ensembl хранятся в таблице SitePosPEP, а соответствующие им позиции в кодирующей нуклеотидной последовательности, хранятся в таблице SitePosCDS. Таблицы SitePosPEP, SitePosCDS, PDB_Site и ENS_Chain связаны между собой через соответствующие идентификаторы.

Таблицы Pfam и Scop содержат информацию о позициях границ доменов в аминокислотной последовательности, содержащейся в поле SeqPEP таблицы

ENS_Chain. Таблица PDBad информирует о тех идентификаторах PDB, которые описывают белки, имеющие последовательность, сходную с представленной в БД SitEx, более чем на 90%. Таблица ExonPDB содержит файлы PDB, сгенерированные для каждого экзона, если для его последовательности известна пространственная структура. Остальные таблицы содержат различную дополнительную информацию о белках и кодирующих их генов, включая полиморфизмы генов.

Описание веб-интерфейса

Разработанный веб-интерфейс обеспечивает доступ пользователей к базе данных SitEx и визуализацию результатов поиска. Реализована возможность проведения гибкого поиска по текстовым полям БД SitEx и поиска по сходству нуклеотидных или аминокислотных последовательностей, а также пространственных структур белков, выполняемого с помощью программ BLAST и 3DPDBScan, соответственно. Для удобства осуществления запросов и представления результатов поиска веб-интерфейс реализован в виде отдельных страниц, соответствующих определенным функциональным возможностям системы или типам данных.

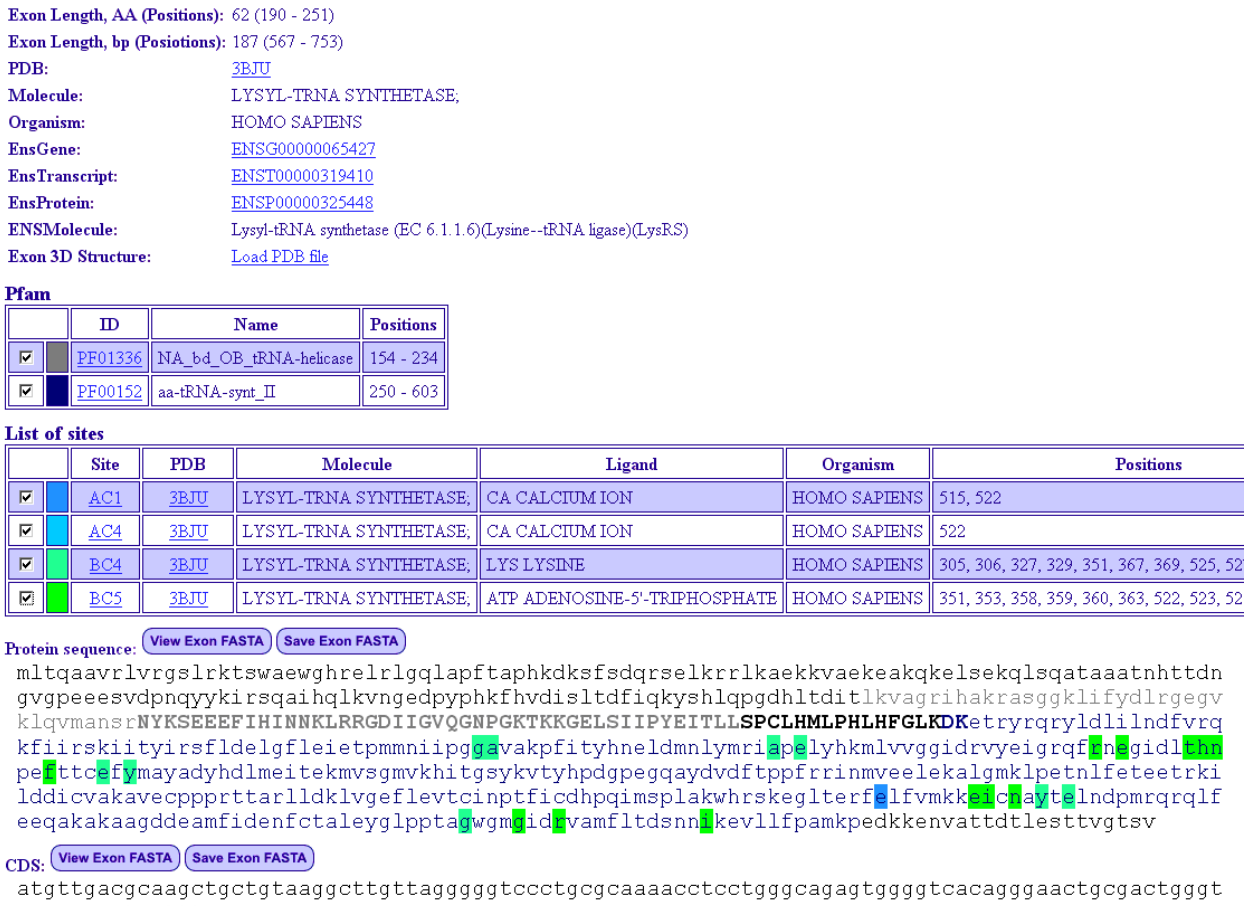

Рисунок 4. Страница описания экзона в базе данных SitEx. Представлен блок описания последовательности, блок для разметки доменов на последовательности, а также блок для разметки аминокислот функциональных сайтов на последовательности экзона либо полипептиде, кодируемом им.

Страница описания экзона (рис. 4) предназначена для представления результатов поиска по БД SitEx, содержащих информацию об экзонах, включая длину экзона, положение его границ в кодирующей нуклеотидной последовательности гена, аминокислотной последовательности белка, а также различную описательную информацию о белке и соответствующем гене, согласно БД PDB и Ensembl. На странице в графическом виде показана разметка на аминокислотных и нуклеотидных последовательностях границ экзонов, белковых доменов, а также позиций функциональных сайтов.

Страница функционального сайта предназначена для представления результатов поиска по БД SitEx, содержащих информацию о функциональных сайтах, включая данные о белках, доменах, разметке позиций функциональных сайтов в аминокислотных и кодирующих нуклеотидных последовательностях, коэффициентах разрывности сайтов и т.д. Для удобства представления, также как и на «странице описания экзона» используется графическое изображение аминокислотных и нуклеотидных последовательностей с выделением позиций функциональных сайтов, границ белковых доменов и экзонов.

Страница статистики предоставляет текущую информацию о содержании БД. Ниже, согласно данной странице, приведена статистика по настоящей версии БД:

- 14 организмов (75% сайтов представлено белками человека, 10% белками мыши, 5% белками крысы, 5% белками быка, а остальные белки представлены единично);
- 715 лигандов;
- 2021 (из 4014) уникальных последовательностей[1];
- 9994 (из 10887) уникальных сайта[2];
- классификация белков (Таблица 1);
- классификация сайтов (Таблица 2).

Таблица 1. Классификация белков в SitEx

Наименование	Кол-во записей
Белки мышц	225
Белки крови	25
Белки клеточного цикла	328
Ферменты (с EC-номером, киназы, синтазы)	2069
Белки иммунной системы	274
Мембранные белки	73
Рецепторы	213
Белки репликации, транскрипции, трансляции	73
Белки теплового шока	22
Транспортные белки	313
Белки опухолей	161
«Цинковые пальцы», RING пальцы	177
Другие белки и предшественники	614

[1] Уникальная последовательность – та, которая не повторяется в выборке.

[2] Уникальный сайт – сайт, который не повторяется в одной и той же структуре по аминокислотному составу

Классификация белков проводилась по ключевым словам, включающим ткань, функцию, локализацию и процесс в названиях белков, извлеченных из базы данных Ensembl. Всего было выделено 13 групп белков, среди которых максимально представленными оказались ферменты (Таблица 1).
Лиганд-связывающие сайты также классифицировались по ключевым словам, коду функциональных сайтов в БД PDB и номенклатурным окончаниям. Все лиганды были разбиты на 14 групп по типу лиганда, среди которых наиболее представленными оказались неорганические лиганды (Таблица 2).

Таблица 2. Классификация лигандов в SitEx

Наименование	Кол-во записей
Ионы металлов	2917
Анионы кислот	2401
Органические кислоты	595
Нуклеотидфосфаты	799
Фосфосахара	308
Белки	73
Аминокислоты и их соединения	164
Коферменты	89
Спирты и их производные	665
Атомы и неорганические соединения	351
Амины и амиды	1112
Порфирины	59
Более мелкие классы (алкалоиды, кетоны, пигменты и прочее)	958
Неизвестный лиганд	396

Страница Exon BLAST Search предназначена для осуществления поиска по БД по сходству нуклеотидных или аминокислотных последовательностей в формате FASTA с использованием программы BLAST. Для осуществления такого поиска была проведена индексация аминокислотных и нуклеотидных последовательностей из БД SitEx с использованием инструментов программы BLAST.

Страница 3D Exon Search предоставляет интерфейс для загрузки файла в формате PDB с указанием полипептидной цепи анализируемого белка. Для осуществления поиска по базе SitEx, основанного на сходстве пространственных структур анализируемого белка и фрагментов белков, кодируемых отдельными экзонами, вызывается программа 3DPDBScan. Результатом поиска является интерактивная таблица с идентификаторами экзонов БД SitEx, содержащая стандартные показатели качества структурного выравнивания. Предоставляется возможность перехода на другие страницы интерфейса для получения детальной информации об экзонах и функциональных сайтах, а также возможность графической визуализации совмещенных в результате выравнивания пространственных структур с последующим сохранением в формате PDB.

Глава 3. Статистический анализ особенностей кодирования функциональных сайтов белков в генах эукариот.

Исследование распределений длин экзонов, кодирующих и некодирующих функциональные сайты

Для сравнительного анализа распределений длин экзонов, кодирующих и некодирующих функциональные сайты, было создано две соответствующих выборки экзонов (ЭКФС и ЭНФС) с использованием БД SitEx (рис. 5).

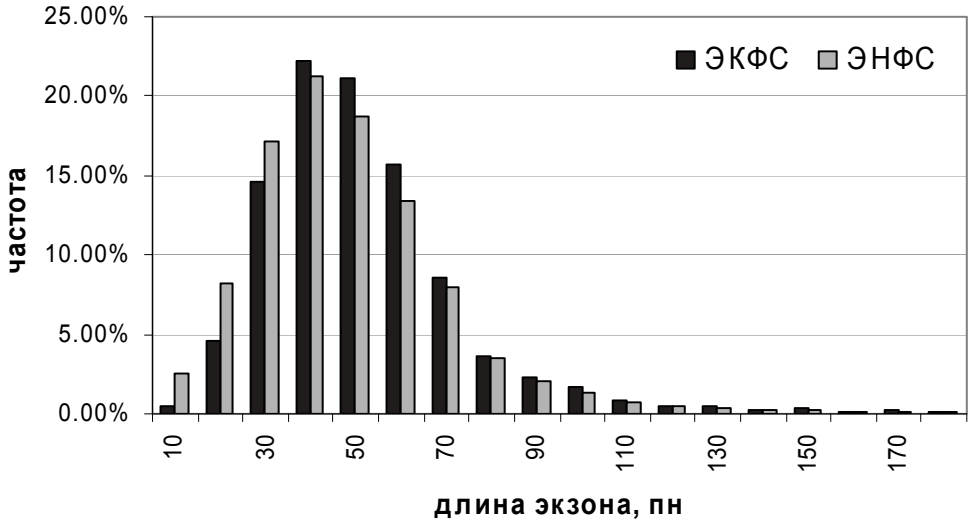

Рисунок 5. Распределения длин экзонов из выборок ЭНФС и ЭКФС.

Выборка ЭКФС включала в себя 6444 экзона, а выборка ЭНФС - 10679 экзонов из 2021 гена. Статистический анализ с помощью χ^2 двух распределений длин экзонов из выборок ЭКФС и ЭНФС показал их значимое различие (χ^2=582.8, p<0.01). При этом средняя длина экзонов из ЭКФС превышала среднюю длину экзонов из ЭНФС. Средние длины экзонов составили ≈159 п.н. и ≈137 п.н., соответственно. Согласно критерию Манна-Уитни, средние значения длин экзонов из этих выборок отличаются со значимостью $p=10^{-6}$. Таким образом, длина ЭКФС, в среднем значимо превышает длину ЭНФС.

Исследование разрывности функциональных сайтов

Расчет коэффициентов разрывности ФС показал, что 27% всех сайтов кодируются одним экзоном и еще 37% кодируются сближенными в последовательности экзонами (CoefE = 0), при этом 95.5% сайтов разрывны по аминокислотной последовательности (CoefA > 0). Достоверно показано, что коэффициенты разрывности сайтов коррелируют между собой (ρ≈0.4, p<0.05).

Для статистической проверки гипотезы о том, что разрывность функциональных сайтов по экзонам значимо меньше, чем ожидается по случайным причинам, оценивалось ожидаемая и наблюдаемая представленность границ экзонов в области функционального сайта при их картировании на аминокислотную последовательность белка. В данном случае в качестве области

функционального сайта рассматривался фрагмент аминокислотной последовательности, ограниченный крайними аминокислотными остатками ФС (рис. 7).

Ожидаемое распределение количества границ экзонов в области функционального сайта рассчитывалось методом 10-кратного повторения случайного выбора позиций границ экзонов в последовательности. Распределения наблюдаемого и ожидаемого количества экзонов сравнивались с помощью критерия χ^2 (χ^2=22.4, p<0.01, df=6).

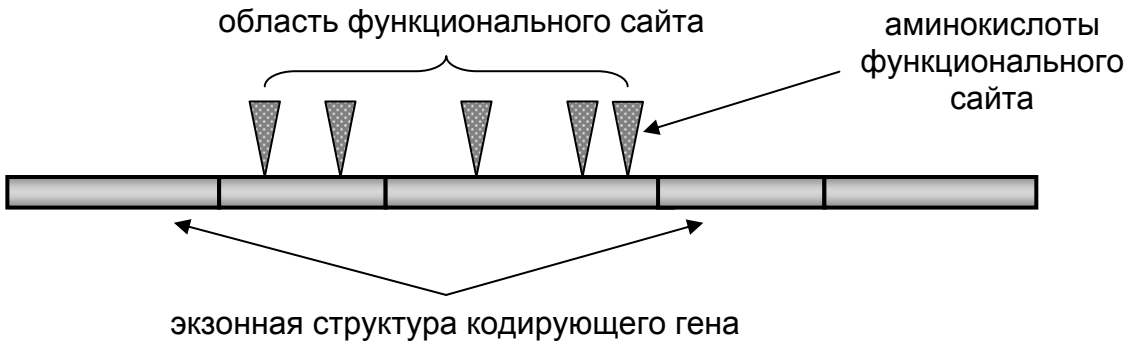

Рисунок 7. Область ФС, отображенная на экзонной структуре кодирующего гена.

Показано, что наблюдаемое количество экзонов, кодирующих фрагменты аминокислотных последовательностей, соответствующих области функциональных сайтов, в среднем значимо меньше количества экзонов, ожидаемых по случайным причинам (значение статистики Манна-Уитни U=52988.5; N_1=427; N_2=390; p<<0.01). Аналогичный анализ проводился для отдельно взятых групп функциональных сайтов (сайтов связывания аминов, спиртов, органических кислот, сложных органический соединений), наиболее представленных в БД SitEx. Во всех случаях наблюдаемое количество экзонов, кодирующих область функционального сайта, было значимо меньше ожидаемого.

На основе проведенного анализа можно заключить, что функциональные сайты белков статистически чаще кодируются одним или близко расположенными экзонами.

Анализ частот кодонов во фрагментах ДНК, кодирующих аминокислотные остатки функциональных сайтов белков

В настоящее время остается актуальной задача изучения влияния кодонного состава на эффективность трансляции белков как у прокариот, так и у эукариот (Zhou et al., 2013). Для анализа частот кодонов в фрагментах ДНК, кодирующих и не кодирующих функциональные сайты белков, использовались данные из БД SitEx. На первом шаге анализа было построено распределение встречаемости аминокислот в рассматриваемых функциональных сайтах белков. Среди наиболее часто встречающихся оказались гидрофильные аминокислоты, предпочтительно располагающиеся на поверхности белков (гистидин, цистеин, тирозин и др.), что характерно для функциональных сайтов (Liao et al., 2013; Betts&Russel, 2003).

Известно, что частоты встречаемости кодонов в последовательностях ДНК вблизи границ экзонов и в остальной части кодирующей последовательности отличаются. В частности, это связывают с сигналами сплайсинга, которые обуславливают богатое содержание пуринов (Gelfand, 1989). Однако существуют работы, в которых авторы выдвигают гипотезу о том, что в таких районах отбор может быть направлен также на нуклеотиды A и T (Parmley et al., 2006). Можно предположить, что на кодонный состав могут влиять не только сигналы сплайсинга, но и другие факторы, такие как кодирование функциональных сайтов. Для проверки данной гипотезы была подсчитана относительная частота встречаемости различных кодонов как в составе функционального сайта, так и на границах экзонов. Для расчета частоты кодонов вблизи границ экзонов рассматривали участки, ограниченные только пятью кодонами на 5'-конце и 3'-конце экзона. Для составления контрольного распределения также рассматривали частоту встречаемости кодонов во фрагментах последовательности экзона, в которых исключены пограничные районы.

Встречаемость кодонов анализировалось с помощью метода матриц 2*J с использованием критерия χ^2. При этом сравнение частот встречаемости кодонов в ДНК проводилось для каждой из 20 канонических аминокислот, за исключением метионина и триптофана, которые кодируются только единственным кодоном. Было показано, что участки ДНК, кодирующие и не кодирующие функциональные сайты в районах 5`-концов экзонов в геноме человека отличается друг от друга по распределению частот встречаемости кодонов, кодирующих аспарагин, пролин, глутамин, глутаминовую кислоту и цистеин (рис. 8).

Данные отличия могут быть объяснены наблюдаемой повышенной частотой представленности аденина и тимина в третьей позиции кодонов, кодирующих перечисленные выше аминокислоты. В частности, полученный результат согласуется с гипотезой Parmley об эволюционном отборе, направленном на нуклеотиды A и T вблизи 5`-конца экзонов. Кроме того, на 5'-конце экзонов в участках, кодирующих функциональные сайты, наблюдалась повышенная частота встречаемости следующих кодонов, содержащих нуклеотиды A, T в третьей позиции: TT<u>T</u> (Phe), AT<u>T</u> (Ile), AA<u>A</u> (Lys), TT<u>A</u>, TT<u>G</u> (Leu), AC<u>A</u>, AC<u>T</u> (Thr), TA<u>T</u> (Tyr), GG<u>T</u> (Gly), CG<u>A</u>, CG<u>T</u> (Arg), AG<u>T</u>, TC<u>T</u>, TC<u>A</u> (Ser). Это может быть следствием влияния генетических сигналов (в частности, сайтов сплайсинга) и кода функциональных сайтов друг на друга.

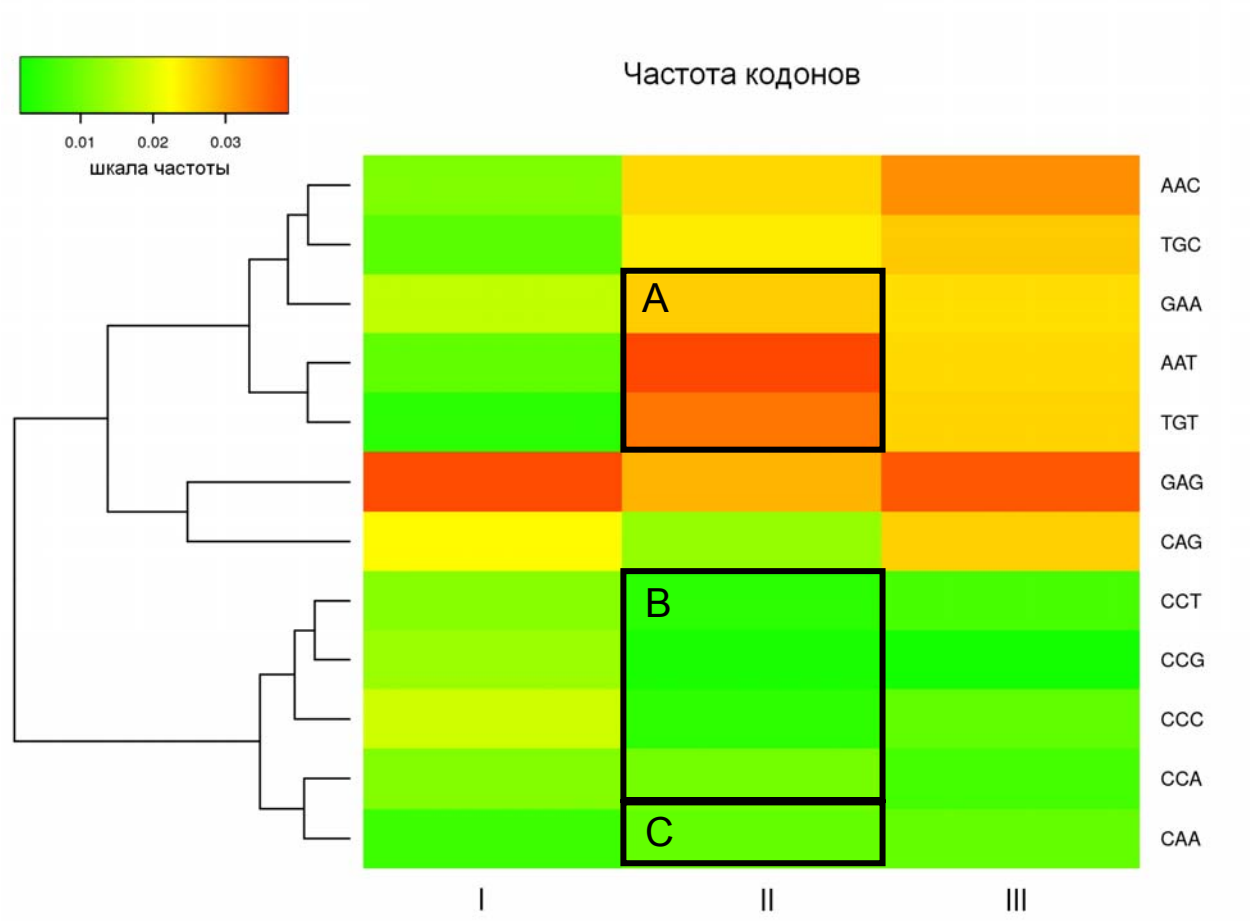

Рисунок 8. Встречаемость кодонов, кодирующих аспарагин, пролин, глутамин, глутаминовую кислоту и цистеин в различных участках последовательности. I – участок последовательности на 5'-конце экзона, ограниченный пятью кодонами; II – часть функционального сайта, кодируемая на 5'-конце экзона, ограниченном пятью кодонами; III – часть сайта, кодируемая между концами экзона, ограниченными пятью кодонами. A – соответствует частотам кодонов, содержащих в третьей позиции A или T, с наибольшей представленностью этих кодонов в рассматриваемом фрагменте ДНК; B – соответсвует кодонам, кодирующим пролин, можно видеть, что наибольшая частота соответствует кодону CCA; C – соответствует кодону CAA, доля которого при кодировании глутамина возрастает за счет снижения встречаемости другого кодона CAG.

Частота фаз экзонов в функциональных сайтах на границе экзонов

Для анализа частот встречаемости различных фаз экзонов, имеющих в крайней 5'-позиции кодон, кодирующий аминокислоту функционального сайта, была создана выборка экзонов из последовательностей генов 14 позвоночных организмов, представленных в БД SitEx. Была подсчитана встречаемость фаз 0, 1, 2 в кодонах на 5'-конце экзонов, которые кодируют аминокислоту функционального сайта (I), и остальных экзонов (II). Всего в анализе участвовало 40 000 экзонов, 1867 из которых содержат на 5'-конце экзона кодон, кодирующий аминокислоту функционального сайта.

Сравнение частот встречаемости фазы 0 между этими двумя группами с помощью парного критерия Вилкоксона показало статистически значимое

различие между распределениями частот для фаз 0 и суммарных частот остальных ($p<8.3*10^{-6}$ с учетом поправки Бонферрони ($Z=4.86$) и $p<8.3*10^{-6}$ ($Z=4.47$) соответственно). При этом среднее и медиана в I группе для фазы 0 были ниже, а для фазы 1 и 2 – выше. Частоты встречаемости различных фаз в выборках представлены на рисунке 9.

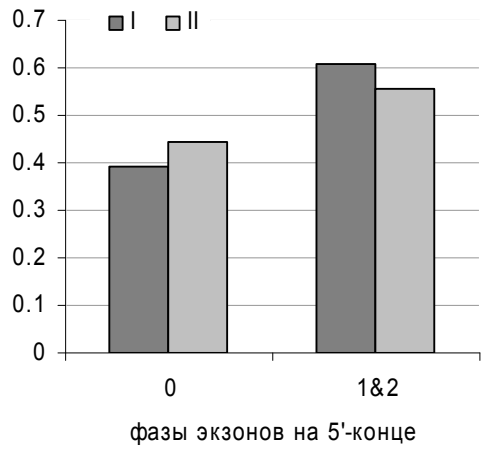

Рисунок 9. Распределение средней частоты фаз среди экзонов, кодирующих аминокислоту функционального сайта на 5'-конце (I) и некодирующих ее (II).

Ранее было показано (Vibranovski et al., 2005), что фаза 0 более часто встречается среди экзонов, имеющих более древнее происхождение, в связи с явлением перетасовки экзонов как одним из основных путей возникновения последовательностей, кодирующих белки с новыми функциями, а фазы 1 и 2 чаще встречаются среди экзонов, имеющих более позднее возникновение. На основе этого можно предположить, что существуют ограничения на перетасовку экзонов, которые кодируют функциональные сайты белка.

ВЫВОДЫ

1. Создана база данных SitEx, содержащая разметку в белковых и геномных последовательностях эукариот границ экзонов, доменов, функциональных сайтов белков и однонуклеотидных полиморфизмов. База данных интегрирована с программами BLAST и 3DPDBScan для поиска участков в первичных и пространственных структурах белков, имеющих сходство с фрагментами белка, кодируемыми одним экзоном в базе данных SitEx.
2. Впервые показано, что функциональные сайты белков имеют тенденцию к кодированию одним или близко расположенными в последовательности гена экзонами. При этом значение показателя разрывности функциональных сайтов по экзонам значимо меньше, чем ожидаемое по случайным причинам.
3. Впервые показано, что длина экзонов, кодирующих функциональные сайты, в среднем значимо превышает длину экзонов, некодирующих функциональные сайты.
4. Впервые показано, что распределение частот представленности различных фаз кодонов, расположенных в районах 5`-концов экзонов, статистически значимо отличаются между кодонами, соответствующими аминокислотным остаткам в позициях функционального сайта белка и не соответствующими им. При этом, оказалось, что фаза 0 кодонов, кодирующих аминокислоты в позициях функциональных сайтов белков, представлена значимо реже по сравнению с кодонами, не соответствующими аминокислотным остаткам функциональных сайтов, что может свидетельствовать об ограничении перетасовки экзонов, при которой происходит разрыв функциональных сайтов белка.

5. Впервые показано отличие частот использования кодонов в участках ДНК, кодирующих функциональные сайты, от участков, не кодирующих функциональные сайты, в районах 5`-концов экзонов в геноме человека. Статистически значимые отличия были получены для кодонов, кодирующих часто встречающиеся в функциональных сайтах аспарагин, пролин, глутамин, глутаминовую кислоту и цистеин. Отличия были обусловлены повышенной частотой встречаемости аденина и тимина в третьей позиции кодонов в участках ДНК, кодирующих функциональные сайты на 5`-конце экзонов. Полученные закономерности могут лежать в основе механизма интерференции генетических сигналов (в частности, сайтов сплайсинга) и кода функциональных сайтов.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи в рецензируемых журналах:

1. Орлов Ю.Л., Брагин А.О., **Медведева И.В.**, Гунбин К.В., Деменков П.С., Вишневский О.В., Левицкий В.Г., Ощепков Д.Ю., Подколодный Н.Л., Афонников Д.А., Гроссе И., Колчанов Н.А. ICGenomics: программный комплекс анализа символьных последовательностей геномики // Вавиловский журнал генетики и селекции. – 2012. – Том 16, 4/1. – с. 732-741.
2. **Medvedeva I.V.**, Demenkov P.S., Kolchanov N.A,, Ivanisenko VA. SitEx: a computer system for analysis of projections of protein functional sites on eukaryotic genes // Nucleic Acids Res. – 2012. – Vol. 40(D1) – p. D278-283.
3. **Медведева И.В.**, Деменков П.С., Иванисенко В.А.. Анализ распределения аденозин-фосфат связывающих сайтов белков на экзонной структуре гена // Информационный Вестник ВОГиС. – 2009. – Том 13, №1. – с. 122-127.

Свидетельства:

Медведева И.В., Деменков П.С., Иванисенко В.А. (2013) Свидетельство о государственной регистрации базы данных № 2013621254. Позиции аминокислот функциональных сайтов белков в экзонной структуре кодирующих генов (СайтЭкс)/Protein functional sites positions in exon structure of the coding genes (SitEx).

Тезисы конференций:

1. **Medvedeva I.V.**, Demenkov P.S., Ivanisenko V.A. Influences of protein functional site encoding features on protein evolution in Eukaryota. // Abstracts of the Eighth International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2012), Novosibirsk, Russia, June 25- 29, 2012, p.209.
2. **Medvedeva I.V.**, Demenkov P.S., Ivanisenko V. A. Computer system SitEx for analyzing protein functional sites in eukaryotic gene structure. // Abstracts of the Seventh International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2010), Novosibirsk, Russia, June 20- 27, 2010, p.182.
3. **Medvedeva I.V.**, Demenkov P.S., Ivanisenko V. A. Protein functional site projection on exon structure of gene. // Abstracts of the Sixth International

Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2008), Novosibirsk, Russia, June 22- 28, 2008, p.159.
4. **Medvedeva I.V.,** Demenkov P. S., Ivanisenko V. A. (2007) Analysis of protein functional site distribution on gene structure. Proceedings of the 2007 international conference on bioinformatics and computational biology (BIOCOMP'07). Vol. 2, pp. 452-455.
5. **Медведева И.В.** Анализ картирования функциональных сайтов белков на экзонной структуре гена. Материалы докладов XIV Международной конференции студентов, аспирантов и молодых ученых «Ломоносов». Москва. 2007. стр. 58.
6. **Медведева И.В.** Анализ распределения просайтов функциональных сайтов в пространственных структурах белков. Материалы XLIV Международной студенческой конференции «Студент и научно-технический прогресс». Биология. Новосибирск. 2006. стр. 146.

Подписано к печати
Формат бумаги 60 х 90 1/16. Печ. л. 2. Уч.изд.л. 1,4
Тираж 100 экз. Заказ 101.

Ротапринт Института цитологии и генетики СО РАН
630090, Новосибирск, проспект академика Лаврентьева, 10.