РОССИЙСКАЯ АКАДЕМИЯ НАУК

Сибирское отделение

Федеральное государственное бюджетное учреждение науки

Институт цитологии и генетики

На правах рукописи

Медведева Ирина Вадимовна

КОМПЬЮТЕРНЫЙ АНАЛИЗ ЗАКОНОМЕРНОСТЕЙ КОДИРОВАНИЯ ФУНКЦИОНАЛЬНЫХ САЙТОВ БЕЛКОВ В ГЕНАХ ПОЗВОНОЧНЫХ

(03.01.09) «математическая биология, биоинформатика»

Диссертация на соискание учёной степени кандидата биологических наук

Научный руководитель:

доцент, к.б.н. Иванисенко В.А.

Новосибирск – 2014

Содержание

Введение

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 [1]. Однако дальнейшие исследования показали, что корреляция между границами доменной и экзонной структур наблюдается не всегда [2]. 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 взаимосвязь между структурной организацией функциональных сайтов и особенностями молекулярной эволюции генома 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.

Цели и задачи работы

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, 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;
- Распределение частот представленности различных фаз кодонов в районах 5`-концов экзонов, статистически значимо отличается между кодонами, кодирующими и не кодирующими аминокислоты в позициях функционального сайта белка;
- 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 работы

Основные результаты работы были представлены на следующих конференциях:

- Восьмая Международная конференция по биоинформатике регуляции и структуры генома (BGRS'2012). Россия, Новосибирск, июнь 25-29, 2012, устный доклад.
- 19th Annual International Conference on Intelligent Systems for Molecular Biology and 10th European Conference on Computational Biology. Австрия, Вена, июль 17-19, 2011, постер
- 2011 International German/Russian Summer School on Integrative Biological Pathway Analysis and Simulation. Германия, Билефельд, июль 4-7, 2011, устный доклад.

- Седьмая Международная конференция по биоинформатике регуляции и структуры генома (BGRS'2010). Россия, Новосибирск, июнь 20-27, 2010, постер.
- Школа Молодых Ученых (YSS`2010). Россия, Новосибирск, июнь 28-29, 2010, устный доклад
- International Autumn School for Young Scientists on Computational Systems Biology and Bioinformatics 2008. Россия, Новосибирск, сентябрь 24, 2008, устный доклад
- Шестая Международная конференция по биоинформатике регуляции и структуры генома (BGRS'2008). Россия, Новосибирск, июнь 22-28, 2008, устный доклад.
- The 2007 International Conference on Bioinformatics & Computational Biology (BIOCOMP`07). США, Лас-Вегас, июнь 25-28, 2007, постер.
- Международная научная конференция студентов, аспирантов и молодых учёных "Ломоносов-2007". Москва, Россия, апрель 8-12, 2007, устный доклад.

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.

Статьи в рецензируемых журналах:

- 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.

- **Medvedeva I.V.**, Demenkov P.S., Kolchanov N.A., Ivanisenko V.A. 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.
- **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:

- **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.
- **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.
- **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.
- **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.

- **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.
- **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.

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 3DPDBScan 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 четырех 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.

Автор участвовал в работах по грантам Министерства образования и науки (гранты 14.740.11.0001, 07.514.11.4003, 8740); междисциплинарных интеграционных проектах СО РАН (94, 111, 119); РФФИ (11-04-92712); FP7: EU-FP7 SYSPATHO No. 260429; программ РАН (A.II.5, A.II.6, B.21, B.26) и гранте Леонарда Эйлера DAAD.

Список сокращений

3D – пространственная структура

ФС – функциональный сайт белка

ЭКФС – экзон, кодирующий хотя бы часть функционального сайта

ЭНФС – экзон, не кодирующий функциональный сайт

АФС – аминокислота функционального сайта

Глава 1. Обзор литературы

1.1 Пространственная структура белка

Благодаря своей сложной структуре и огромному разнообразию, белки участвуют во множестве процессов: инициации транскрипции, ферментативном катализе, передаче сигналов, распознавании чужеродных молекул, образовании мембранных каналов, сокращении мышечных клеток и многих других. Это было бы невозможно без разнообразия пространственных структур белков. В свою очередь, пространственная структура белка, зависит от физико-химических свойств аминокислот, составляющих последовательность белка.

1.1.1 Физико-химические свойства аминокислот

Последовательность белка кодируется 20 различными каноническими аминокислотами (обозначения представлены в таблице 1.1).

Таблица 1.1. Однобуквенные и трехбуквенные обозначения аминокислот

A	Ala	Аланин
C	Cys	Цистеин
D	Asp	Аспарагиновая кислота
E	Glu	Глутаминовая кислота
F	Phe	Фенилаланин
G	Gly	Глицин
H	His	Гистидин
I	Ile	Изолейцин
K	Lys	Лизин
L	Leu	Лейцин
M	Met	Метионин
N	Asn	Аспарагин
P	Pro	Пролин
Q	Gln	Глутамин
R	Arg	Аргинин
S	Ser	Серин
T	Thr	Треонин
V	Val	Валин
W	Trp	Триптофан
Y	Tyr	Тирозин

Аминокислоты можно классифицировать на группы по их физико-химическим свойствам (Рисунок 1.1). По свойствам боковых радикалов аминокислоты разделяют на несколько классов: 1) неполярные (A, V, L, I, F, P, M, C); 2) положительно заряженные (K, R); 3) отрицательно заряженные (E, D); 4) полярные незаряженные (S, T, N, Q, Y, W, H); 5) G, имеющий в боковой цепи только один атом водорода, обладает отличными свойствами и его относят к отдельному классу или к первому из указанных [3]. В силу столь малого объема боковой цепи, которая практически не создает стерических трудностей при конформационных изменениях полипептидной цепи, глицин необходим для обеспечения гибкости белка. Минимальный объем глицина также накладывает сильные ограничения на возможность его замены на другие аминокислоты в случае, когда он располагается внутри белковой глобулы. Такие замены не могут проходить без подвижек внутри всей молекулы, в силу того, что заменяющие аминокислоты имеют больший объем по сравнению с глицином, что, как правило, ведет к нарушению пространственной структуры белка.

Рис. 1.1. Классификация аминокислот У. Тэйлора по их физико-химическим свойствам на основе метода кругов Эйлера (1986) [4,5].

Между боковыми цепями полипептида действуют такие слабые взаимодействия, как: ионные, водородные, Ван-дер-ваальса [6]. Кроме того, на стабилизацию пространственной структуры влияют цистеиновые мостики (связи S-S). Водородные связи образуются между группами атомов акцептор–донор. Две боковые цепи, имеющие разный заряд – положительный и отрицательный - формируют солевой мостик.

Гидрофобность определенных аминокислот обуславливает важный эффект в процессе сворачивания белка: чтобы избежать контакта с водой гидрофобные боковые цепи полипептида разворачиваются внутрь белка, формируя гидрофобную сердцевину. В гидрофобной среде атомы основной цепи образуют водородные связи и таким образом формируются, элементы вторичной структуры белка [3]. Кроме того, внутри структуры белка иногда присутствуют полости, заполненные водой и изолированные от растворителя, с молекулами воды которых полярные боковые цепи также могут взаимодействовать. Подобные полости также часто являются областями связывания различных лигандов (атом, ион или молекула, непосредственно связанные с боковыми группами аминокислот в составе белка). Среднестатистический мономерный белок имеет на своей поверхности следующий аминокислотный состав: 58% гидрофобных (неполярных) аминокислот, 29% полярных, 13% заряженных аминокислот; внутри гидрофобного ядра состав аминокислот приблизительно следующий: 60% неполярных, 33% полярных, 7% заряженных аминокислот [7].

Вследствие разнообразия взаимодействий, действующих на пространственную структуру полипептидов, различают несколько уровней организации структур белка:

1) Первичная структура – аминокислотная последовательность белка
2) Вторичная структура – единица пространственной организации полипептидов
3) Третичная структура (3D-структура) – пространственная структура белка.

4) Четвертичная структура – взаимная пространственная ориентация комплекса белков либо нескольких полипептидных цепей.

1.1.2 Вторичная структура полипептидов

Важнейшей характеристикой структуры белка является его вторичная структура, образуемая за счет водородных связей между атомами основной цепи. Другой особенностью вторичной структуры является наличие фиксированных конформаций основной цепи, при которых конформации боковых цепей неважны. Наиболее широко распространены α–спираль и β-лист.

Спирали могут различаться по направлению вращения (право- и левозакрученные), периоду (количеству аминокислотных остатков) и шагу (длине витка). Направление спираль считается от N-конца к C-концу полипептида. α–спираль имеет период 3.6, т.е. группа C=O *i* аминокислотного остатка в последовательности соединяется водородной связью с группой H-N *i+4* остатка. В белках в основном встречается правозакрученная (против часовой стрелки) α–спираль как наиболее стабильная. Известны такие спирали: 2_7 (в белках не встречается), 3_{10}(связь *i* – *i+3*), 4_{13}(α–спираль), 5_{16} (π-спираль, связь *i* – *i+5,* встречается в белках очень редко). Нижний индекс обозначает число атомов основной цепи между группами донора и акцептора, образующими водородную связь, поддерживающую соответствующую вторичную структуру [3, 8]. Для разных типов вторичных структур существует предпочтительность аминокислот образовывать ту или иную структуру. Например, такие аминокислоты как аланин (A), глутамат (E), лейцин (L), метионин (M) встречаются чаще других в α–спиралях. С другой стороны, пролина, глицин и тирозин встречаются редко в спиралях [3].

Регулярная структура, образованная водородными связями между удаленными участками белка, формирует β-лист. β-структура может быть параллельной, антипараллельной и смешанной. Поверхность β-листа

складчатая, а сам лист имеет небольшую скрученность вправо за счет стерически выгодных конформаций [8, 9, 10].

Помимо регулярных вторичных структур существуют и нерегулярные: β-изгибы и петли. β-изгибы формируются между участками полипептида, задействованных в формировании антипараллельного β-листа. Петли обычно располагаются на поверхности белка и могут участвовать в образовании функционального сайта белка. Большая часть петель обладает стабильной структурой, однако, есть и неупорядоченные петли [8, 9].

Статистические закономерности встречаемости определенных аминокислотных остатков в различных участках вторичной структуры белка: в составе α–спирали, β-листа, нерегулярной структуры или гидрофобного ядра приведено в Приложении 1 [8].

Между вторичными струкутрами существуют взаимодействия, в частности, α–спирали за счет амфипатичности могут взаимодействовать друг с другом гидрофобными фрагментами, образуя «пучок прутиков». Во взаимодействия между вторичными структурами могут быть вовлечены как ковалентные связи (S-S мостики), слабые взаимодействия, а также стекинг, или π-π взаимодействия между ароматическими аминокислотами. Около 60% всех ароматических аминокислот белка вовлечены в π-π взаимодействия, при этом их большая часть осуществляется со сдвигом в параллельной плоскости, а меньшая – перпендикулярно друг к другу. Они играют значительную роль при сворачивании белка [11].

1.1.3 Классификация структур белков

Чем больше расшифровывалось пространственных структур белков, тем тем понятнее становилось, что белки, даже разные по функции и по последовательности, имеют общие элементы пространственной структуры. Так было введено понятие *мотива укладки* - взаимная пространственная ориентация вторичных структур в составе пространственной структуры белка. *Укладка белка* – это структура, образованная атомами основной

полипептидной пептидной цепи. Таким образом, в основу классификации структур белков легла классификация мотивов укладки. Всего насчитывается 1000-2000 мотивов укладки, хотя по некоторым оценкам их кличество может возрасти до 7000 [12,13,15,16]. На сегодняшний день выделеляют четыре основные группы структур, описывающие укладку большей части всех белков [10]:

1) только α – вторичная структура включает α–спирали, но не β-листы
2) только β – вторичная структура включает β-листы, но не α–спирали
3) α./β – чередование α–спиралей и β-листов,
4) α+β - α–спирали и β-листы присутсвуют в структуре, но не чередуются

Наиболее известные классификации представлены в ресурсах SCOP[16] и CATH[17].

1.1.4 Доменная структура белка

Доменная структура белка определяется взаимным расположением доменов в пространственной и первичной структурах одного белка. Ее исследование позволяет получить важную информацию о функции белка. В белках различают структурные, функциональные и эволюционные домены [18]. При этом разные типы доменов могут либо совпадать, либо не совпадать друг с другом.

Структурный домен определяют как обособленную в пространстве часть белка, способную к самосборке в нативную структуру, имеющую сравнительно мало контактов с другими частями белка и собственное гиброфобное ядро.

Функциональный домен - минимальная часть полипептидной цепи, способную к самосборке в нативную структуру и обладающую той же целевой функцией, что и в составе полноразмерного белка [18].

Эволюционный домен - непрерывный участок полипептидной цепи, эволюционирующий существенно медленнее других участков, является эволюционной единицей в перетасовке доменов.

В 1981 году Го также определил термин «модуль» [19, 20]. Это структурная единица, определяемая диаметром в пределах 15-35 Å. Эта структура также рассматривалась как эволюционная единица (см. раздел 1.3.1). Кроме этого, существуют свидетельства того, что модули могут функционировать независимо, вследствие чего было предположено, что модуль – первоначальная функциональная единица белка [21].

Для проведения биоинформатических исследований наиболее часто используются домены из базы данных Pfam [22]. Понятие домена, используемое в Pfam, базируется на поиске консервативных участков гомологичных последовательностей белка из различных организмов. Ядро множественного выравнивания аминокислотных последовательностей для каждого из функциональных семейств, определенных в PFAM, задавалось путем ручного анализа экспертов, с учетом функциональной аннотации каждого из гомологов. Затем, каждое из таких ядер подвергалось автоматическому расширению путем добавления выравнивания новых гомологов. При выравнивании учитывалось также сходство пространственных структур белков [23].

1.1.5 Существующие компьютерные ресурсы по пространственной структуре белков и анализу ее особенностей

Первые пространственные структуры белка (миоглобина и гемоглобина) были расшифрованы в конце 1950х годов Джоном Кендрю [24] и Максом Перуцем [25] с помощью рентгеноструктурного анализа. В 1980х годах Карлом Вютрихом и Ричардом Эрнстом были разработаны методы определения трехмерной структуры биологических молекул с помощью ядерно-магнитного резонанса [26, 27, 10]. Также разновидностью электронной микроскопии, проводимой при низких температурах, является криоэлектронная микроскопия, применяемая для распознавания структур крупных белковых комплексов с середины 1980х годов [10, 28].

Знание пространственной структуры помогает определить положение функциональных сайтов, элементов вторичной структуры и отдельных доменов. С 1990х годов расшифрованные пространственные структуры белков стали помещаться в единый банк структур – Protein Data Bank (PDB) – в специальном формате данных, включающим координаты атомов [29]. До февраля 2009 года не было единого формата данных. Помимо координат атомов и информации о структурных элементах, отмеченных выше, формат включает в себя информацию об авторе, организме, молекулах растворителя, подробностях эксперимента, последовательности, отсутствующих в структуре атомах, лигандах и идентификаторах в других базах данных.

На основе PDB было создано множество ресурсов, однако основные из них посвящены классификации пространственных структур. В частности, SCOP (поддерживается экспертным курированием базы)[16] и CATH (поддерживается автоматическим курированием)[17]. Среди баз данных, посвященных доменам белков, можно выделить PROSITE [30], BLOCKS [31], PRINTS [32], SUPERFAMILY [33], CDD [34], TIGRFAM [35], Panther [36], ProDom [37], EVEREST [38], Pfam [22] и SMART (Simple Modular Architecture Research Tool [39]. Большая их часть основывается на информации о консервативных участках последовательности различной протяженности, некоторые аннотируются экспертами, другие – автоматически.

1.2 Структурная организация функциональных сайтов белков

Традиционно функции белков подразделяют на каталитическую, структурную, защитную, регуляторную, сигнальную, транспортную, рецепторную, моторную и запасающую. В пост-геномную эру, с развитием экспериментальных высокопроизводительных транскриптомных, протеомных и метаболомных технологий появилась возможность полногеномного профилирования молекулярно-генетических взаимодействий и экспрессии белков. Это позволило более полно описывать

биохимическую и системную функцию белка, включая клеточный, тканевой и организменный уровень [40, 41]. Развитие методов кристаллизации белков в сочетании с методами рентгеноструктурного анализа обеспечили расшифровку десятков тысяч пространственных структур белков из различных организмов. Информации о пространственных структурах белков, в свою очередь, послужила основой для изучения структурной организации функциональных сайтов белков и биохимических механизмов их функционирования.

Функциональный сайт белка – группа аминокислотных остатков, непосредственно участвующая во взаимодействиях белка с лигандами/рецептором или биохимических реакциях, обеспечивающих выполнение его функции. Функциональные сайты обладают свойством компактности в пространственной структуре, т.е., их аминокислотные остатки сближены в пространстве между собой, но могут быть удаленно распределены по последовательности. Среди различных типов функциональных сайтов можно выделить следующие:

1) активные центры (посредством которых катализируются химические реакции, включают в себя каталитические сайты и сайты связывания субстрата). Существует классификация ферментов, разработанная совместно с IUPAC, отражающая также функциональное деление активных центров [42];
2) связывающие лиганды (могут связывать либо макромолекулы - белки, ДНК, РНК, - либо небольшие молекулы);
3) аллостерические (сайты, связывание лигандов с которыми может изменять конформацию белка и, таким образом, регуляторно воздействовать на функцию других удаленных в пространстве сайтов);
4) регуляторные (могут регулировать ферментативную активность белков с помощью активаторов и ингибиторов). Аллостерические сайты являются подгруппой регуляторных сайтов;

5) сайты посттрансляционной модификации белков. К сайтам посттрансляционных ковалентных модификаций относятся сайты фосфорилирования, метилирования, ацетилирования, гликозилирования, присоединения жирной кислоты и др. Описано более 100 видов посттрансляционных модификаций белков [43].

Лиганд-связывающие сайты подразделяются на множество сайтов, в зависимости от типа лиганда (Таблица 1.2). Лигандами могут быть другие белки, ДНК, РНК, или ионы тяжелых металлов, аденин-содержащие кофакторы, органические кислоты и т. п. При классификации таких сайтов могут учитываться размер лиганда, характер химических групп, входящих в состав лиганда и т. д. [44].

Таблица 1.2. Классификация лиганд-связывающих сайтов, представленных в базе данных PDBSite [45].

Тип	кол-во разнообразных групп	кол-во
связывающие ионы металлов	17	1506
связывающие неорганические вещества и неметаллосвязывающие	9	657
связывающие органические лиганды	133	1130
белок-белковые взаимодействия	995	1002
сайты белок-ДНК	1324	1329
сайты белок-РНК	752	755
сайты лекарственных лигандов	14	28
неклассифицированные	0	2303

Структурная организация сайтов во многом определяется той функцией, которую они выполняют. В частности, лиганд-связывающие сайты преимущественно располагаются во впадинах на молекулярной поверхности белка, что способствует увеличению площади контакта лиганда с белком и афинности связывания. Можно выделить следующие структурные особенности расположения лиганд-связыващих сайтов [8]:

1. в воронке на торце β-цилиндра. Такое углубление способствует окружению субстрата одновременно многими боковыми цепями белка;
2. в месте стыка доменов.

Активные центры, как правило, не доступны растворителю, они расположены в углублениях на поверхности белка и часто оказываются доступны для взаимодействия с субстратом и осуществления каталитической реакции только после конформационных изменений структуры активного сайта [46]. Каталитические сайты и субстрат-связывающие сайты могут пересекаться между собой в структуре активного центра, либо располагаться раздельно друг от друга, образуя распределенный активный центр. Например, активный центр многих сериновых протеаз, липаз и сериновых карбоксипептидаз помимо каталитического сайта включает в себя расположенную по соседству впадину, роль которой заключается в образовании водородных связей с атомами-акцепторами и образовании комплекса с субстратом [47]. При наличии гидрофобных частей лиганда связывание, как правило, осуществляется гидрофобными аминокислотными остатками [48].

Активные центры ферментов и сайты связывания ДНК, РНК расположены на поверхности, имеющей высокий электростатический потенциал. То, насколько важен электростатический потенциал, показывает пример сериновых протеаз [48], чьи каталитические триады могут различаться по аминокислотному составу, положению в структуре различных укладок, но проявляют высокую консервативность по отношению к заряду, т.е. имеют одинаковый заряд каталитического кармана.

Лиганд-свзяывающие сайты при связывании с лигандом могут оказывать аллостерический эффект друг на друга. Примером может служить связывание гемоглобином кислорода. Молекулы кислорода выступают аллостерическими регуляторами при связывании с ионом железа гемма, входящего в состав гемоглобина. Связывание хотя бы одной молекулы кислорода приводит к конформационным изменениям белка и повышает его аффинность для последующего связывания других молекул кислорода. Высокоафинная форма гемоглобина называется R-формой (англ. relaxed), а кислород является положительным гомотропным эффектором. Верно и

обратное, частичная потеря молекул кислорода приводит к снижению уровня сродства гемоглобина к кислороду. CO_2, H^+ и дифосфоглицерат (ДФГ) являются естественными гетеротропными эффекторами, стабилизирующими низкоафинную форму гемоглобина Т (англ. tense). ДФГ также оказывает аллостерический эффект, который интерферирует с эффектом, вызываемым молекулами кислорода. При связывании гемоглобином молекул кислорода стерическое влияние связанного с гемоглобином ДФГ, расположенного в центральной полости тетрамера гемоглобина, ослабляется [10, 49].

Способность белка связывать лиганд зависит от физико-химических свойств аминокислот. Алифатические аминокислоты – аланин, валин, лейцин, изолейцин - обычно не вступают во взаимодействия, но могут осуществлять распознавание лиганда и обеспечивать его связывание. Такие аминокислоты как: лизин, аргинин, глутамин, глутаминовая кислота – являются дифильными, т.е. гидрофильными и гидрофобными одновременно. Ароматические аминокислоты помимо свойства гидрофильности также обладают способностью образовывать стэкинг и Т-стэкинг взаимодействия. В частности, они могут связывать, например, полипролиновые участки молекул (домен SH3) [50]. Случай Т-стэкинга можно рассмотреть на примере ингибитора тромбина: ароматическое кольцо ингибитора расположено под углом к плоскости ароматического кольца триптофана [51]. Кроме того, важную роль в этих взаимодействиях играет гистидин: часто встречается в случае обмена доменами, белок-ДНК взаимодействиях и каталитических остатках [52]. Одним из частных случаев химической связи рассматривается образование водородной связи перпендикулярно к ароматической молекуле. Например, комплекс белка GGBP, связывающего глюкозу между фенилаланином и триптофаном [53]. Еще одним типом взаимодействий являются пи-катионные, т. е. образующиеся между ароматическим аминокислотным остатком и N-H связью лиганда [11].

Роль аминокислотных остатков в каталитических сайтах до сих пор активно изучается [54, 55]. В частности, в сериновых протеазах активный

сайт, как правило, состоит из глутаминовой и аспартановой кислоты, гистидина и серина. Гистидин и кислота являются акцепторами атома водорода серина, серин же временно связывает N-конец новообразованного пептида.

В случае реакции образования тирозил-АМФ из тирозина и АТФ на ферменте тирозил-тРНК синтетазе за связывание в процессе переходного состояния отвечают гистидин и треонин за счет водородных связей.

1.3. Влияние мутаций на структуру и функцию белка

Обычно рассматривают три основных воздействия мутаций на белок, это влияние мутаций на его функцию, термодинамическую стабильность и синтез [56]. Мутации в белке могут быть повреждающими (запрещенными), нейтральными, либо благоприятными [57]. При этом замечено, что чем более белок функционально нагружен (например, участвует в белок-белок взаимодействиях), тем больше мутаций оказываются запрещенными [56].

Аминокислотные остатки в различных позициях могут по-разному влиять на структуру и функцию конкретного белка. Некоторые аминокислотные остатки важны только для функции белка, другие необходимы для поддержания пространственной структуры. Третьи не так важны для структуры, но опосредованно влияют на внутри- или межмолекулярные взаимодействия. Про четвертые нельзя сказать ничего. Самые важные аминокислотные остатки являются консервативными, часто формируя консервативные паттерны. Мутации, приводящие к замене аминокислотных остатков в функциональном сайте, как правило, не фиксируются. В случае взаимодействия аминокислотных остатков на поверхности между белками наблюдается их коэволюция, т.е. если на поверхности одного белка произошла мутация, то на аминокислотный остаток другого белка будет действовать отбор, и может возникнуть компенсаторная мутация [58]. Мутации аминокислотных остатков вне функциональных сайтов также могут

влиять на стабильность и афинность связывания через аллостерический эффект [59].

Остатки, погруженные внутрь гидрофобного ядра, как правило, сохраняют гидрофобные свойства. Аминокислотные остатки на поверхности белка наоборот могут свободно мутировать. Петли на поверхности белка также нечувствительны к вставкам и делециям. В то же время взаимодействия между аминокислотными остатками ограничивают возможности молекулярной изменчивости [10].

Очевидно, что одна и та же мутация может иметь множественный эффект на функцию и структуру белка. Мутации, нейтральные с точки зрения функции, могут быть благоприятны для повышения стабильности пространственной структуры белка, однако при этом могут исключаться последующие мутации, которые могли бы быть функционально выгодны, но запрещены с точки зрения стабильности структуры. В то же время такие мутации могут приводить к увеличению количества функциональных сайтов белка или расширению области его специфичности. До определенного момента такие мутации могут не подвергаться отбору, но впоследствии являться ключевыми при адаптивной эволюции новых функций данного белка [60]. Подобные мутации могут влиять не только на эволюцию этого белка, но и на эволюцию всего функционального процесса, в котором задействован белок [61].

Следует также выделить мутации, ведущие к образованию новых функциональных сайтов. Новообразованный функциональный сайт может конкурировать за субстрат или нарушать конформацию ранее существующего в белке функционального сайта. Как правило, устойчивое сосуществование этих двух функций наблюдается в случае, если лиганды этих сайтов различны по размеру или заряду. Обычно такая функциональная дивергенция основывается на предшествующей дупликации или различных компенсаторных механизмах, например, регуляции экспрессии белка. Если устойчивого существования не наблюдается, в большинстве случае такие

мутации приводят к приобретению новой функции и утрате прежней. Также наблюдается возникновение возможности для белок-белковых взаимодействий [61].

1.4 Базы данных, посвященные функциональным сайтам белков

Описание функциональных сайтов в базах данных, как правило, содержит информацию о позициях аминокислотных остатков, входящих в его состав, и их взаимного пространственного расположения в структуре белка. Многие базы данных также содержат функциональные мотивы (как в первичной, так и в пространственной структуре), например, PROSITE [30]. Экспериментальные данные о положении функциональных сайтов в пространственной структуре белков часто приводятся в базе данных PDB [29], являющийся официальным депозитарием координат атомов пространственных структур белков. На основе экспертного реферирования литературы создана Catalytic Site Atlas (CSA) – база данных каталитических сайтов белков с известной пространственной структурой [62, 63]. Существуют также вторичные базы данных, содержащие информацию о функциональных сайтах, полученную в результате компьютерного анализа экспериментальных данных. Наиболее известны среди них следующие:

SitesBase – база данных на основе информации из PDB, посвященная сайтам распознавания и связывания лигандов [64].

PDBSite – база данных функциональных сайтов различных типов (сайты посттрансляционной модификации, каталитические, сайты связывания органических и неорганических лигандов, сайты белок-белковых и белок-ДНК взаимодействий и т. д.) на основе информации о сайтах из PDB и о контактных сайтах на основе информации о структуре гетерокомплексов [45].

eF-Site – база данных функциональных сайтов, посвященная электростатическим свойствам молекулярной поверхности функционального сайта [65].

sc-PDB - база данных сайтов связывания лигандов, построенная на основе скрининга структур PDB на предмет поиска полостей на поверхности белка, которые могли бы связывать небольшие лиганды, в частности, с целью фармакологического применения [66].

FireDB – представляет собой интегрированную информацию о функциональных сайтах как из PDB, так и из CSA [67].

LigASite – база данных таких функциональных сайтов белков, для которых известна как апо-структура (структур белка до связывания с лигандом), так и холо-структура (структура белка после связывания с лигандом) [68].

Эти базы данных содержат информацию об особенностях организации функциональных сайтов в структуре белка, которая может быть расширена информацией об экзон-интронной организации генов, кодирующих соответствующие белки. Однако ресурсов по функциональным сайтам, предоставляющих информацию о взаимоотношении функциональных сайтов с экзон-интронной структурой гена до сих пор не создано.

1.3 Эволюция структуры и функции белков

Всестороннее изучение особенностей кодирования функциональных сайтов помимо анализа структурно-функциональной организации белков также требует исследования экзон-интронной структуры гена. Геномы позвоночных характеризуется большим объемом (обычно более 1Гб), при этом большая часть последовательности ДНК является некодирующей [69]. На рисунке 1.2 показана общая структура гена эукариот и его экзон-интронная структура.

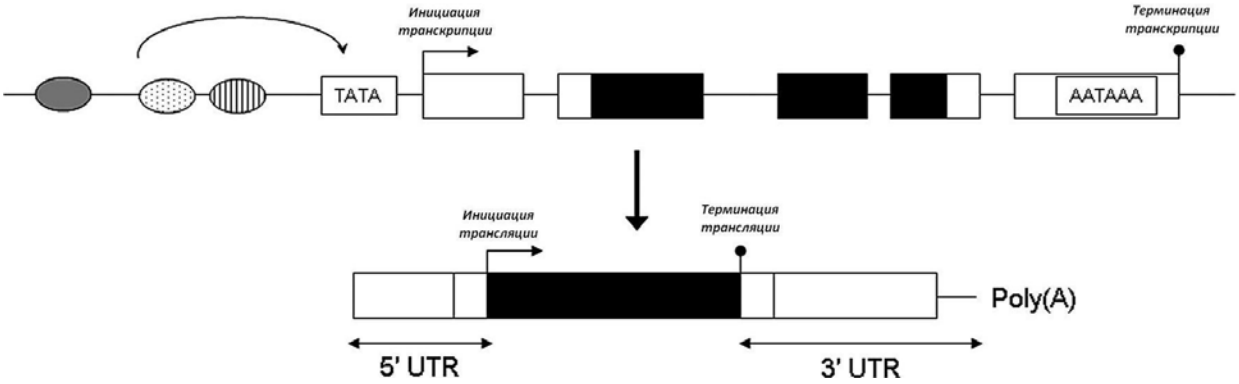


Рис. 1.2. На рисунке представлена общая структура гена эукариот и его экзон-интронная организация. Овалами отмечены сайты связывания транскрипционных факторов. Поли(А) – поли(А)-тракт на 3'-конце молекулы РНК, 5'UTR – 5' нетранслируемая лидерная область, 3'UTR – 3' нетранслируемая трэйлерная область. В качестве примера промоторной области приведен TATA-бокс. Черным цветом показана кодирующая белок область, белым – некодирующая [70].

Кодирующая область ДНК представлена экзонами, которые имеют разные длины. Статистика распределений длин экзонов представлена в таблице 1.3. При этом значения параметров распределения длин экзонов близки для человека, мыши и крысы. [71]. Число экзонов в структуре одного гена эукариот колеблются в количестве 2-3 у одноклеточных и в среднем до 6-9 у позвоночных [72,73].

Таблица 1.3. Распределение длин экзонов в генах различных организмов [71].

Организм	Количество экзонов	Средняя длина (п.о.)	Дисперсия	Минимум	Медиана	Максимум
Homo sapiens	232892	170.9	262.33	1	124	17106
Mus musculus	198344	179.5	257.51	1	126	11544
Rattus norvegicus	152885	181.9	269.03	1	129	11853

1.3.1 Пути эволюции генов эукариот

Существует несколько механизмов возникновения новых генов: ***дупликация гена*** (или экзона) с последующей дивергенцией функции [57],

перетасовка экзонов (а также слияние генов, разрыв гена) и ***горизонтальный перенос генов*** (между геномом эукариот и геномами внутриклеточных органелл (митохондрий, пластид), прокариот и эукариот [74]). Основным понятием в анализе последовательностей генов является понятие ***гомологии***. Гомологичные гены различают двух типов: ***ортологи*** (гены разных видов, имеющие одного предка) и ***паралоги*** (гены, образованные путем дупликации внутри одного генома). При этом ортологи обычно обладают сходными функциями. В то же время паралоги в большинстве случаев представляют различные функции [75].

Под действием отбора происходит дивергенция функций белков, кодируемых как генами-ортологами, так и генами-паралогами. Как правило, большинство различий по последовательности ортологов не связаны с приобретением белков новой функции. Помимо изменения функциональности, различия могут быть также вызваны адаптацией к другим внутриклеточным компартментам, тканям или коадаптацией к измененным рецепторам или лигандам, с которыми белок взаимодействует [61].

Дупликация генов является частым событием. По некоторым оценкам в различных геномах от 1% до 10% генов сосуществуют со своими почти точными копиями [76], также присутствует вариация количества копий генов между отдельными особями [61]. Известно, что вследствие дупликации генов у эукариот наблюдается кластеризация паралогов в геноме [69]. В то же время тандемные дупликации приводят к увеличению частоты неравного кроссинговера [57].

Эволюция генов путем диверсификации функции кодируемых белков может происходить по одной из трех моделей (рис. 1.3) [61].

1) *Модель Оно.* Согласно этой модели дупликация гена является нейтральным событием, а мутации свободно накапливаются в образовавшейся копии до тех пор, пока белок, кодируемый паралогом, не приобретет функцию, благоприятно сказывающуюся на каком-либо метаболическом пути. Это одна из первых моделей, позволившая в

упрощенном виде объяснить возникновение генов с новыми функциями. Однако ей присущ ряд недостатков. В частности, модель Оно опирается на представление о том, что один ген кодирует один белок с единственной функцией. Кроме того, согласно модели в силу нейтральной эволюции в гене накапливаются запрещенные мутации с такой же вероятностью, как и разрешенные, что противоречит наблюдаемым закономерностям. Существуют также и другие события, которые не укладываются в рамки модели. В частности, в работе [77] были рассмотрены возможные случаи, когда дупликация гена может подвергаться положительному отбору в условиях увеличенной потребности в продукте гена.

2) *Дивергенция прежде дупликации.* Эта модель основывается на том, что многие белки способны проявлять несколько функций в той или иной степени, даже если они нисколько не свойственны основной функции этого белка. В этом случае дупликация позволяет данному гену-паралогу специализировать одну из функций, которая в ходе отбора становится основной. Таким образом, дупликация может подвергаться положительному отбору тогда, когда возрастает потребность во второстепенной функции белка, кодируемого данным геном.
3) *Субфункционализация.* Данная модель предполагает, что после дупликации запрещенные мутации могут накапливаться в обеих копиях паралогов, в этом случае оба гена подвергаются положительному отбору на функцию, представленную геном-предком. Чаще встречается у регуляторных элементов, но подходит и для белок-кодирующих последовательностей.

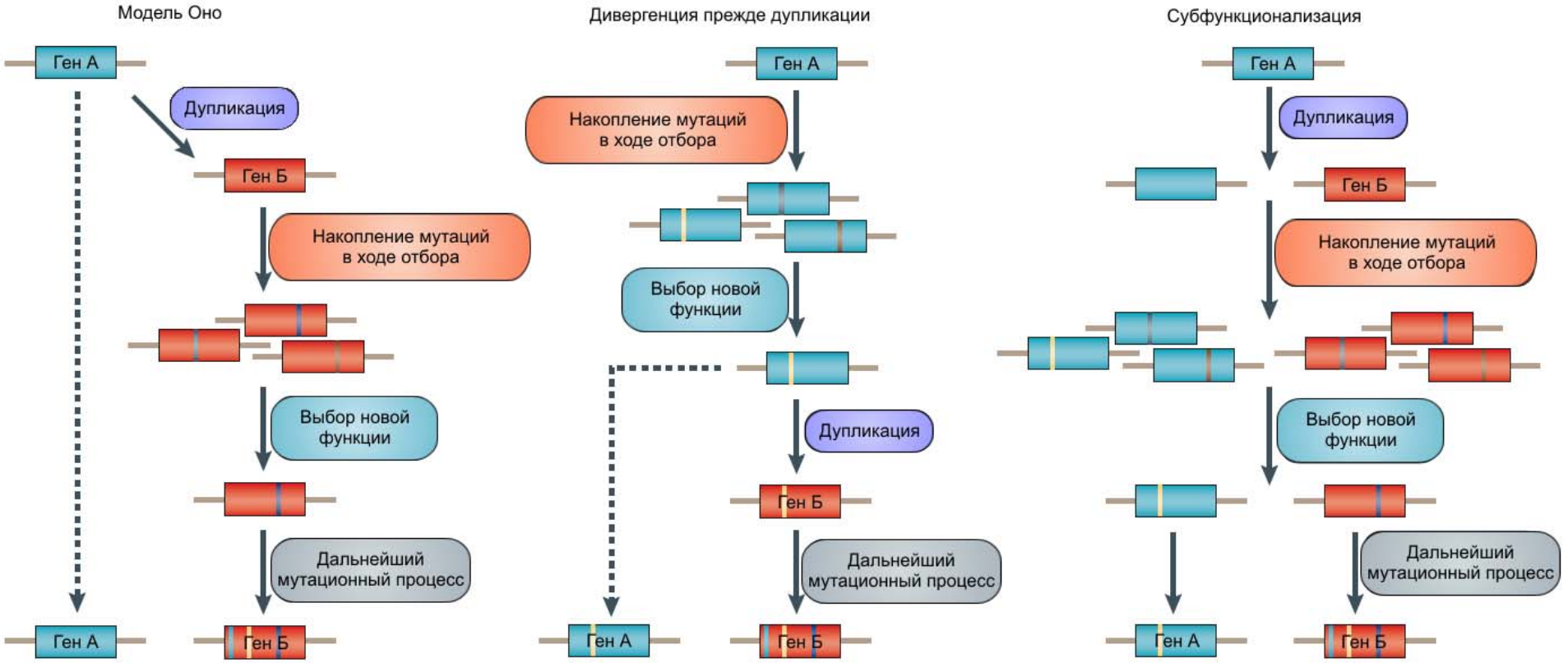


Рисунок 1.3. Основные модели дивергентной эволюции белков, в которую вовлечены дупликации кодирующего гена [61].

Перетасовка экзонов как еще один способ эволюции гена был предложен У. Гилбертом (Gilbert) в 1978. Под перетасовкой понимают слияние и разрыв экзонов, а также их дупликацию. Считается, что механизмом перетасовки экзонов является двойной неравный кроссинговер [56]. Частота кроссинговера обуславливается соотношением длины экзонов и интронов. Так как экзоны занимают 1% длины генома человека, а интроны 24% [78], то рекомбинация с большей вероятностью осуществляется между экзонами и затрагивает некодирующую часть генома, а не в экзонах. Было высказано предположение, что интроны – горячие точки рекомбинации [59]. Вследствие этого можно считать экзонную перетасовку одним из главных факторов эволюции белков. При этом показано, что случаи перетасовки экзонов происходили только на поздних ступенях эволюции, и потому процесс перетасовки не мог значительно влиять на возникновение древних белков. Считается, что процесс перетасовки экзонов возник после появления сплайсосомных интронов [79,80]. Помимо двойного неравного кроссинговера причиной перетасовки экзонов могут являться дупликации [61] и взаимодействия с мобильными элементами [60,61].

Чтобы можно было утверждать, что имел место случай перетасовки экзонов, необходимо, чтобы выполнялись два условия [80]:

1) два гомологичных участка последовательности присутствуют в окружении негомологичных участков;
2) перенос гомологичного участка осуществлялся с помощью рекомбинации между интронами.

Случаи перетасовки экзонов распространены среди многоклеточных организмов для генов, кодирующих внеклеточные белки и рецепторы. На основе этого факта была выдвинута гипотеза, что именно этот механизм эволюции способствовал развитию явления многоклеточности [74].

Вертикального переноса наследственного материала, то есть от предка, недостаточно для описания эволюции эукариот, в частности, одноклеточных микроорганизмов. Причина этого кроется в том, что эукариотические геномы

химерны, так как возникли в результате как вертикального, так и **горизонтального переносов** [81]. Анализ доступных данных показал, что существует два типа горизонтального переноса: 1) горизонтальный перенос генетического материала из органелл в ядро, лишь некоторые события которого произошли на заре возникновения эукариотических геномов; 2) аномальные события горизонтального переноса, в которые вовлечены различные виды организмов как донора, так и реципиента.

Горизонтальный перенос между геномами прокариот и эукариот. Примером может служить встройка в геном *Drosophila melanogaster* большей части генома облигатного внутриклеточного симбионта *Wolbachia*[82]. Как правило, такие случаи горизонтального переноса связаны с эндосимбиозом или способности простейших к фаготрофии. Однако также известно существование горизонтального переноса и в геномах животных [83]. Известны случаи переноса генов и от эукариотического генома к прокариотическому [84] . Однако эти случаи намного более редки.

Горизонтальный перенос между геномами эукариот. Засвидетельствованы несколько случаев горизонтального переноса генов отвечающих за патогенность или осмотрофный тип питания между различными видами грибов от патогенного к непатогенному. Это привело к конвергентной адаптации непатогенного вида гриба к патогенности [85].

1.3.2 Частота использования кодонов в последовательностях ДНК

Оптимальными кодонами принято считать те кодоны, которые соответствуют наиболее распространенным тРНК и трансляция которых соответственно проходит быстрее. Среди кодонов выделяют часто встречающиеся и редко встречающиеся. На использование кодонов влияют такие факторы, как инициация трансляции, сайты сплайсинга, нуклеосомный контекст.

Известно, что использование кодонов вблизи границ экзонов и интронов отличается, так как кодоны в этих участках ДНК несут сигнал сплайсинга.

Посчитаны консенсусы для донорного и акцепторного сайтов сплайсинга (рис. 1.4) [86].

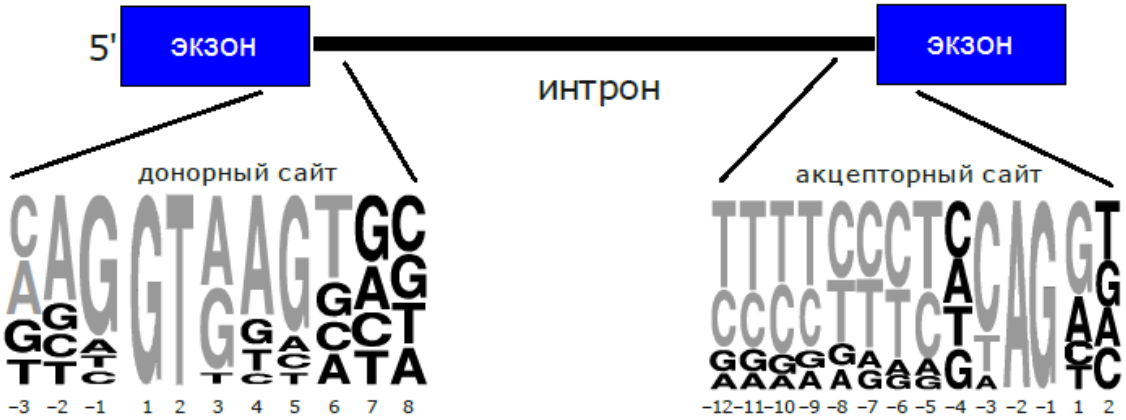


Рисунок 1.4. Консенсус последовательности для донорного и акцепторного сайтов сплайсинга отображен с помощью весовых матриц PSM.

Дальнейшие исследования показали, что хотя последовательность насыщена динуклеотидами AG в области донорного сайта сплайсинга, кодоны встречаются, как правило, AT-богатые [87].

Последние исследования на различных организмах показывают, что использование кодонов связано и со структурно-функциональными особенностями кодируемых белков. Так, можно выделить следующие факторы [88]:

1) кодоны во фрагментах ДНК, кодирующих аминокислоты, располагающиеся в гидрофобном ядре пространственной структуры белка, более консервативны и чаще оптимальны, чем те, которые кодируют аминокислоты, расположенные на его поверхности;

2) использование оптимальных кодонов во фрагментах ДНК, кодирующих аминокислоты, расположенных внутри элемента вторичной структуры белка, не коррелирует со скоростью эволюции последовательности;

3) слабоструктурированные участки пространственной структуры менее консервативны, чем элементы вторичной структуры;

4) наблюдается корреляция между скоростью экспрессии гена, его скоростью эволюции и термостабильностью кодируемого белка;

5) участки белок-белковых взаимодействий подвергаются меньшей скорости эволюции, чем другие на поверхности пространственной структуры белка.

Оптимальные и часто встречающиеся кодоны могут отличаться у разных видов, при этом различается и интенсивность их использования [89].

1.3.3 Эволюция пространственной структуры белка: конвергенция и дивергенция

Пространственная укладка аминокислотной цепи белка подчиняется ряду закономерностей [23]:

1. белок сворачивается в уникальную трехмерную структуру;
2. пространственная укладка последовательности белка избирательна, т.е. произвольно выбранный полипептид не обязательно образует известную укладку;
3. белки, взятые из протеомов различных организмов и имеющие сходство последовательностей 25-30%, имеют одинаковую третичную структуру и принадлежат к одному семейству укладок
4. некоторые пары белков, имеющие одинаковую укладку, имеют такое же сходство последовательностей, какое могут иметь произвольные последовательности – 8-9%
5. внутри одного семейства укладок белковые последовательности имеют только 3-4% одинаковых аминокислот на совпадающих позициях.

На формирование и сохранение пространственной структуры белка влияет жесткий стабилизирующий отбор, так как белок стремиться приспособиться к наиболее стабильной и консервативной укладке. Последние же данные показывают, что белковая укладка не обязана быть ни физически, ни биологически инвариантной [90].

Небольшое количество аминокислот, одинаковых и располагающихся в сходных участках пространственной структуры белков с одинаковой укладкой, относится к числу наиболее функционально важных. Так, показано

сходство между функциональными сайтами разных белков для укладки класса α/β, таких как фосфорибозилантранилат изомераза, индолглицеролфосфат синтаза, рибулозо-бисфосфаткарбоксилаза и др. [3].

Было замечено, что существуют участки третичной структуры белка (петли и другие периферические структуры), являющиеся структурно подвижными. Такие белки были названы хамелеонами, а подвижные части вариабельными. Как правило, расположение этих участков в структуре изменяется при изменении функционального состояния белка (например, при связывании с лигандом) [90].

Белки сгруппированы в различные семейства, объединяемые общей укладкой, на основе гомологии и сходства пространственной структуры [91]. При определении белкового семейства используются различные пороги для определения сходства последовательности и разнообразные методы подсчета. Широко применяется порог в 30%. В частности, при составлении базы данных SCOP (Structural Classification of Proteins) [16], используется два критерия кластеризации белков: 1) аминокислотные последовательности, имеющие сходство более 30%; 2) белки, обладающие низким уровнем сходства последовательности, но близкие по функциям и структуре (например, глобины с уровнем сходства 15%).

1.4 Проекция пространственной структуры белка на структуру кодирующего гена

Экспериментальные методы позволили идентифицировать не только структуры белков, но структурные домены, входящие в их состав, т.е. минимальные по длине фрагменты последовательности, обладающие способностью к самостоятельной сборке в глобулярную пространственную структуру, идентичную структуре этого фрагмента в составе полноразмерного белка. В настоящее время существует множество свидетельств того, что структурные домены также могут сохранять функциональные свойства, которые им были присущи в составе

полноразмерных белков [2]. С появлением данных о пространственной структуре белков возникли вопросы о том, как же соотносятся между собой структурная организация беков и структура кодирующих их генов.

1.4.1 Соответствие доменной структуры белка и экзонной структуры кодирующего гена

Соответствия между экзоном и доменом могут быть классифицированы следующим образом [2]:

1. один экзон – один домен
2. несколько экзонов – один домен
3. один экзон – несколько доменов
4. нет точного соответствия (например, домен составлен из частей экзонов).

Показано, что после перетасовки кодирующая экзонная структура домена белка может изменяться за счет вставок или делеций интронов [80].

Для мультидоменных белков была рассмотрена зависимость между сложностью организма и количеством комбинаций с доменами других суперсемейств. В этом случае суперсемейства были эквивалентны суперсемействам базы данных SCOP. Как видно из таблицы 1.4, количество таких суперсемейств увеличивается со сложностью организма, также увеличивается и количество белков, содержащих в своем составе различные комбинации доменов [92]. Существование комбинаций доменов может обуславливаться потребностью в многофункциональных белках. Одним из лидеров суперсемейств по количеству различных доменов является EGF. Интересно отметить, что происхождение этого суперсемейства связывают с перетасовкой генов [93].

В литературе, наравне с доменом, также выделяют супра-домен, как эволюционную единицу, т.е. такую парную комбинацию доменов из различных суперсемейств, которая подвергалась дупликации в пределах некоего гена [94].

Таблица 1.4. Распределение комбинированных друг с другом суперсемейств по геномам [92].

Геномы	% суперсемейств, скомбинированных с одним суперсемейством	% суперсемейств, скомбинированных с двумя суперсемействами	% суперсемейств, скомбинированных с тремя суперсемействами
Archaeoglobus fulgidus	38	7	5
Methanobacterium thermoautotrophicum	40	8	5
Bacillus subtilis	45	8	4
Escherichia coli	43	10	7
Saccharomyces cerevisiae	38	5	4
Caenorharbditis elegans	35	10	11
Drosophila melanogaster	36	9	11

Считается, что на перетасовку доменов оказывает влияние их размер. Полагают, что более крупные домены являются менее мобильными по сравнению с меньшими по размеру [95]. В частности, это согласуется с гипотезой о том, что вставки дополнительных интронов в экзоны способствуют осуществлению негомологичной рекомбинации [95].

1.4.2 Фазы экзонов и интронов и их роль в эволюции

Для анализа корреляции структур экзона и домена используется сопоставление границ экзона в гене и проекции домена кодируемого белка на нуклеотидной последовательности. Было выявлено, что существуют различные комбинации по фазе интронов и экзонов, т.е. по номеру нуклеотида в кодоне согласно рамке считывания. Фаза интронов определяется по месту разрыва кодирующего кодона интроном: 0 – если разрыв находится на границе триплетов, 1 – если сдвиг на 1 позицию, 2 – если сдвиг на два нуклеотида (рис. 1.5). У одного экзона можно выделяют две фазы, так как он ограничен двумя кодирующими кодонами, участвующими в разрыве на 5’-конце и 3’-конце.

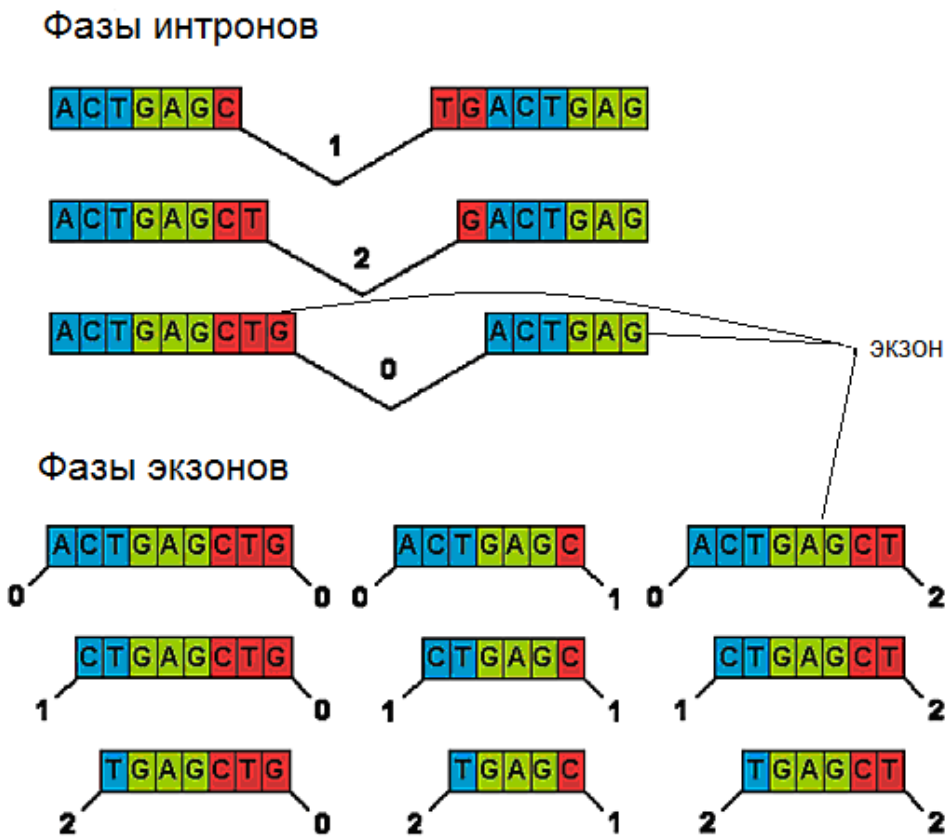


Рисунок 1.5. Фазы интронов и экзонов соответственно границам вставки [78]. Красным показан кодон, по которому проходит вставка интрона между экзонами.

Частота статистически значимых совпадений границ проекции домена и экзона увеличивается в ходе эволюции от червей и насекомых к рыбам и млекопитающим, что отражает постепенно увеличивающуюся корреляцию экзон-домен. Огромное количество данных по этой корреляции дает повод для некоторых предположений и вопросов, увеличивающих интерес к экзонной теории перекомбинации. Например, способствует ли экзон-доменная корреляция распространению доменов и облегчению распределения доменов внутри генома? Может ли эта мобильность домена быть движущей силой формирования мозаичных белков?

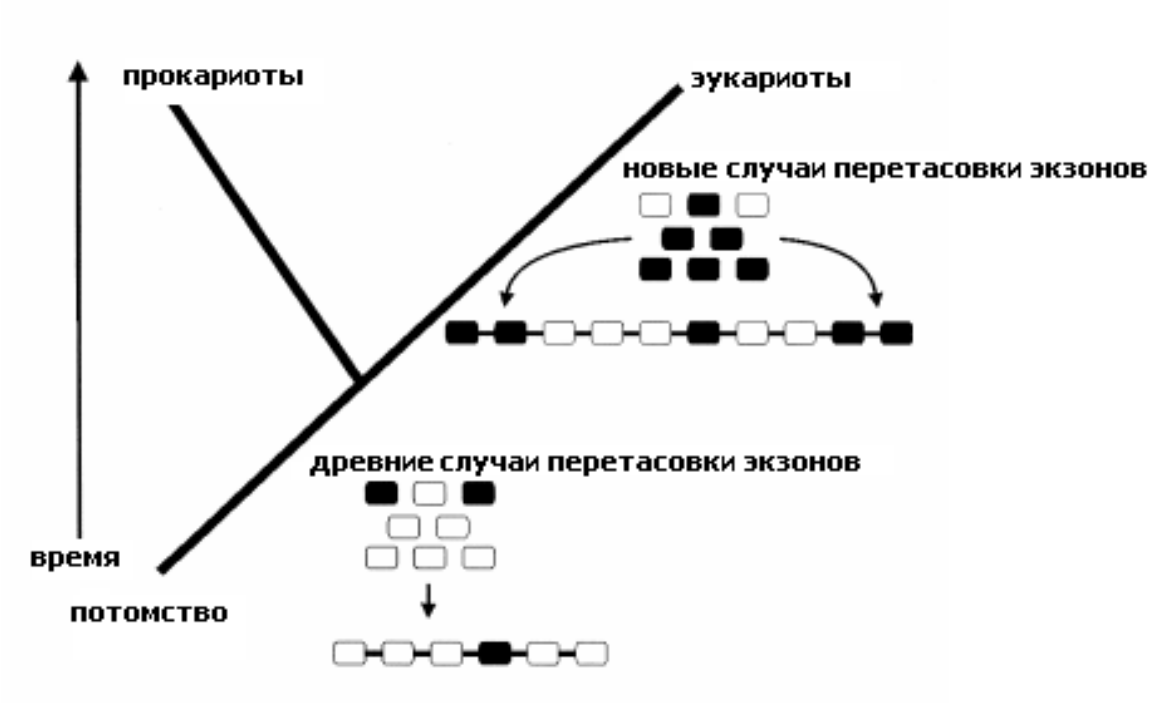


Рисунок 1.6. Эволюция прокариот и эукариот с точки зрения интрон-экзонных границ. Интроны показаны черной соединительной линией. Экзоны с границами 0-0 – белыми квадратами, с границами 1-1 – черными [96].

Группой авторов [97] исследовался кластер доменов, чьи границы расположены очень близко к экзон-интронным границам. С использованием геномов девяти видов животных, было показано, что эти ограниченные одним экзоном домены демонстрируют высокую подвижность в пределах генома. Экзон-ограниченные домены в среднем более многочисленны и присутствуют в большем количестве генов, чем другие домены.

Например, у *C. elegans* 19% всех доменов совпадают с границами экзонов, тогда как у человека таких доменов 42% [97]. Если в мультидоменных белках имеются повторяющиеся домены, процент совпадения границ экзонов и доменов увеличивается до 50%. Эти и другие данные говорят о постепенном усложнении генома и об эволюционном прогрессе, связанном с перетасовкой и дупликациями [97].

Правило вырезания рамки для экзонной рекомбинации обуславливает участие в перетасовках доменов, ограниченных одной и той же фазой интронов. Это может быть одной из причин, по которой интрон-экзонные границы начала и конца экзона часто совпадают по нуклеотидной позиции в

триплете. При этом у древних видов чаще встречается позиция 0-0, а у новых – 1-1 (рис. 1.6, 1.7) [2].

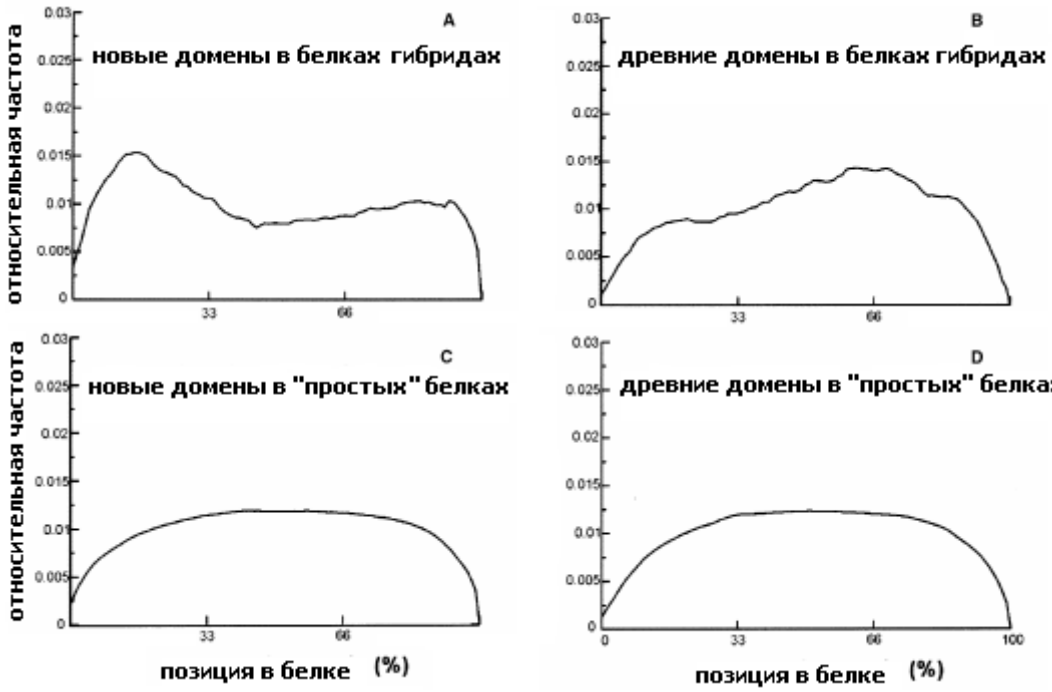


Рисунок 1.7. Распределение древних и новых доменов в зависимости от позиции в белке [91]. Гибридный белок содержит и древние, и новые домены. Простой белок содержит домены только одного типа. ***A.*** *Новые домены в гибридных белках.* ***B.*** *Старые домены в гибридных белках.* ***C.*** *Новые домены в простых белках.* ***D.*** *Гибридные домены в простых белках.*

Подобная корреляция является строгой для интрон-экзонной структуры некоторых генов позвоночных (протеазы коагуляции крови, фибринолиза, белки комплемента, различные селектины, белок соединения хряща, фактор H, тенацины [78]). В то же время существуют группы генов, для которых корреляции границ экзонов и доменов не были выявлены. Например, у ламининов.

Авторами [96] проводился анализ корреляции границ проекции модуля белка на нуклеотидную последовательность гена и границ экзона. Оказалось, что границы интронов коррелируют именно с такими модулями, но только если оба находятся в фазе 0. На этом фоне было выдвинуто предположение, что перетасовка древних генов у прокариот осуществлялся по фазе 0,

поэтому интроны ими были потеряны или их не было вовсе, но рекомбинация все же происходила.

1.4.3 Интегрированные базы данных

За последние несколько лет были интегрированы многие базы данных, содержащие информацию по нуклеотидным, аминокислотным последовательностям, а также пространственным структурам белков. Так были созданы ресурсы XdomView [98], ExDom [99] и Structural Exon Database (SEDB) [100], осуществляющие разметку границ экзонов, а также доменов на пространственной структуре белка (Таб. 1.5).

XdomView – это веб-приложение для визуализации границ доменов белка на его третичной структуре. Для разметки доменной структуры извлекалась информация из баз данных SCOP, CATH, DALI [91], 3DEE [101], MMDB [102]. Для аминокислотной последовательности производилась разметка экзонной структуры кодирующего гена на основе БД ExInt [103]. Разметка экзонной структуры также производилась для гомологичных белков, поиск которых осуществлялся с помощью программы BLAST [104]. При этом информация не хранится в базе данных, а генерируется после запроса пользователя.

SEDB осуществляет поиск гомологичных заданным белкам последовательностей по GeneBank [105] с помощью BLAST, осуществляя структурное выравнивание между найденными последовательностями и последовательностью PDB по методу CE [106] и экстрагируя информацию об экзонно-интронной структуре из базы данные EID [107]. Кроме этого, программа позволяет проводить множественное выравнивание белков по соответствующим им границам экзонов и визуализировать фазы экзон-интронных границ.

ExDom сопоставляет доменную структуру белка и экзон-интронную структуру кодирующего его гена. В случае если для белка известна пространственная структура, на нее проецируют границы экзонов.

Таблица 1.5. Сравнительные характеристики ресурсов, посвященных проекции структуры белка на структуру кодирующего его гена.

Ресурс/Свойство	**XdomView**	**SEDB**	**ExDom**
Проекция структуры PDB	+	+	+
Проекция доменной структуры	+	+	+
Проекция доменов SCOP	+	-	-
Проекция экзонов на PDB структуру	+	-	-
Проекция функциональных сайтов	-	-	-
Взаимно-однозначное соответствие кодирующего гена и пространственной структуры	-	-	+
Многозначная проекция структуры PDB на гены	+	+	-

1.5 Заключение к литературному обзору

В литературном обзоре рассмотрены основные принципы строения, функционирования и организации белков. Приведены физико-химические свойства аминокислот в составе функциональных сайтов, классификация функциональных сайтов белков, в том числе по типу связываемых лигандов. Рассматривается эволюция структуры и функции белка в связи с эволюцией кодирующей структуры гена. Описываются причины наблюдаемой корреляции границ экзонной и доменной структур.

Сделан обзор существующих баз данных и веб-ресурсов по пространственной структуре белков и анализу ее особенностей, по методам и ресурсам, осуществляющим проекцию пространственной структуры белков на кодирующую структуру гена, а также баз данных, содержащих информацию по функциональным сайтам белков. По итогам такого обзора можно сделать вывод об отсутствии ресурса, который бы интегрировал различные и разрозненные данные о функциональных сайтах белка и их проекции на кодирующую структуру гена.

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 предоставлющего графическую визуализацию результатов.

2.1 Описание использованных баз данных

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). В разделе приводится описание форматов данных этих ресурсов.

2.1.1 Ensembl

Ресурс Ensembl посвящен хранению организованной биологической информации о последовательностях генов организмов, геном которых секвенирован полностью или почти полностью [108, 109]. При этом Ensembl посвящен информации преимущественно о геномах эукариот, в частности,

хордовых. Позднее в ресурс вошли еще 5 веб-сайтов, посвященных бактериальным геномам (Ensembl Bacteria), геномам простейших (Ensembl Protista), грибов (Ensembl Fungi), растений (Ensembl Plants) и животных (Ensembl Metazoa). В основе Ensembl лежит автоматическая аннотация известных генов и предсказание новых генов на основе функциональной аннотации InterPro [110], информации о болезнях, связанных с мутациями OMIM [111], данных по экспрессии белков на основе метода SAGE (Serial analysis of gene expression) [112] и информации о семействе генов. Ensembl также содержит информацию о генах, предсказанных по гомологии и по методу скрытых марковских моделей. Геномный браузер позволяет просматривать гены на протяжении всей длины хромосомы. Помимо этого, хранится информация о генетических маркерах, генах сцепленных с заболеваниями, об однонуклеотидных полиморфизмах, CpG-островах, повторах, сравнительном анализе генов. Транскрипты основываются на курируемых базах данных UniProt/Swiss-Prot и NCBI RefSeq, а также UniProt/TrEMBL [113]. Таким образом, транскрипты в базе данных Ensembl могут быть, известными (known), предсказанными (novel) и смешенного типа (merged).

Идентификаторы Ensembl организованы следующим образом:

- ENSG### ген
- ENST### транскрипт
- ENSP### белок
- ENSE### экзон

Для генов других организмов к идентификаторам добавляются первые буквы латинского алфавита: ENSDA**R**T### (*Da**r**io **r**erio*).

В случае альтернативного сплайсинга для одного гена в Ensembl может содержаться информация о транслируемых и нетранслируемых транскриптах. Для каждого транскрипта имеется полная последовательность, включающая экзон-интронную разметку. Также для гена представлено попарное выравнивание с последовательностями известных ортологов и

паралогов. Страница, описывающая белок, включает информацию о последовательности белка, о кодирующей экзон-интронной структуре и о доменной структуре на основе таких баз данных, как Pfam, Prosite, InterPro.

2.1.2 Protein Data Bank (PDB)

Общая информация об этом банке данных описывалась выше (см. 1.1.3). Поскольку это база данных трехмерных структур белков, то основную часть файла в формате PDB занимает описание пространственных координат атомов основной цепи и аминокислотных остатков. Помимо этого, файл содержит информацию о наименовании молекулы, первичной и вторичной структурах белка, о лигандах и комплексах, ссылки на другие базы данных, содержащие информацию о последовательности белка, библиографические ссылки и подробности проведения эксперимента.

Файл базы данных PDB – форматированный текстовый файл, в котором каждая строка начинается с названия поля. Название поля имеет длину до 6 латинских символов. Есть поля, присутствующие в файле в обязательном порядке (таб. 2.1): краткий заголовок (HEADER), наименование (TITLE), характеристика молекул (COMPND), организм (SOURCE), ключевые слова (KEYWDS), информация о характере эксперимента при получении пространственной структуры (EXPDTA), авторы эксперимента(AUTHOR), изменения в файле с момента первого опубликования (REVDAT), публикации и разрешение структуры (REMARK 2 и 3 соответственно), первичная структура (SEQRES), конец файла (END). Остальные поля являются опциональными.

Данные, которые заносятся в каждое поле, строго регламентированы. Информация о функциональном сайте белка хранится в полях:

- REMARK 800, которое предоставляет информацию об идентификаторе сайта; способе распознавания сайта (экспертное или с помощью компьютерных программ); трехбуквенном идентификаторе лиганда
- HETNAM – содержит расшифровку трехбуквенного идентификатора

- SITE – содержит информацию о положении аминокислот в последовательности белка, для которой известна пространственная структура

Таблица 2.1. Группы данных, представленных в файле формата PDB

Описание	HEADER, OBSLTE, TITLE, SPLIT, CAVEAT, COMPND, SOURCE, KEYWDS,EXPDTA, NUMMDL, MDLTYP, AUTHOR, REVDAT, SPRSDE, JRNL
Комментарии	REMARK 0-999
Последовательность	DBREF, SEQADV, SEQRES MODRES
Дополнительные молекулы	HET, HETNAM, HETSYN, FORMUL
Вторичная структура	HELIX, SHEET
Химические связи	SSBOND, LINK, CISPEP
Различные сайты	SITE
Кристаллографические данные	CRYST1
Система координат	ORIGXn, SCALEn, MTRIXn
Координаты атомов	MODEL, ATOM, ANISOU, TER, HETATM, ENDMDL
Координаты связанных атомов	CONECT
Маркеры файла	MASTER, END

2.1.3 SCOP

Согласно SCOP, структуры белков объединены в семейства, структурные суперсемейства, укладки и классы [16].

Классы – самая старшая единица иерархии – представлены восемью группами структур: 1) состоящие только из альфа-спиралей, 2) только из бета-листов, 3) состоящие из чередующихся альфа и бета структур, 4) состоящие из разделенных альфа и бета структур, 5) мультидоменные белки, 6) мембранные белки и белки поверхности клетки, 7) малые белки, 8) суперзакрученные структуры (coil-coiled).

Структурная укладка объединяет домены, имеющие одинаковые элементы вторичной струкутры в одинаковой топологической ориентации.

Надсемейство объединяет домены, имеющие помимо структурного сходства общее эволюционное происхождение.

Структурное семейство – самая младшая единица иерархии - объединяет домены по двум признакам: 1) значительное сходство по последовательности; 2) функциональное сходство (рис. 2.1).

```
a.1.2.3
| | | |
| | | семейство
| | надсемейство
| укладка
класс
```

Рисунок 2.1. Идентификатор структурного семейства в SCOP.

Поскольку структурная классификация белков основывается на пространственных структурах, описанных в формате PDB, то координаты атомов и сама последовательность белка, для которой известна пространственная структура и структурная укладка, совпадают. Каждый структурный домен в БД SCOP имеет стандартный идентификатор:

```
d(идентификатор pdb)(цепь полипептида в pdb).
```

Структурный домен может полностью включать в себя одну цепь белка. В то же время на одной цепи белка может лежать несколько структурных доменов.

2.2 Описание программных средств

При создании системы были использованы программные средства для сравнения последовательностей и пространственных структур.

2.2.1 Формат данных FASTA

Формат FASTA применяется для хранения нуклеотидной или аминокислотной последовательности. Первая строка, начинаясь с «>», содержит описание последовательности, как правило, в следующем порядке:

- наименование последовательности (обязательно)
- тип последовательности

- длина последовательности

Во второй строке записана последовательность, как правило, разбитая на строки по 60-80 символов в каждой. Пустые строки могут как игнорироваться, так и считаться окончанием последовательности. Цифры, дефисы, пробелы и т.п. обычно игнорируются. Файл может содержать несколько последовательностей.

2.2.2 BLAST

BLAST (Basic Local Alignment System Tool) – программа для сравнения последовательностей с помощью оптимизированного алгоритма локального выравнивания и матриц замен [114]. Она позволяет осуществлять поиск нуклеотидных и аминокислотных последовательностей, имеющих высокое семантическое сходство, по заданной базе последовательности. Результаты поиска упорядочены по очкам (Score) и статистической значимости результата (E-value). Значимым обычно считается результат со значением E-value менее 0,05.

Программа позволяет использовать как стандартные БД (UniProt, PDB, GenBank, RefSeq etc.), так и базу собственных последовательностей в FASTA формате. В первом случае формат FASTA включает в начале строки “gi|”, вот втором идентификатор считается локальным и вначале строки используется “lcl|”. При этом заголовок последовательности должен быть уникальным, а через пробел может включаться любая информация в произвольном формате. Программа находится в свободном доступе и может быть установлена на сервер (таблица 2.2).

Таблица 2.2. Программы, включенные в пакет BLAST

Программа	Анализируемая последовательность	База последовательностей
blastn	нуклеотидная	нуклеотидная
blastp	аминокислотная	аминокислотная
blastx	аминокислотная	транслированная нуклеотидная
tblastn	транслированная нуклеотидная	аминокислотная
tblastx	транслированная нуклеотидная	транслированная нуклеотидная

BLAST позволяет установить различные параметры поиска гомологичных последовательностей, которые зависят, например, от длины последовательности (таблица 2.3). При выравнивании аминокислотных последовательностей показателем количественной оценки качества выравнивания является его вес.

Таблица 2.3. Рекомендуемое использование в BLAST матриц аминокислотных замен при различной длине полипептида [116].

Длина полипептида	>85 aa	50-85 aa	35-50 aa	<35 aa
E-value	10	10	1000	1000
Матрица	BLOSUM 62	BLOSUM 80	PAM 70	PAM 30

Аминокислоты с близкими биохимическими свойствами, такими как заряд, полярность и т.д. характеризуются большей вероятностью парных замен. Некоторые аминокислоты, например цистеин, глицин, триптофан очень редко заменяются в процессе эволюции. Для того чтобы учесть неравную вероятность замен были разработаны специальные матрицы, которые получили название матрицы замен. Эти матрицы содержат оценки частных весов для любой пары замены аминокислоты. Первыми были матрицы аминокислотных замен PAM, учитывающие эволюционное расстояние между последовательностями. Другим широко используемым семейством матриц весов являются матрицы BLOSUM. Они построены на основе выравниваний последовательностей с определенной степенью сходства [115].

2.2.3 ClustalW

Программа ClustalW [117] используется для множественного выравнивания последовательностей ДНК и белков. Имеет два возможных применения:

1) Множественное выравнивание последовательностей. Входной формат данных – NBRF/PIR, EMBL/UniProt, Pearson (FASTA), GDE,

ALN/ClustalW, GCG/MSF, RSF, формат выходных данных – ALN, GCG/MSF, PHYLIP, PIR, GDE. Наиболее часто используемый формат входных данных – FASTA, формат выходных данных - ALN. Формат ALN – это формат выравненных последовательностей. Файл начинается со слова «CLUSTAL» и информации об используемой версии программы. Выравнивание пишется блоками по 60 символов последовательности в каждой строке. Каждый блок начинается с наименования последовательности согласно имени, заданном в FASTA формате, а в конце строки пишется полное количество отображенных букв последовательности. Информация о совпадении последовательности отмечается тремя символами: "*" – символы последовательности идентичны; ":" – наблюдаются консервативные замены; "." – наблюдаются полуконсервативные замены. Если при выравнивании обнаруживаются пропуски нуклеотидов или аминокислот, то в последовательности они обозначаются "-" (рис. 2.2).

```
CLUSTAL W 2.1 multiple sequence alignment


FOSB_MOUSE      ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS 60
FOSB_HUMAN      ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPVVDPYDMPGTSYSTPGMSGYSSGGASGS 60
                ********************************.***************:*.**:****** 
```

Рис 2.2. Пример выравнивания с помощью программы CLUSTALW в формате ALN.

2) Построение филогенетического дерева. Построение дерева происходит на основе данных в формате PIR или PHYLIP. Дерево может быть построено на основе кластеризации несколькими способами: методом ближайшего соседа (NJ) [118], или методом невзвешенного попарного среднего (UPGMA) [119]. При построении может учитываться коррекция расстояний Кимуры (учитывает, что замена нуклеотидов или аминокислот между последовательностями может не быть единичной) и использование игнорирования пропусков в выравнивании.

2.2.4 3DPDBScan

Для возможности поиска пространственного сходства между полипептидами в Институте Цитологии и Генетики СО РАН разработана программа 3DPDBScan. Программа использует файлы в формате PDB. PDB3DScan основана на алгоритме SSM [120]. 3D структура белка представляется в виде элементов вторичной структуры (альфа-спиралей и бета-листов). Это позволяет проводить быстрое сравнение 3D структуры белка с базой данных 3D структур, описанной выше (раздел 1.1.3). Такое представление 3D структуры накладывает ограничения на длину полипептидных последовательностей. Полипептиды, вторичная структура которых состоит менее чем из двух элементов, игнорируются программой.

Основная проблема включения коротких фрагментов белков при 3D сравнении заключается в установлении критериев, позволяющих избежать избыточного количества выравниваний. Например, структурные выравнивания α-спиралей или β-стрендов разных белков в большинстве случаев будут иметь низкое значение RMSD (квадратный корень из минимального значения среднего по квадратам расстояний между соответствующими атомами двух молекул).

Для поиска структурных аналогов между короткими последовательностями разработан метод в рамках программы PDBSiteScan [121], основанный на сравнении заданного полипептида с набором структурных шаблонов. При этом сначала сравнивается расположение атомов основной цепи (N, Ca и C), а затем подбирается структурно сходный аминокислотный остаток.

2.3 Алгоритм создания БД SitEx

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, приведенных в данных записях, рассчитываемое с помощью глобального парного выравнивания с применением программы CLUSTALW. В случае нескольких найденных последовательностей белка в Ensembl (в случае альтернативного сплайсинга) взаимооднозначное соответствие устанавливалось с той последовательностью, для которой было найдено максимальное сходство. 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.

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) (рис. 2.3). Все программы для интеграции данных написаны на языке Perl и языка запросов MySQL.

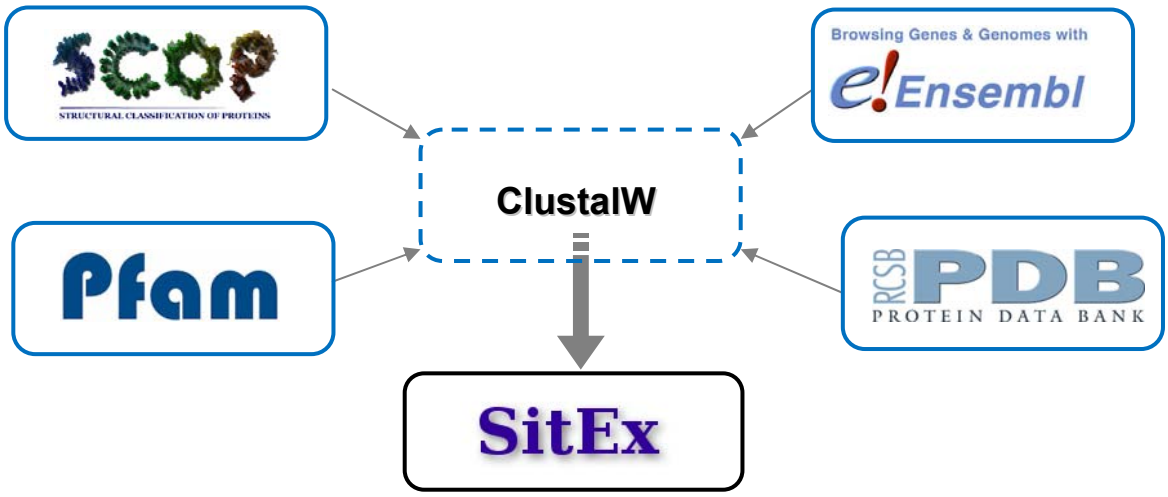


Рисунок 2.3. Схема интегрирования данных на основе компьютерных программ и баз данных, необходимых для разработки SitEx.

2.4 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.

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. 2.4). 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.

Рисунок 2.4. Пример значений коэффициента разрывности сайта по экзонам.

2.5 Описание структуры базы данных SitEx

БД SitEx является реляционной базой данных, для создания которой использовалась система управления MySQL. Структура БД представлена на рис. 2.5.

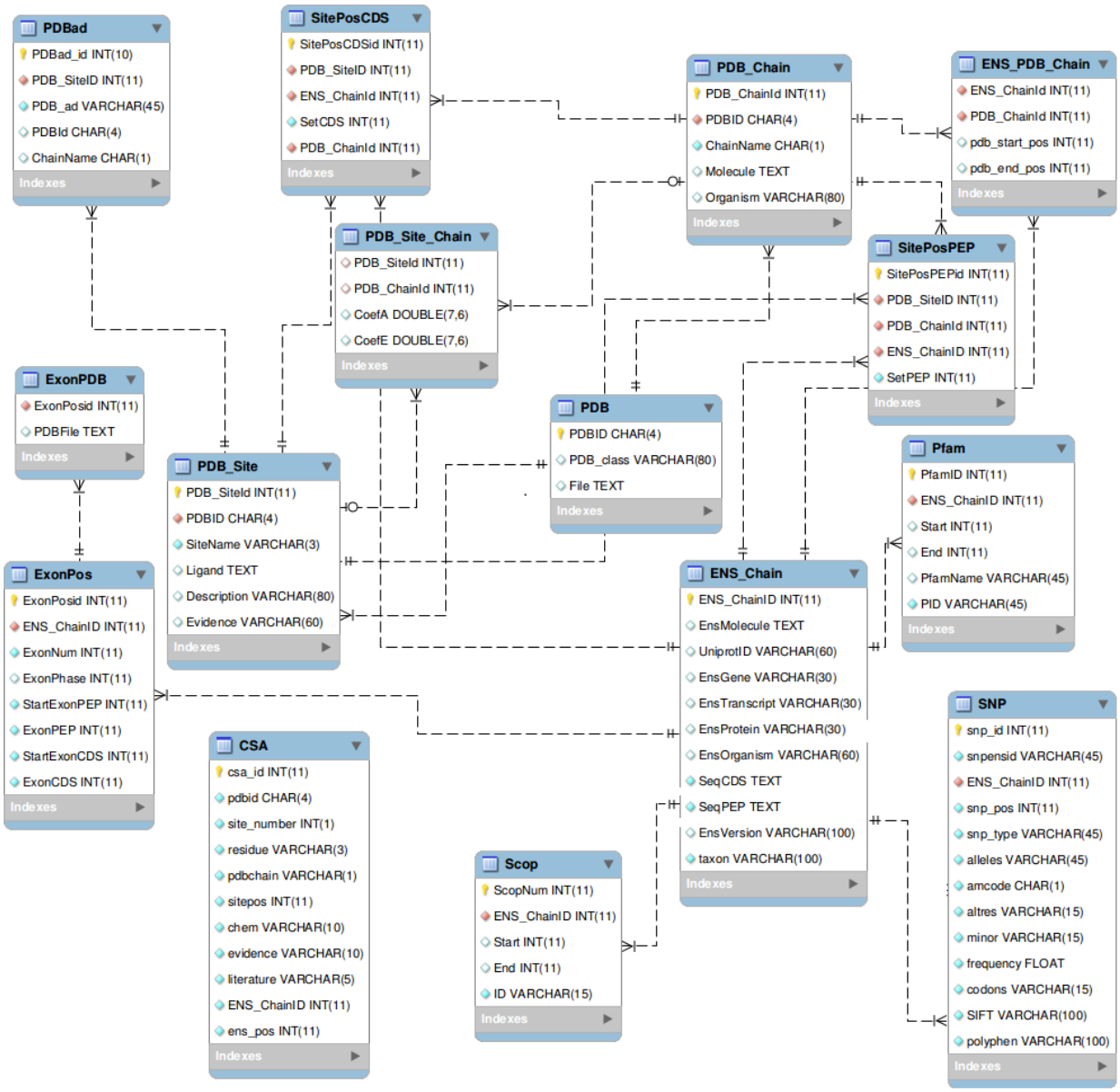


Рисунок 2.5. Схема структуры базы данных SitEx. Заголовки соответствуют названиям таблиц. Желтым цветом отмечены ключевые поля таблиц. Красным – поля, присутствующие в других таблицах. Рядом с наименованием поля указан его тип. INT(11) – численное значение; VARCHAR – строковое значение; TEXT – текстовое значение.

Вся описательная информация о функциональном сайте хранится в таблице PDB_Site. Информация об идентификаторах Ensembl, кодирующей и аминокислотной последовательностях хранится в таблице ENS_Chain. Описание белка согласно PDB внесено в таблицу PDB_Chain. Последние две таблицы связаны между собой вспомогательной таблицей ENS_PDB_Chain. Поскольку функциональный сайт может принадлежать нескольким белковым цепям в структуре, представленным PDB, то связь между сайтом и цепями

белка описана в таблице PDB_Site_Chain. Позиции аминокислот функционального сайта на последовательности белка согласно Ensembl хранятся в таблице SitePosPEP, а позиции нуклеотидов, входящих в кодоны, кодирующие функциональный сайт белка, хранятся в таблице SitePosCDS. В этих же таблицах идентификатор функционального сайта из таблицы PDB_Site связан с идентификатром белка из таблицы ENS_Chain.

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.

2.6 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 (рис. 2.6, рис. 2.7).

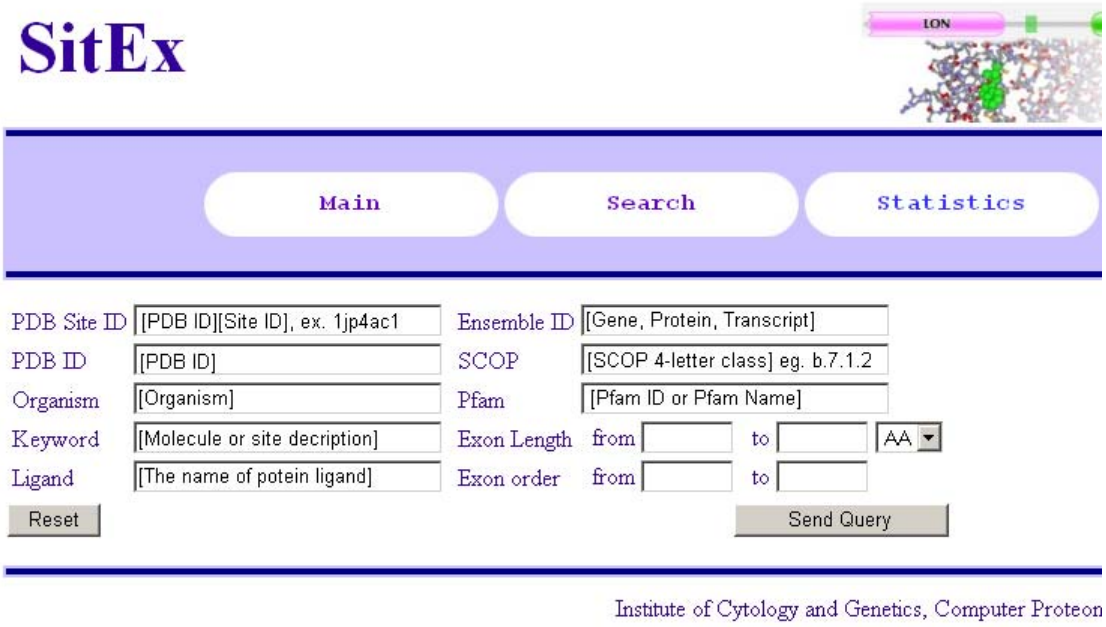


Рисунок 2.6. Страница поискового запроса.

SiteID	PDB	Site name	Organism	Ligand
2B6FAC1	2B6F	AC1	HOMO SAPIENS	ATP ADENOSINE-5'-TRIPHOSPHATE

ExonID	EnsTranscript	EnsGene	Length
Exon 1	ENST00000381962	ENSG00000172070	70
Exon 2	ENST00000381962	ENSG00000172070	68

SiteID	PDB	Site name	Organism	Ligand
3BJUBC5	3BJU	BC5	HOMO SAPIENS	ATP ADENOSINE-5'-TRIPHOSPHATE

ExonID	EnsTranscript	EnsGene	Length
Exon 1	ENST00000319410	ENSG00000065427	49
Exon 2	ENST00000319410	ENSG00000065427	53
Exon 3	ENST00000319410	ENSG00000065427	56
Exon 4	ENST00000319410	ENSG00000065427	31
Exon 5	ENST00000319410	ENSG00000065427	62

Рисунок 2.7. Результат запроса по ключевому слову "ATP" в строке "Ligand". Отображается функциональный сайт, информация о последовательности, список экзонов, которые кодируют последовательность и их длина, измеренная в кодонах.

Страница описания экзона 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 (рис. 2.8).

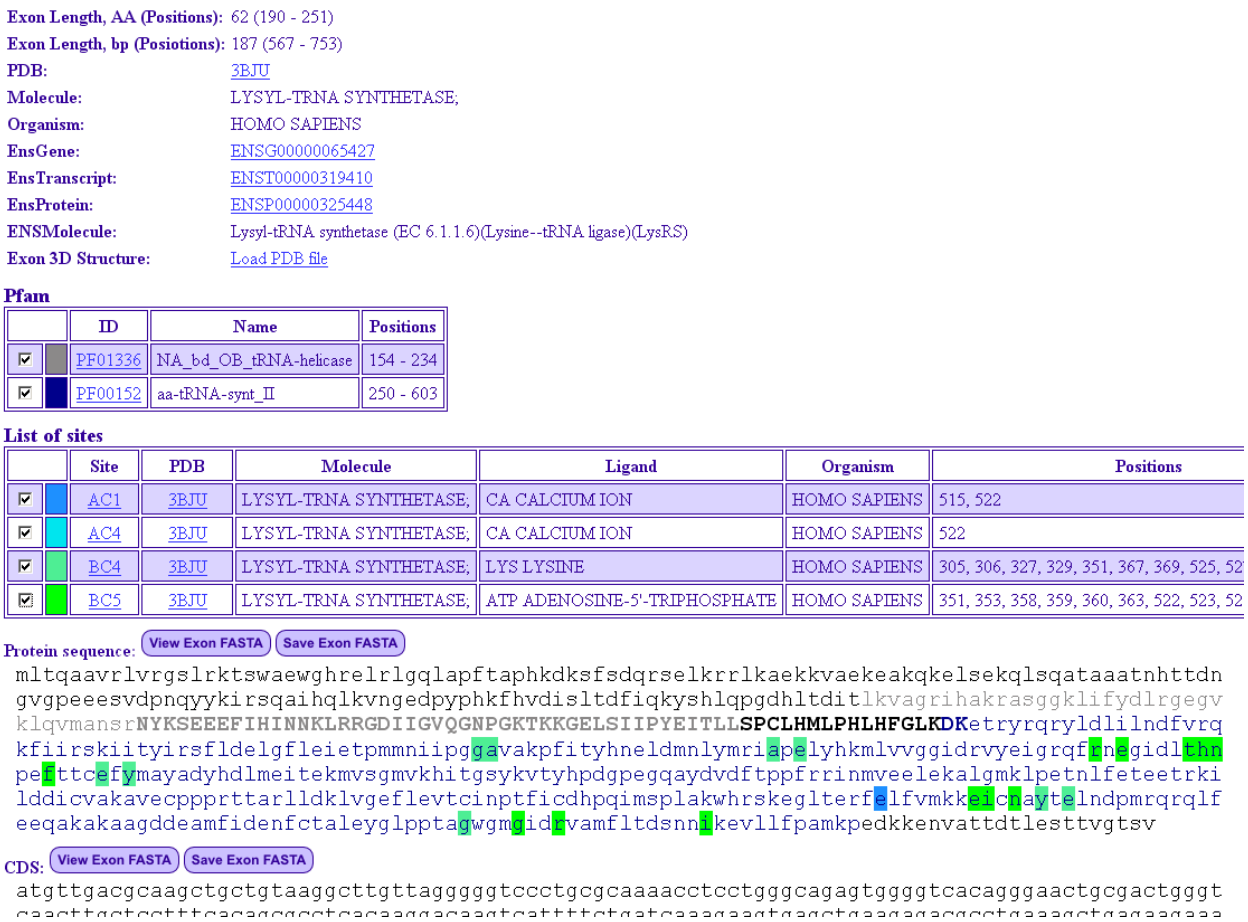


Рисунок 2.8. Страница описания экзона. Представлен блок описания последовательности, блок для разметки доменов на последовательности, а также блок для разметки аминокислот функциональных сайтов на последовательности экзона либо полипептиде, кодируемом им.

Страница описания 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 (рис. 2.9).

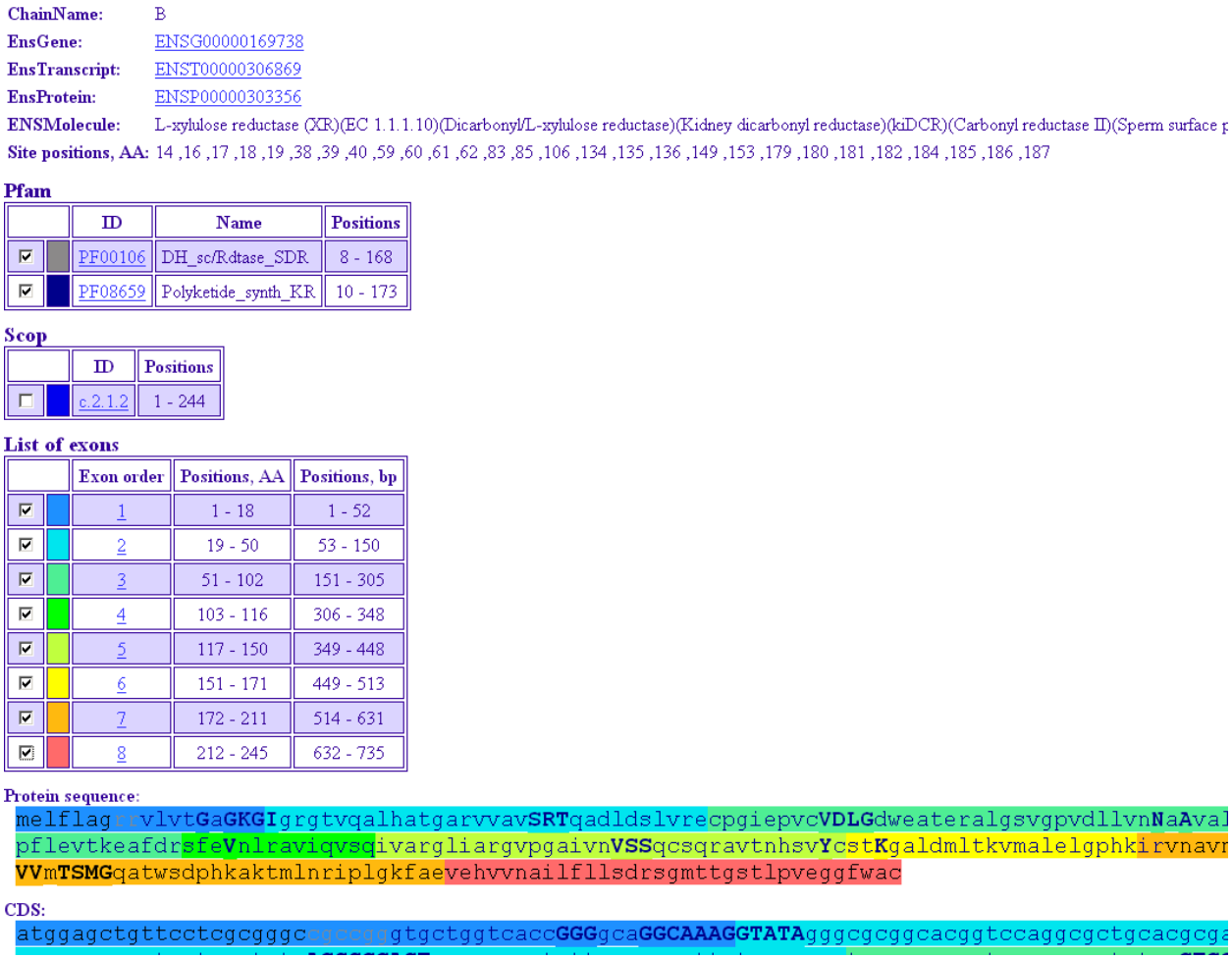


Рисунок 2.9. Фрагмент страницы описания функционального сайта. Представлен блок описания последовательности, блок для разметки доменов на последовательности, а также блок для разметки экзонов на последовательности, в которой размечены аминокислоты функционального сайта.

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]

[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 (Таблица 2.4);
- классификация сайтов (Таблица 2.5).

Таблица 2.4. 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, участвующие в репликации, транскрипции, трансляции	73
Белки теплового шока	22
Транспортные белки	313
Белки опухолей	161
«Цинковые пальцы», RING пальцы	177
Другие белки и предшественники	614

Таблица 2.5. Классификация лигандов в SitEx

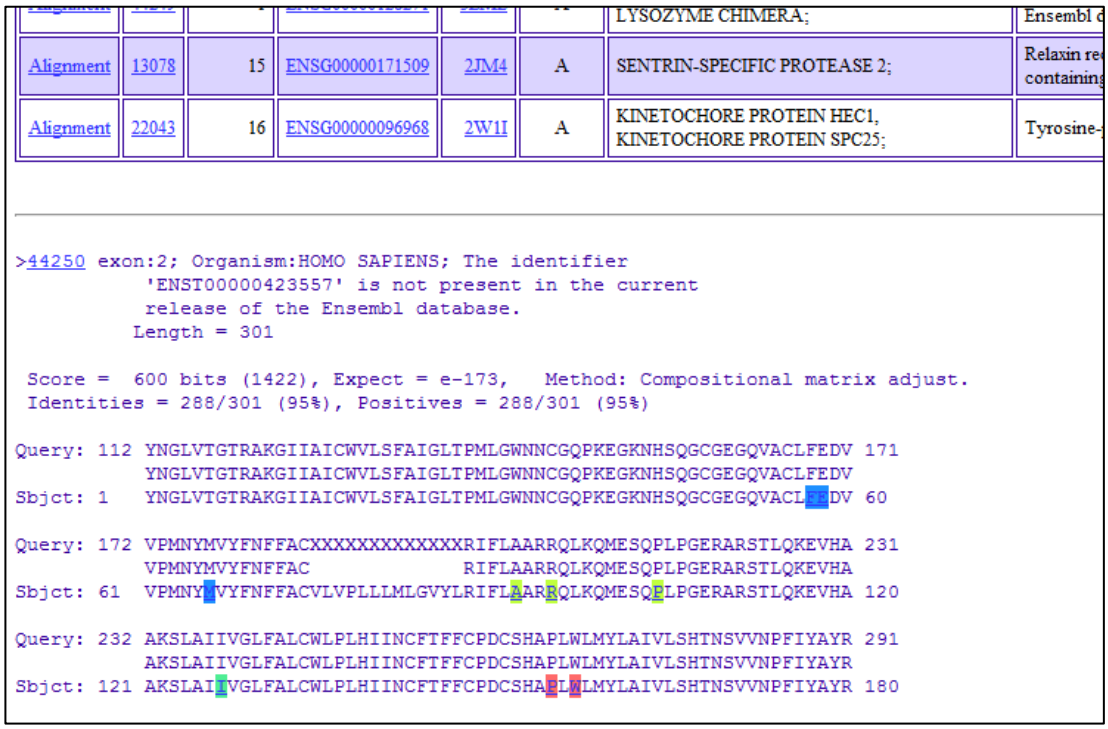
Наименование	Кол-во записей
Ионы металлов	2917
Анионы кислот	2401
Органические кислоты	595
Нуклеотидфосфаты	799
Фосфосахара	308
Белки	73
Аминокислоты и их соединения	164
Коферменты	89
Спирты и их производные	665
Атомы и неорганические соединения	351
Амины и амиды	1112
Порфирины	59
Более мелкие классы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

Классификация белков проводилась по ключевым словам, включающим ткань, функцию, локализацию и процесс в названиях белков, извлеченных из базы данных Ensembl (Приложение 2, таблица 1). Всего было выделено 13

групп белков, среди которых максимально представленными оказались ферменты.

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 (Приложение 2, Таблица 2), среди которых наиболее представленными оказались неорганические лиганды.

Страница Exon BLAST Search предназначена для осуществления поиска по БД по сходству нуклеотидных или аминокислотных последовательностей в формате FASTA с использованием программы BLAST (рис. 2.10). Для осуществления такого поиска была проведена индексация аминокислотных и нуклеотидных последовательностей из БД SitEx с использованием инструментов программы BLAST..



						LYSOZYME CHIMERA;	Ensembl d
Alignment	13078	15	ENSG00000171509	2JM4	A	SENTRIN-SPECIFIC PROTEASE 2;	Relaxin re containing
Alignment	22043	16	ENSG00000096968	2W1I	A	KINETOCHORE PROTEIN HEC1, KINETOCHORE PROTEIN SPC25;	Tyrosine-

```
>44250 exon:2; Organism:HOMO SAPIENS; The identifier
          'ENST00000423557' is not present in the current
          release of the Ensembl database.
         Length = 301

 Score =  600 bits (1422), Expect = e-173,   Method: Compositional matrix adjust.
 Identities = 288/301 (95%), Positives = 288/301 (95%)

Query: 112 YNGLVTGTRAKGIIAICWVLSFAIGLTPMLGWNNCGQPKEGKNHSQGCGEGQVACLFEDV 171
           YNGLVTGTRAKGIIAICWVLSFAIGLTPMLGWNNCGQPKEGKNHSQGCGEGQVACLFEDV
Sbjct: 1   YNGLVTGTRAKGIIAICWVLSFAIGLTPMLGWNNCGQPKEGKNHSQGCGEGQVACLFEDV 60

Query: 172 VPMNYMVYFNFFACXXXXXXXXXXXXXRIFLAARRQLKQMESQPLPGERARSTLQKEVHA 231
           VPMNYMVYFNFFAC             RIFLAARRQLKQMESQPLPGERARSTLQKEVHA
Sbjct: 61  VPMNYTVYFNFFACVLVPLLLMLGVYLRIFLAARRQLKQMESQPLPGERARSTLQKEVHA 120

Query: 232 AKSLAIIVGLFALCWLPLHIINCFTFFCPDCSHAPLWLMYLAIVLSHTNSVVNPFIYAYR 291
           AKSLAIIVGLFALCWLPLHIINCFTFFCPDCSHAPLWLMYLAIVLSHTNSVVNPFIYAYR
Sbjct: 121 AKSLAITVGLFALCWLPLHIINCFTFFCPDCSHAPLWLMYLAIVLSHTNSVVNPFIYAYR 180
```

Рисунок 2.10. Фрагмент страницы с результатами поиска программой BLAST. Помимо списка с описанием последовательностей и ссылками на страницы описания экзонов, для каждого выравнивания размечены аминокислоты функциональных сайтов, расположенных в полипептиде, кодируемом найденным экзоном.

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 RMSD (среднеквадратичное отклонение между атомами структур) и Z-score (показатель, учитывающий качество выравнивания последовательностей и совместимость торсионных углов). Предоставляется возможность перехода на другие страницы интерфейса для получения детальной информации об экзонах и функциональных сайтах, а также возможность графической визуализации суперпозиции пространственных структур с последующим сохранением в формате PDB (рис. 2.11).

	Exon ID	Order	EnsGene	PDB	Chain	Molecule	ENSMolecule	Organism	ZScore	RMSD	Aligned	Size	Gaps
◉	41108	2	ENSG00000077312	1NU4	B	U1A RNA BINDING DOMAIN;	U1 small nuclear ribonucleoprotein A (U1 snRNP protein A)(U1A protein)(U1-A)	HOMO SAPIENS	4.1	2.72	56	58	10
○	8344	2	ENSMUSG00000002633	3D1M	A	SONIC HEDGEHOG PROTEIN;	Sonic hedgehog protein Precursor (SHH)(HHG-1)	MUS MUSCULUS	3.9	2.60	61	87	20
○	11825	3	ENSG00000165240	2AW0	A	MENKES COPPER-TRANSPORTING ATPASE;	Copper-transporting ATPase 1 (EC 3.6.3.4)(Copper pump 1)(Menkes disease-associated protein)	HOMO SAPIENS	3.9	3.08	58	71	11
○	5332	4	ENSG00000030110	2IMS	A	APOPTOSIS REGULATOR BAK;	Bcl-2 homologous antagonist/killer (Apoptosis regulator BAK)(Bcl-2-like protein 7)(Bcl2-L-7)	HOMO SAPIENS	3.5	3.16	44	59	3
○	10909	2	ENSG00000123561	2CEO	A	THYROXINE-BINDING GLOBULIN;	Thyroxine-binding globulin Precursor (T4-binding globulin)(Serpin A7)	HOMO SAPIENS	3.5	3.35	50	92	9

Get alignment

Рисунок 2.11. Пример страницы, содержащей результаты структурного выравнивания.

2.7 Применение системы SitEx для анализа особенностей кодирования функциональных сайтов белков.

2.7.1 Сравнение особенностей кодирования сайтов связывания одинаковых лигандов в негомологичных белках человека на примере глицеральдегид-3-фосфатдегидрогеназы

Глицеральдегид-3-фосфатдегидрогеназа (PDB: 1U8F, EC 1.2.1.12) связывает никотинамидаденин-динуклеотид (НАД) в своем активном сайте (AC2), сформированном на границе двух НАД-связывающих доменов и кодируемом всеми восемью экзонами последовательности. Фермент осуществляет следующую каталитическую реакцию:

глицеральдегид-3-фосфат + фосфат + NAD^+ =
1,3-дифосфоглицерат + NADH + H^+.

Глицеральдегид-3-фосфатдегидрогеназа является важным участником многих важных процессов, таких как апоптоз, репликация и репарация [122]. В результате поиска структурного сходства пространственной структуры глицеральдегид-3-фосфатдегидрогеназы с пространственными структурами фрагментов белков, кодируемых отдельно взятыми экзонами, выполненного с помощью 3D Exon Search, был найден полипептид, кодируемый шестым экзоном алкогольдегидрогеназы (PDB: 1D1T, EC 1.1.1.1) (рис. 2.12). Показатели структурного сходства имели следующие значения: Z-score 3.9, RMSD 3.4. Семейство алкогольдегидрогеназ обратимо окисляет первичные и вторичные спирты до альдегидов и кетонов в присутствии НАД [123,124]. Сравнение аминокислотной последовательности найденных фрагментов глицеральдегид-3-фосфатдегидрогеназы и алкогольдегидрогеназы показало отсутствие сходства. Однако в каждом из этих фрагментов располагался сайт связывания НАД. Таким образом, можно предположить, что эти сайты могли быть образованы в результате конвергентной эволюции [125,126].

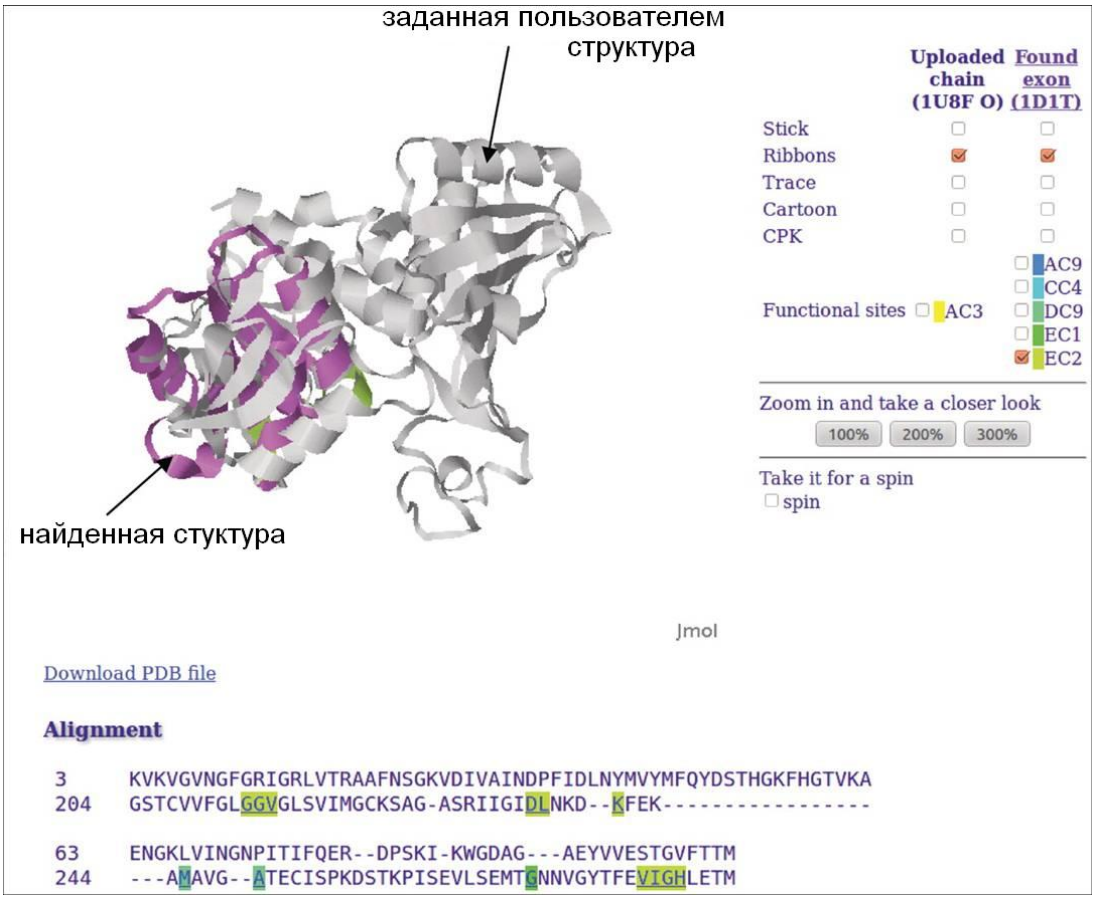


Рисунок 2.12. Структурное выравнивание пространственных структур глицеральдегид-3-фосфатдегидрогеназы 1U8F (помечено серым) и полипептида, кодируемого шестым экзоном алкогольдегидрогеназы (помечено розовым). Функциональные сайты, присутствующие в найденном полипептиде, отмечены на последовательности. Сравнение расположения позиций найденного полипептида с анализируемым представлено внизу.

2.7.2 Поиск сходства между фрагментами белков, кодируемых отдельными экзонами, и аминокислотными последовательностями прокариот на примере уропорфириногендекарбоксилазы *Bacillus subtilis*

В качестве примера использования системы SitEx для поиска сходства между фрагментами белковых последовательностей, кодируемых различными экзонами в геноме человека, и аминокислотной последовательностью заданного пользователем белка была выбрана

уропорфириногендекарбоксилаза *Bacillus subtilis*, так как уропорфириногендекарбоксилаза является ключевым ферментом в синтезе тетрапироллов, в частности, гемма, хлорофлла, сирогема и витамина B_{12} [127].

Поиск с помощью Exon BLAST Search, в котором входными данными была аминокислотная последовательность уропорфириногендекарбоксилазы *Bacillus subtilis*, показал высокое сходство с фрагментами последовательности уропорфириногендекарбоксилазы человека, кодируемых двумя экзонами (экзоны 2 и 5). Длина полипептида, кодируемого вторым экзоном, составила 38 аа, а пятым экзоном – 66 аа, E-value составило 10^{-10} и 10^{-4}, соответственно. Всего ген уропорфириногендекарбоксилазы человека содержит 10 экзонов. Сайт связывания копропорфирина АС1 состоит из 18 аминокислот, кодируемых фрагментами ДНК, расположенных в 8 экзонах. Оказалось, что фрагмент, кодирующий второй экзон, обладающий максимальным сходством содержит наибольшее количество аминокислотных остатков.

Из 17 аминокислот функционального сайта человека 7 отличаются от аминокислот активного центра уропорфириногендекарбоксилазы бактерии (рис. 2.13), при этом замены наблюдаются для гидрофобных аминокислот, а аминокислоты, непосредственно связывающиеся с копропорфирином, остаются консервативными [127] и содержатся в полипептидах, кодируемых в том числе 2 и 5 экзонами уропорфириногендекарбоксилазы человека. Таким образом, выявлены экзоны уропорфириногендекарбоксилазы человека, кодирующие наиболее консервативные и функционально значимые участки последовательности белка.

```
Hs  13   ELKNDTFLRAAWGEETDYTPVWCMRQAGRYLPEFRETRAAQDFFSTCRSPEACCELTLQP  72
         E  N+TFL+AA GE+ D+TPVW MRQAGR  PE+R+ +     F     PE C  +T  P
Bs  11   ETFNETFLKAARGEKADHTPVWYMRQAGRSQPEYRKLKEKYGLFEITHQPELCAYVTRLP  70

Hs  73   LRRFPLDAAIIFSDTLVVPQALGMEVTMVPGKGPSFPEPLREEQDLERLR--DPEVVASE  130
         + ++ +DAAI++ DI+    ++G++V +  G GP   +P+R   D+E+L   DPE    +
Bs  71   VEQYGVDAAILYKDIMTPLPSIGVDVEIKNGIGPVIDQPIRSLADIEKLGQIDPE---QD  127

Hs  131  LGYVFQAIT-LTRQRLAGRVPLIGFAGAPWTLMTYMVEGGGSSTMAQAKRWLYQRPQASH  189
         + YV + I  L  ++L   VPLIGF+GAP+TL +YM EGG S    + K ++Y  P A +
Bs  128  VPYVLETIKLLVNEQL--NVPLIGFSGAPFTLASYMTEGGPSKNYNKTKAFMYSMPDAWN  185

Hs  190  QLLRILTDALVPYLVGQVVAGAQALQLFESHAGHLGPQLFNKFALPYIRDVAKQVKARLR  249
          L+  L D ++ Y+  Q+ AGA+A+Q+F+S  G L    +      YI+ V  ++ + L
Bs  186  LLMSKLADMIIVYVKAQIKAGAKAIQIFDSWVGALNQADYRT----YIKPVMNRIFSELA  241

Hs  250  EAGLAPVPMIIFAKDGHFALEELAQAGYEVVGLDWTVAPKKARECVGKTVTLQGNLDPCA  309
         +     VP+I+F         +      +VVGLDW +   +AR   G T T+QGNLDP
Bs  242  KEN---VPLIMFGVGASHLAGDWHDLPLDVVGLDWRLGIDEAR-SKGITKTVQGNLDPSI  297

Hs  310  LYASEEEIGQLVKQMLDD-FGPHRYIANLGHGLYPDMDPEHVGAFVDAVHKHSR        362
         L A  E I Q  K++LD       +I NLGHG++PD+ PE +      VH++S+
Bs  298  LLAPWEVIEQKTKEILDQGMESDGFIFNLGHGVFPDVSPEVLKKLTAFVHEYSQ        351
```

Рисунок 2.13. Выравнивание функциональных сайтов в последовательностях уропорфириногендекарбоксилазы Homo sapiens (красный) и Bacillus subtilis (зеленый; голубым помечены остатки, связывающиеся с копропорфирином).

2.7.3 Исследование разрывности сайтов в функционально близких доменах белков, кодируемых генами с различной экзонной структурой на примере домена карбоксилазы типа B

Домен карбоксилаза типа B (Pfam: PF00135) встречается в белках с различной функцией и может быть найден в комбинациях с различными доменами других белков. Исследование разрывности сайтов в функционально близких доменах белков, кодируемых генами с различной экзонной структурой проведено на примере домена карбоксилазы типа B из ацетилхолинэстеразы (PDB:2X8B). Этот фермент осуществляет катализ гидролиза ацетилхолина до холина и ацетата в синапсах.

В результате реакции происходит дезактивация ацетилхолина в синаптической щели, что, в частности, необходимо для расслабления мышечных клеток. Используя программу Exon BLAST Search системы SitEx, было установлено сходство между последовательностью белка ацетилхолинэстеразы и последовательностями полипептидов, кодируемых

отдельно взятыми экзонами генов бутирилхолинэстеразы (PDB:1P0I), холестеролэстеразы (PDB:1AQL), карбоксилэстеразы-1 (PDB:2H7C), нейролигина-2 (PDB:3BL8), нейролигина-4 (PDB:3BE8). Все эти белки имеют в структуре домен карбоксилазы типа B. На основе результатов поиска программой BLAST в системе SitEx показано, что домен карбоксилазы типа B кодируется различным количеством экзонов во всех рассмотренных белках (рис. 2.14). Необходимо отметить, что экзон 2 в кодирующей структуре нейролигина-2 не имеет сходства с последовательностями других экзонов, кодирующих домен карбоксилазы, хотя расположен между теми экзонами, которые имеют такое сходство.

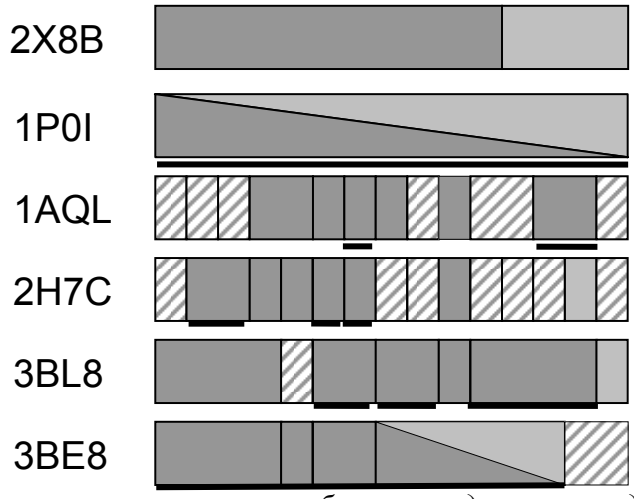


Рисунок 2.14. Экзонная структура участков белков, содержащих домен карбоксилазы типа B. Вертикальными полосами показаны границы экзонов. В белке 2X8B темно-серый цветом показан первый экзон, кодирующий N-терминальную часть домена карбоксилазы типа B, светло-серым показан второй экзон, кодирующий C-терминальную часть домена карбоксилазы типа B. В других белках (1P0I, 1AQL, 2H7C, 3BL8, 3BE8) темно-серым цветом показаны фрагменты, кодируемые отдельными экзонами, имеющие сходство с последовательностью N-терминальной части домена белка 2X8B, светло-серым – показано сходство с C-терминальной частью. Штрихом показаны фрагменты последовательностей белков, кодируемые отдельными экзонами, не имеющие сходства с последовательностью домена. Черными полосками показаны фрагменты, кодируемые экзонами, обладающие значимым структурным сходством со структурой 2X8B (Z-score>3.5).

Для сравнительного анализа разрывности функциональных сайтов, входящих в состав домена карбоксилазы типа B был рассчитан коэффициент CoefE для всех рассматриваемых выше белков (Таб. 2.6).

Таблица 2.6. Сайты связывания лигандов в домене карбоксилазы типа B

PDB ID	*ФС*	*Лиганд*	*CoefE*	*CoefA*	*Кол-во экзонов*	*Кол-во доменов*	*Кол-во цепей белка*
1P0I	AC1	NAG	0	0.6	1	1	1
	AC4	NAG	0	0.4	1	1	1
	AC5	NAG	0	0	1	1	1
	AC7	NAG	0	0.953	1	1	1
	AC8	NAG	0	0.955	1	1	1
	AC9	NAG	0	0.857	2	1	1
	AC3	FUC	0	0.912	1	1	1
	BC5	BUA	0	0.985	1	1	1
1AQL	AC1	NAG	0	0	1	1	1
	AC3	TCH	0/0	0.5/0	1/1	1	2
	AC4	TCH	0.25	0.932	3	1	1
	AC5	TCH	0.714/0	0.982/0	2/1	1	2
2H7C	AC1	NAG	0	0.25	1	1	1
	CC5	COA	0.33/0.25	0.953/0.91	4/3	1	2
	DC1	COA	0.5/0.375	0.972/0.959	4/6	1	2
3BL8	AC1	NAG	0.714	0.991	2	1	1
	AC2	NAG	0.714	0.996	2	1	1
	AC4	NAG	0	0.4	1	1	1
	AC5	NAG	0.33/0	0.968/0	2/1	1	2
	AC7	NAG	0.33	0.977	2	1	1
	AC8	BMA	0	0	1	1	1
	AC9	MAN	0	0.939	2	1	1
	AD1	MAN	0	0	1	1	1
	BC2	MAN	0	0.968	2	1	1
3BE8	AC1	NAG	0	0	1	1	1
	AC2	NAG	0	0.886	1	1	1
	AC5	FLC	0	0.933	2	1	1
	AC9	NA	0	0.5	1	1	1

Проанализированы сайты связывания N-ацетил-D-глюкозамина (NAG), таурохолевой кислоты (TCH), кофактора A (COA), фукозы (FUC), бутановой кислоты (BUA), маннозы (MAN и BMA), лимонной кислоты (FLC). При этом только для NAG-связывающего сайта имелась информация по его расположению во всех рассматриваемых белках. Оказалось, что для NAG-связывающего сайта значение CoefE равно 0 во всех белках за исключением нейролигина-2. NAG-связывающий сайт в нейролигине-2 обладал

специфической особенностью. В отличие от других рассмотренных белков, в нейролигине-2 он кодировался двумя экзонами, между которыми была вставка экзона, не кодирующего функциональный сайт. Следует также отметить особенность кодирования других лигандов, хотя они были представлены только одним из рассмотренных выше белков. ТСН- и CoA-связывающие сайты также оказались разрывными по экзонам. Эти сайты имеют относительно большой размер по сравнению с сайтами связывания FUC, BUA, MAN, FLC, которые кодировались одним или несколькими соседними экзонами.

Таким образом, на примере домена карбоксилазы типа В показано, что функциональные возможности системы SitEx позволяют проводить анализ разрывности сайтов в функциональных доменах белков, кодируемых генами с различной экзонной структурой.

2.8 Заключение

В главе описаны этапы разработки компьютерной системы SitEx. Описано устройство баз данных, входных и выходных форматов данных, используемых в последующем для интеграции данных и во вспомогательных программах. В результате интеграции была создана база данных, содержащая информацию о разметке функциональных сайтов в экзонной структуре генов, являющаяся частью системы SitEx. Система SitEx иммет гибкий интуитивно ясный для пользователя интерфейс, а также содержит программы Exon BLAST Search и 3D Exon Search, позволяющие решать задачи анализа соотношения между экзон-интронной структурой генов и особенностями структурно-функциональной организации белков.

Функциональные возможности системы SitEx были продемонстрированы на примерах решения следующих задач:

1) *Сравнение особенностей кодирования сайтов связывания одинаковых лигандов в негомологичных белках человека на примере глицеральдегид-3-фосфатдегидрогеназы.* В результате анализа показано, что НАД-

связывающие сайты в глицеральдегид-3-фосфатдегидрогеназе и алкогольдегидрогеназе расположены в пространственно сходных участках, но не гомологичных по аминокислотной последовательности, что может быть следствием конвергентной эволюции этих сайтов;

2) *Поиск сходства между фрагментами белков, кодируемых отдельными экзонами, и аминокислотными последовательностями прокариот.* На примере белка уропорфириногендекарбоксилазы человека, имеющего общую функцию с белком *Bacillus subtilis*, с помощью SitEx выявлены экзоны, кодирующие наиболее консервативные и функционально значимые участки последовательности белка;

3) *Исследование разрывности сайтов в функционально близких доменах белков, кодируемых генами с различной экзонной структурой.* С использованием SitEx на примере домена карбоксилазы типа B, часто встречающегося в мозаичных белках, показано, что экзонная структура, кодирующая этот домен, вариабельна, однако, разрывность по экзонам NAG-связывающего сайта проявляет консервативность.

Таким образом, компьютерная система SitEx обеспечивает возможность реализации сложных запросов, позволяющих формировать различные сценарии анализа функциональной организации генов, особенностей кодирования и эволюции функциональных сайтов с учетом экзонной структуры гена, а также задач выявления экзонов, задействованных в эволюционных перетасовках, вставках и делециях.

Глава 3. Статистический анализ закономерностей кодирования функциональных сайтов белков в генах позвоночных

3.1 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) на основе БД SitEx.

Выборка ЭКФС включала в себя 6444 экзона, а выборка ЭНФС - 10679 экзонов. Всего в анализе участвовало 2021 кодирующих последовательностей генов. Распределения длин экзонов представлено на рисунке 3.1. Статистический анализ с помощью $\chi 2$ двух распределений длин экзонов из выборок ЭКФС и ЭНФС показал их значимое различие ($\chi 2$=582.8,

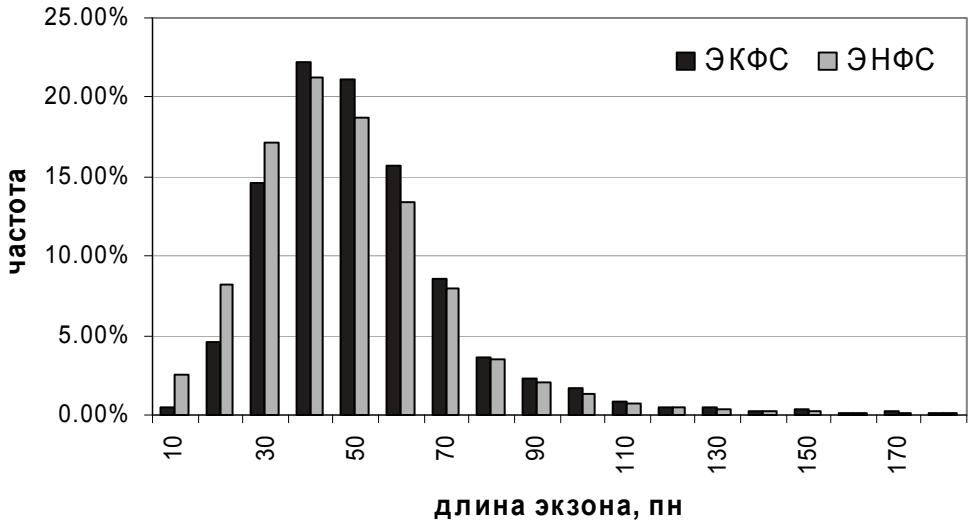


Рисунок 3.1. Распределения длин экзонов из выборок ЭНФС (I) и ЭКФС(II).

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.

3.2 Анализ консервативности экзонов, кодирующих функциональные сайты

Для анализа консервативности экзонов, кодирующих функциональные сайты белков, была сформирована выборка из 955 последовательностей белков эукариотических организмов, имеющих известную пространственную структуру, полученная с помощью случайного выбора из БД PDB. На основе сравнения сходства данных последовательностей с последовательностями фрагментов белков, кодируемых отдельно взятыми экзонами, представленными в БД SitEx, были сформированы две выборки экзонов, кодирующих (ЭКФС) и некодирующих (ЭНФС) функциональные сайты белков. Сравнение последовательностей с БД SitEx осуществлялось с помощью Exon BLAST Search. Для работы Exon BLAST Search задавались следующие параметры: blastp, E-value<1000, PAM 30. В результатах поиска среди последовательностей ЭКФС принимались во внимание только те, в которых выравненный участок содержал аминокислоты функционального сайта. Из таблицы выдачи программы Exon BLAST Search, содержащей выравнивания и значения сходства анализируемой последовательности с последовательностями из БД SitEx, для дальнейшего анализа бралась последовательность с минимальным значением E-value. Для каждой из полученной таким образом выборок ЭКФС и ЭНФС рассчитывалось распределение отношения длины выравнивания к длине экзона. Средние значения для данного показателя оказались равны 0.535 и 0.582 для ЭНФС и ЭКФС, соответственно. Различия между этими распределениями оказались статистически значимыми (значение критерия $\chi 2$ 61.4, df=15, p<<0.001).

Таким образом, показано, что фрагменты белков, кодируемые отдельно взятыми экзонами, проявляют большую гомологию между собой в случае, если они содержат аминокислотные остатки функциональных сайтов.

Аналогичный тест был проведен при поиске сходных фрагментов пространственных структур для каждого белка из сформированной выше выборки с помощью программы 3DPDBScan (3D Exon Search) (раздел 2.2.4). В результате оказалось, что около 85% найденных структур полипептидов в базе данных SitEx, кодируемых отдельно взятым экзоном, содержали аминокислотные остатки функционального сайта. Всего в базе данных SitEx содержится 57% полипептидов, включающих аминокислотные остатки функциональных сайтов. Точный критерий Фишера отверг гипотезу о том, что количество полипептидов, содержащих аминокислоты функционального сайта, в полученной выборке и в базе данных одинаково (значение критерия 0.24, $p << 0.01$).

Структуры полипептидов, кодируемые единственным экзоном, полученные в результате работы программы 3D Exon Search, разделялись на содержащие аминокислотные остатки функциональных сайта и не содержащие их. Сравнение распределений максимальных значений параметра Z-score для этих выборок критерием $\chi 2$ показало значимое различие этого параметра между структурами полипептидов в зависимости от содержания ими аминокислот функционального сайта (значение критерия $\chi 2$ 56.2, df=11, $p << 0.001$). При этом средние значения Z-score составляют 4.6 и 5.3 для выборок полипептидов, не содержащих и содержащих функциональный сайт, соответственно.

Таким образом, показано, что фрагменты белков, кодируемые отдельно взятыми экзонами, проявляют большее структурное сходство между собой в случае, если они содержат аминокислотные остатки функциональных сайтов.

3.3 Исследование разрывности функциональных сайтов

Чтобы оценить разрывность функциональных сайтов по экзонам, необходимо оценить характеристики исследуемой выборки функциональных сайтов. Один функциональный сайт кодируется в среднем нуклеотидной последовательностью 2.6 экзонов, всего ЭКФС в последовательности в среднем 4 (рис. 3.2). 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 ($\rho \approx 0.4$, $p < 0.05$).

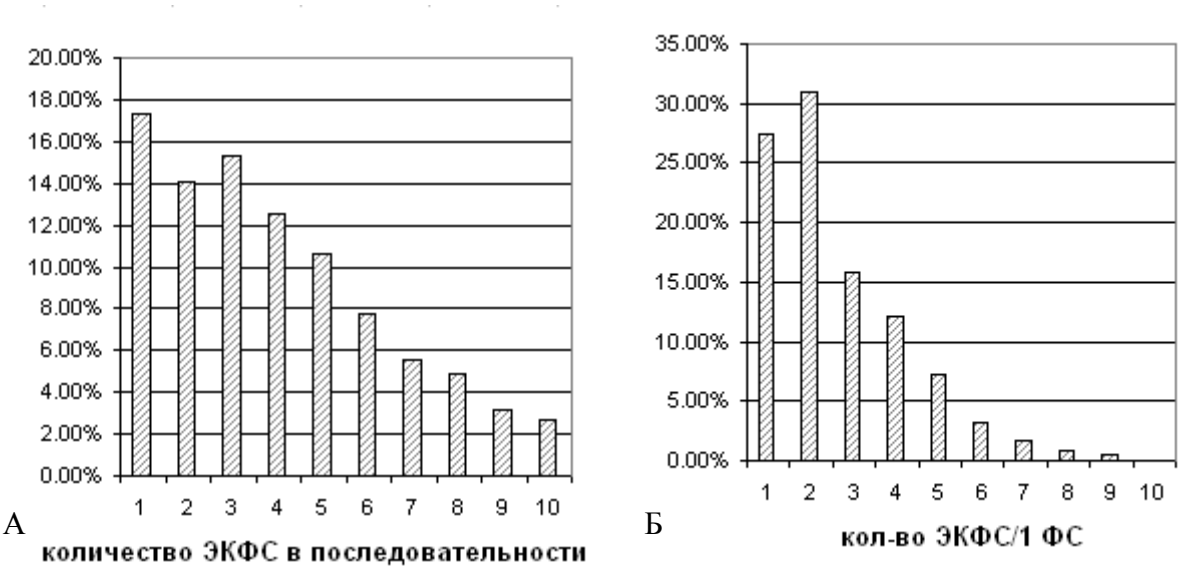


Рисунок 3.2. А. Распределение количества экзонов, кодирующих все функциональные сайты в одной последовательности. Б. Распределение количества экзонов, кодирующих один функциональный сайт, в выборке.

Для статистической проверки гипотезы о том, что разрывность функциональных сайтов по экзонам значимо меньше, чем ожидается по случайным причинам, оценивалось ожидаемое и наблюдаемое количество 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. 3.3).

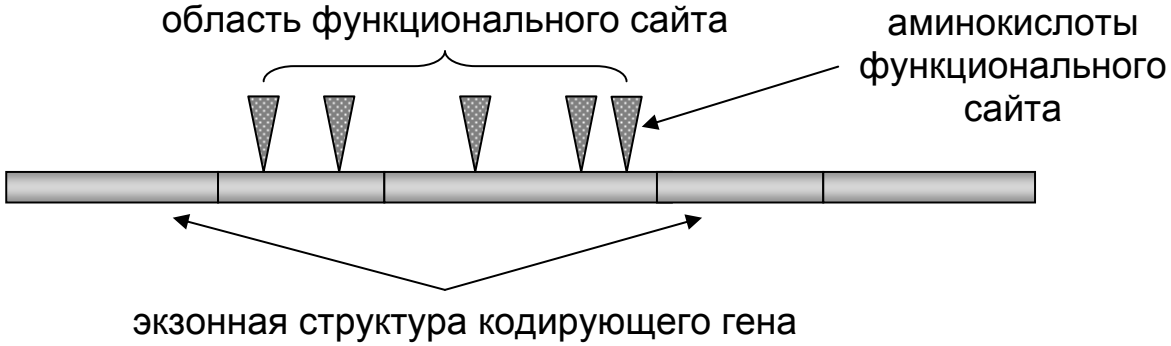


Рисунок 3.3. Область ФС, отображенная на экзонной структуре кодирующего гена.

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 $\chi 2$ ($\chi 2=22.4$, $p<0.01$, $df=6$).

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; N1=427; N2=390; $p<<0.01$). Аналогичный анализ проводился для отдельно взятых групп функциональных сайтов, наиболее представленных в БД SitEx. Во всех случаях наблюдаемое количество экзонов, кодирующих область функционального сайта, было значимо меньше ожидаемого, а распределения были различны согласно тесту Манна-Уитни (для аминокислот $U=24$, $N_1=21$, $N_2=15$, $p<0.01$; для органических кислот $U=85.5$, $N_1=23$, $N_2=17$, $p=0.002$; для аминов $U=86$, $N_1=27$, $N_2=16$, $p=0.0008$; для спиртов $U=32.5$, $N_1=27$, $N_2=13$, $p<0.01$; для сложных органических соединений $U=43.5$, $N_1=19$, $N_2=19$, $p<0.01$) (рис. 3.4).

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

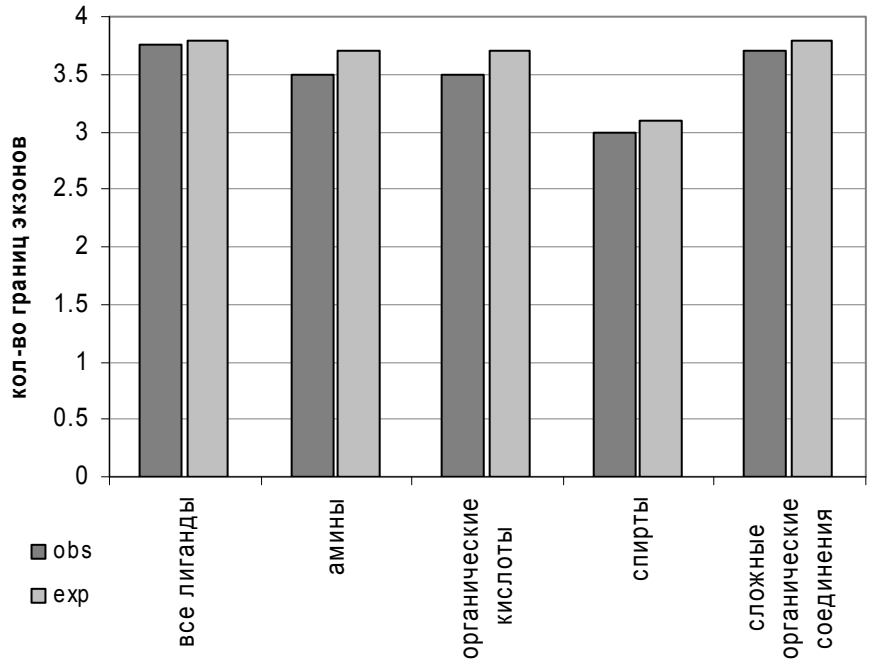


Рисунок 3.4. Наблюдаемое(obs) и ожидаемое по случайным причинам (exp) среднее количество экзонов в области функциональных сайтов и вне ее.

3.4 Анализ частот кодонов в 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 [128]. 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 (рис. 3.5). Встречаемость аминокислоты рассчитывалась как отношение частоты данной аминокислоты во всех функциональных сайтах к частоте ее встречаемости во всех последовательностях белков. Среди наиболее часто встречающихся оказались гидрофильные аминокислоты, предпочтительно располагающиеся на поверхности белков (гистидин, цистеин, тирозин и др.), что характерно для функциональных сайтов (см. раздел 1.2.4).

His	2,28	Trp	1,47	Met	0,95	Ile	0,67
Cys	1,98	Asn	1,44	Thr	0,91	Ala	0,6
Tyr	1,69	Phe	1,25	Glu	0,88	Leu	0,6
Asp	1,58	Gly	1,11	Ser	0,86	Val	0,58
Arg	1,55	Lys	0,97	Gln	0,81	Pro	0,39

Рисунок 3.5. Относительная встречаемость аминокислот в функциональных сайтах, представленных в БД SitEx.

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 [86]. Однако существуют работы, в которых авторы выдвигают гипотезу о том, что в таких районах отбор может быть направлен также на нуклеотиды A и T [87]. 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 $\chi 2$. 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 (Приложение 3). 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. 3.6).

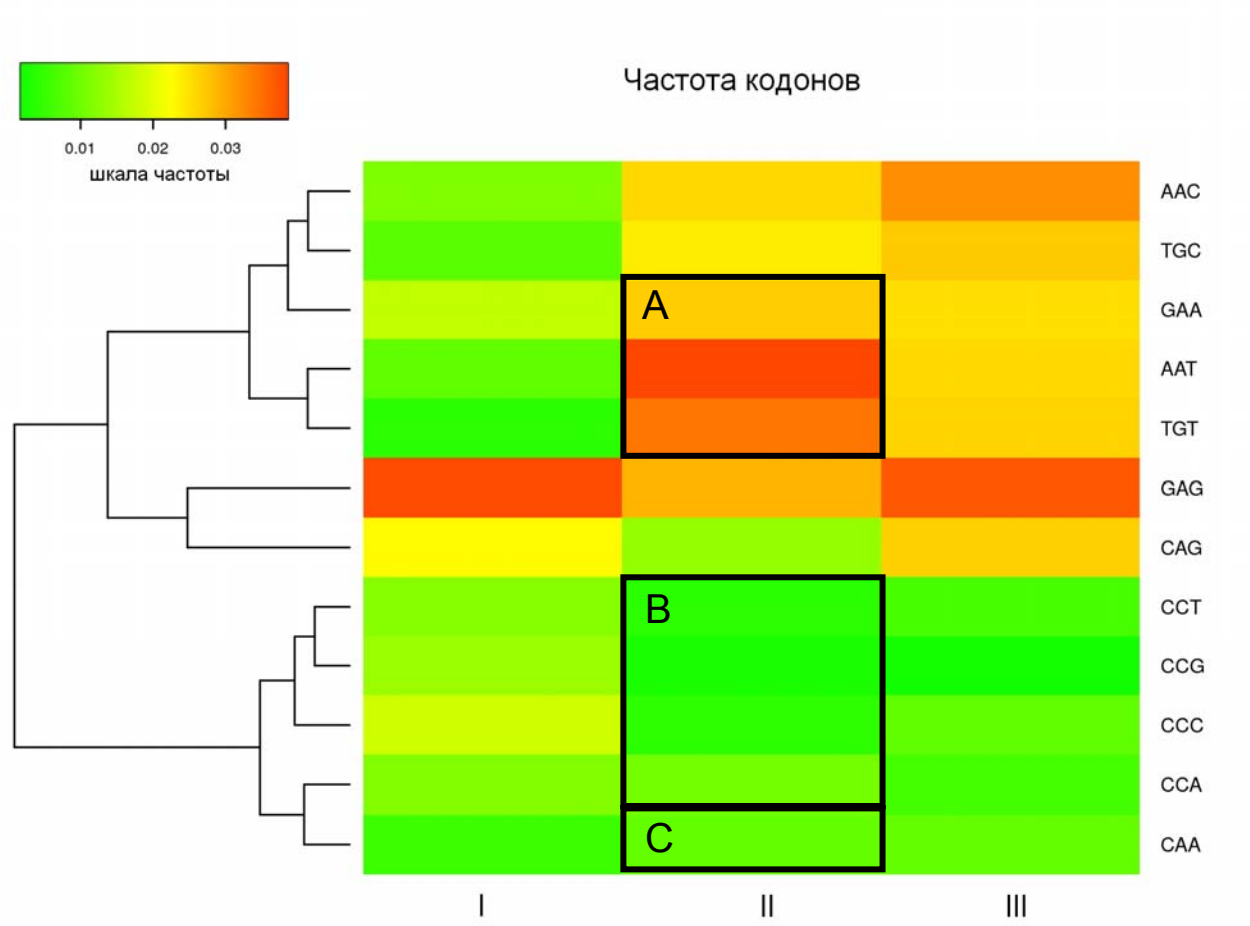


Рисунок 3.6. Встречаемость кодонов, кодирующих аспарагин, пролин, глутамин, глутаминовую кислоту и цистеин в различных участках последовательности. I – участок последовательности на 5'-конце экзона, ограниченный пятью кодонами; II – часть функционального сайта, кодируемая на 5'-конце экзона, ограниченном 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.

Данные отличия могут быть объяснены наблюдаемой повышенной частотой представленности аденина и тимина в третьей позиции кодонов, кодирующих перечисленные выше аминокислоты (Приложение 4). 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: TTT (Phe), ATT (Ile), AAA (Lys), TTA, TTG (Leu), ACA, ACT (Thr), TAT (Tyr), GGT (Gly), CGA, CGT (Arg), AGT, TCT, TCA (Ser). Это может быть следствием влияния генетических сигналов (в частности, сайтов сплайсинга) и кода функциональных сайтов друг на друга.

3.5 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 (Приложение 5).

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 ($p<8.3*10\text{-}6$ с учетом поправки Бонферрони ($Z=4.86$) и $p<8.3*10\text{-}6$ ($Z=4.47$) 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 3.7.

Ранее было показано [96], 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.

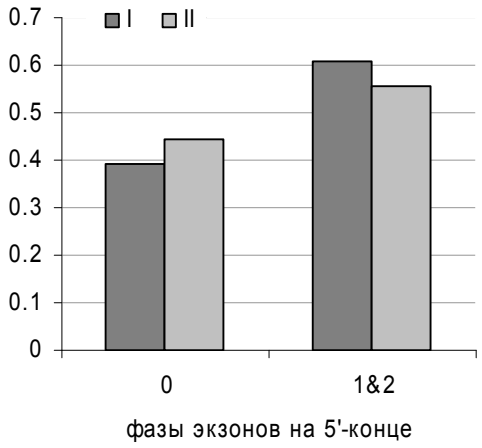


Рисунок 3.7. Распределение средней частоты фаз среди экзонов, кодирующих аминокислоту функционального сайта на 5'-конце (I) и некодирующих ее (II).

Обсуждение

В работе проведен анализ особенностей кодирования функциональных сайтов белков в генах позвоночных. В частности, исследованы разрывность функциональных сайтов в кодирующей структуре гена, длина экзонов, кодирующих аминокислотные остатки функционального сайта, проанализирован состав кодонов во фрагментах ДНК, кодирующих аминокислотные остатки функциональных сайтов, а также представленность фаз на 5’-границах экзонов, в зависимости от содержания участков ДНК, кодирующих функциональные сайты.

В рамках решениях поставленных задач разработана база данных SitEx, содержащая информацию о позициях аминокислот функционального сайта в экзонной структуре кодирующего гена. База данных SitEx интегрирована с программами BLAST для поиска гомологии заданного белка с полипептидами, кодируемыми отдельно взятыми экзонами, представленными в базе данных. Также SiteEx интегрирована с программой 3DPDBScan для поиска структурной гомологии таких пептидов с заданной пространственной структурой белка или полипептида. Компьютерная система может быть использована для изучения структурно-функциональной организации генов; особенностей кодирования и эволюции функциональных сайтов с учетом экзонной структуры гена; выявления экзонов, задействованных в эволюционных перетасовках; планирования белково-инженерных экспериментов по направленной эволюции белков; дизайном новых искусственных белков, состоящих из фрагментов, кодируемых отдельными экзонами из разных генов и т.д. Система SitEx доступна по адресу: http://www-bionet.sscc.ru/sitex/ (возможен доступ вне ИЦиГ СО РАН).

На основе построенной базы данных о разметке функциональных сайтов белков на экзонной структуре гена была сформирована выборка для последующего статистического анализа.

Анализ разрывности функциональных сайтов, показывал, что функциональные сайты в большей степени кодируются фрагментами ДНК, расположенными в одном или близко расположенных экзонах. В то же время, экзоны, кодирующие функциональный сайт, как правило, длиннее тех, которые его не кодируют. Такое различие показывает существование эволюционного отбора на определенную длину экзонов, которые кодируют функциональный сайт.

Показано, что аминокислотные остатки функциональных сайтов реже кодируются на границе экзона соответствующими позиции кодонами в фазе 0, которая указывает на более древние вставки интронов, но чаще в фазах 1 и 2. Это может указывать на то, что аминокислоты функциональных сайтов, кодируемые фрагментами ДНК на границе экзона и интрона, в ходе эволюции реже сохраняются, но могут приобретаться.

Анализ использования кодонов во фрагментах ДНК, кодирующих функциональные сайты белка, показал неравномерное использование кодонов вдоль последовательности экзона. При этом в середине последовательности экзона использовались часто встречающиеся кодоны, вблизи же экзонных границ во фрагментах ДНК, кодирующих некоторые аминокислоты, использовались реже встречающиеся кодоны. Ранее в литературе было показано наличие нескольких кодов в ДНК и возможность их интерференции [129, 130], поэтому полученный результат может указывать на перекрытие генетических кодов и функциональных сайтов белков.

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.

Список литературы

1. Gilbert W. Why genes in pieces? // Nature. - 1978. - Vol. 271. - P. 501.
2. Kaessmann H., Zollner S., Nekrutenko A., Li W.H. Signatures of domain shuffling in the human genome // Genome Res. - 2002. - Vol. 12(11). - P. 1642-1650.
3. Branden C., Tooze J. Introduction to protein structure. - Garland Publishing, 1998. - 410 p.
4. Taylor W.R. The classification of amino acid conservation // J Theor Biol. - 1986.- Vol.-119(2). - p. 205-218.
5. Livingstone C.D., Barton G.J. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation // Comput Appl Biosci. – 1993. – Vol. 9. – p. 745–756.
6. Волькенштейн М. В. Биофизика. – М.: Наука, 1988. - 591 с.
7. Miller S., Janin J., Lesk A.M., Chothia C. Interior and surface of monomeric proteins // J Mol Biol. – 1987. – Vol. 196(3). – p. 641-56.
8. Физика белка: курс лекций с цветными и стереоскопическими иллюстрацями и задачами: учебное пособие / А. В. Финкельштейн, О. Б. Птицын. – 4-е изд., испр. и доп. – М.: КДУ, 2012. – 524 с. : табл., ил. [32] с. цв. ил.
9. Bourne, P.E. and Weissig, H. Structural Bioinformatics. - Wiley-Liss, 2003. - 649 p.
10. Lesk A. M. Introduction of protein science: architecture, function and genomics. – Oxford University Press, 2010. – 455 p.
11. Meyer E. A., Castellano R. K., Diederich F. Interactions with aromatic rings in chemical and biological recognition // Angew Chem Int Edit. – 2003. – Vol. 42. – p. 1210–1250.
12. Zhang C., DeLisi C. Estimating the number of protein folds // J Mol Biol. – 1998. – Vol. 284. – p. 1301–1305.

13. Wolf Y.I., Grishin N.V., Koonin E.V. Estimating the number of protein folds and families from complete genome data // J Mol Biol. – 2000. – Vol. 299(4). – p. 897-905.
14. Govindarajan S., Recabarren R., Goldstein R. A. Estimating the total number of protein folds // Proteins. – 1999. – Vol. 35. – p. 408–414.
15. Cossio P., Trovato A., Pietrucci F., Seno F., Maritan A., Laio A. Exploring the universe of protein structures beyond the Protein Data Bank // PLoS Comput Biol. – 2010. – Vol. 6(11). – e1000957.
16. Hubbard T. J., Murzin A. G., Brenner S. E., Chothia C.. SCOP: a structural classification of proteins database // Nucleic Acids Res. – 1997. – Vol. 25(1). – p. 236-239.
17. Sillitoe I., Cuff A.L., Dessailly B.H., Dawson N.L., Furnham N., Lee D., Lees J.G., Lewis T.E., Studer R.A., Rentzsch R., Yeats C., Thornton J.M., Orengo C.A. New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures // Nucleic Acids Res. – 2013. – Vol. 41 (Database issue). – p. D490-D498.
18. Ponting C. P., Russell R. R. The natural history of protein domains // Annu. Rev. Biophys. Biomol. Struct. – 2002. – Vol. 31. – p. 45–71.
19. Takahashi K., Noguti T., Hojo H., Yamauchi K., Kinoshita M., Aimoto S., Ohkubo T., Go M. A mini-protein designed by removing a module from barnase: molecular modeling and NMR measurements of the conformation // Protein Eng. – 1999. – Vol. 12(8). – p. 673-680.
20. Go M. Correlation of DNA exonic regions with protein structural units in haemoglobin // Nature. – 1981. – Vol. 291(5810). – p. 90-92.
21. Yanagawa H., Yoshida K., Torigoe C., Park J. S., Sato K., Shirai T., Go M. Protein anatomy: functional roles of barnase module // J. Biol. Chem. – 1993. – Vol. 268. – p. 5861-5865.
22. Sammut S. J., Finn R. D., Bateman A. Pfam 10 years on: 10,000 families and still growing // Brief Bioinform. – 2008. – Vol. 9. – p. 210-219.

23. Chothia C. One thousand protein families for the molecular biologist // Nature. – 1992. – Vol. 357. – p. 543–544.
24. Kendrew J. C., Bodo G., Dintzis H. M., Parrish R. G., Wyckoff H., Phillips D. C. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis // Nature. – 1958. – Vol. 181(4610). – p. 662-666.
25. Perutz M. F., Rossmann M. G., Cullis M. G., Muirhead H., Will G. North ACT Structure of haemoglobin. A three-dimensional Fourier synthesis at 5.5Å resolution, obtained by X-ray analysis // Nature. – 1960. – Vol. (185). – p. 416–422.
26. Anil-Kumar, Ernst R. R., Wüthrich K. A two-dimensional nuclear Overhauser enhancement (2D NOE) experiment for the elucidation of complete proton–proton cross-relaxation networks in biological macromolecules // Biochem. Biophys. Res. Comm. – 1980. – Vol. 95. – p. 1−6
27. Wagner. G., Wüthrich K. Sequential resonance assignments in protein ^{1}H nuclear magnetic resonance spectra: basic pancreatic trypsin inhibitor // J. Mol. Biol. – 1982. – Vol. 155. – p. 347−366.
28. Dubochet J., McDowall A. W. Vitrification of pure water for electron microscopy // J. Microsc. – 1981. – Vol. 124. – p. RP3–RP4.
29. Sigrist C. J. A., Cerutti L., de Castro E., Langendijk-Genevaux P. S., Bulliard V., Bairoch A., Hulo N. PROSITE, a protein domain database for functional characterization and annotation // Nucleic Acids Res. – 2010. – Vol. 38(Database issue). – p. 161-166.
30. Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. The Protein Data Bank // Nucleic Acids Res. – 2000. – Vol. 28(1). – p. 235-242.
31. Henikoff J. G., Greene E. A., Taylor N., Henikoff S., Pietrokovski S. Using the blocks database to recognize functional domains // Curr Protoc Bioinformatics. – 2002. – Vol. 00:2.2. – 2.2.1-2.2.32.
32. Attwood T. K., Bradley P., Flower D. R., Gaulton A., Maudling N., Mitchell A. L., Moulton G., Nordle A., Paine K., Taylor P., Uddin A., Zygouri C.

PRINTS and its automatic supplement, preprints // Nucleic Acids Res. – 2003. – Vol.31. – p. 400–402.

33. Wilson D., Pethica R., Zhou Y., Talbot C., Vogel C., Madera M., Chothia C., Gough J. SUPERFAMILY–sophisticated comparative genomics, data mining, visualization and phylogeny // Nucleic Acids Res. – 2009. – Vol. 37. – p. D380–D386.
34. Marchler-Bauer A., Anderson J. B., Chitsaz F., Derbyshire M. K., Weese-Scott C., Fong J. H., Geer L. Y., Geer R. C., Gonzales N. R., Gwadz M., He S., Hurwitz D. I., Jackson J. D., Ke Z., Lanczycki C. J., Liebert C.A., Liu C., Lu F., Lu S., Marchler G. H., Mullokandov M., Song J. S., Tasneem A., Thanki N., Yamashita R. A., Zhang D., Zhang N., Bryant S. H. CDD: specific functional annotation with the Conserved Domain Database // Nucleic Acids Res. – 2009. – Vol. 37. – p. D205–D210.
35. Selengut J. D., Haft D. H., Davidsen T., Ganapathy A., Gwinn-Giglio M., Nelson W. C., Richter A. R., White O. TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes // Nucleic Acids Res. – 2007. – Vol. 35. – p. D260-D264.
36. Mi H., Lazareva-Ulitsky B., Loo R., Kejariwal A., Vandergriff J., Rabkin S., Guo N., Muruganujan A., Doremieux O., Campbell M. J., Kitano H., Thomas P. D. The PANTHER database of protein families, subfamilies, functions and pathways // Nucleic Acids Res. – 2005. – Vol. 33. – p. D284–D288.
37. Bru C., Courcelle E., Carrere S., Beausse Y., Dalmar S., Kahn D. The ProDom database of protein domain families: more emphasis on 3D // Nucleic Acids Res. – 2005. – Vol. 33. – p. D212-D215.
38. Portugaly E,. Linial N., Linial M. EVEREST: a collection of evolutionary conserved protein domains // Nucleic Acids Res. – 2007. – Vol. 35. – p. D241–D246.
39. Letunic I., Doerks T., Bork P. SMART 6: recent updates and new developments // Nucleic Acids Res. – 2009. – Vol. 37. – p. D229-D232.

40. Eisenberg D., Marcotte E. M., Xenarios I., Yeates T. O. Protein function in the post-genomic era // Nature. - 2000. – Vol. 405(6788). – p. 823-826.
41. Jacq B. Protein function from the perspective of molecular interactions and genetic networks // Brief. Bioinform. – 2001. – Vol. 2. – p. 38-50.
42. Webb E. C. Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes / Academic Press. – San Diego, 1992. – 862 c.
43. Альбертс Б., Брей Д., Льюис Дж., Рэфф М., Робертс К., Уотсон Дж.. Молекулярная биология клетки. Том 2. – М.: Мир, 1994. – с. 539.
44. Pazienza R., Teresa M. AI*IA 2007: Artificial Intelligence and Human-Oriented Computing. – Rome, 2007. – 859 p.
45. Ivanisenko V.A., Pintus S.S., Grigorovich D.A., Kolchanov N.A. PDBSite: a database of the 3D structure of protein functional sites // Nucleic Acids Res. – 2005. – Vol. 33(Database issue). – p. D183-D187.
46. Биохимия: Учеб. для вузов / Под ред. Е.С. Северина. – М.: ГЭОТАР-МЕД, 2004. – 779 с.
47. Whiting A. K., Peticolas W. L. Details of the acyl-enzyme intermediate and the oxyanion hole in serine protease catalysis // Biochemistry. – 1994. – Vol. 33. – p. 552–561.
48. Burgoyne N.J., Jackson R.M. Chapter 7. Predicting protein function from surface properties // Editor D.G. Rigden. From Protein Structure to Function with Bioinformatics. –2009. – p. 167 -186
49. Кольман Я., Рём К.-Г.. Наглядная биохимия. – М.: Мир, 2000. – 469 с.
50. Macias M. J., Wiesner S., Sudol M. WW and SH3 domains, two different scaffolds to recognize proline-rich ligands // FEBS Lett. – 2002. – Vol. 513. – p. 30–37.
51. Obst U., Banner D.W., Weber L., Diederich F. Molecular recognition at the thrombin active site: Structure-based design and synthesis of potent and

selective thrombin inhibitors and the x-ray crystal structures of two thrombin-inhibitor complexes // Chem. Biol. – 1997. – Vol. 4. – p. 287 – 295.

52. Cauet E., Rooman M., Wintjens R., Lievin J., Biot C. Histidine-aromatic interactions in proteins and protein-ligand complexes: quantum chemical study of X-ray and model structures // J. Chem. Theory Comput. – 2005. – Vol. 1. – p. 472-483.
53. Vyas N. K., Vyas M. N., Quiocho F. A. Sugar and signal-transducer binding sites of the Escherichia coli galactose chemoreceptor protein // Science. – 1988. – Vol. 242. – p. 1290-1295.
54. Kumar S., Kumar N., Gaur R.K. Amino acid frequency distribution at enzymatic active site // IIOAB journal. – 2011. – Vol. 2(4). – p. 23-30
55. Bartlett G.J., Porter C.T., Borkakoti N., Thornton J.M. Analysis of catalytic residues in enzyme active sites // J Mol Biol. – 2002. – Vol. 324(1). – p. 105-121.
56. Chothia C., Gough J. Genomic and structural aspects of protein evolution. Biochem J. – 2009. – Vol. 419. – p. 15-28.
57. Оно С. Генетические механизмы прогрессивной эволюции. – М.:Мир, 1973. – 222 с.
58. Afonnikov D. A., Oshchepkov D. Yu., Kolchanov N. A.. Detection of conserved physico-chemical characteristics of proteins by analyzing clusters of positions with co-ordinated substitutions // Bioinformatics. – 2001. – Vol. 17(11). – p. 1035-1046
59. Studer R. A., Dessailly B. H., Orengo C. A. Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes // Biochem J. – 2013. – Vol. 449(3). – p. 581-594.
60. Bloom J. D., Arnold F. H. In the light of directed evolution: pathways of adaptive protein evolution // Proc. Natl. Acad. Sci. – 2009. – Vol. 106. – p. 9995-10000.
61. Soskine M., Tawfik, D. S. Mutational effects and the evolution of new protein functions // Nature Reviews Genetics. – 2010. – Vol. 11. – p. 572-582.

62. Porter C. T., Bartlett G. J., Thornton J. M. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data // Nucl. Acids. Res. – 2004. – Vol. 32. – p. D129-D133.
63. Torrance J. W., Bartlett G. J., Porter C. T., Thornton J. M. Using a Library of Structural Templates to Recognise Catalytic Sites and Explore their Evolution in Homologous Families // J Mol Biol. – 2005. – Vol. 347. – p. 565-581
64. Gold N. D., Jackson R. M. SitesBase: a database for structure-based protein-ligand binding site comparisons // Nucleic Acids Res. – 2006. – Vol. 34. – p. D231–D234.
65. Kinoshita K., Furui J., Nakamura H. Identification of protein functions from a molecular surface database, eF-site // J. Struct. Func. Genomics. – 2002. - Vol. 2. – p. 9-22.
66. Kellenberger E., Muller P., Schalon C., Bret G., Foata N., Rognan D. sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank // J. Chem. Inf. Model. – 2006. – Vol. 46. – p. 717–727.
67. Lopez G., Valencia A., Tress M. FireDB–a database of functionally important residues from proteins of known structure // Nucleic Acids Res. – 2007. – Vol. 35. – p. D219-223.
68. Dessailly B., Lensink M., Orengo C., Wodak S. LigASite: a database of biologically relevant binding sites in proteins with known apo-structures. Nucleic Acids Res. – 2008. – Vol. 36. – p. D667-673.
69. Koonin E. V. Evolution of genome architecture // Int. J. Biochem. Cell Biol. – 2009. – Vol. 41. – p. 298-306.
70. Lynch M. The origins of eukaryotic gene structure // Mol Biol Evol. – 2006. – Vol. 23. – p. 450-468.
71. Gudlaugsdottir S., Boswell D. R., Wood G. R., Ma J. Exon size distribution and the origin of introns // Genetica. – 2007. – Vol. 131. – p. 299-306.
72. Deutsch M., Long M. Intron-exon structures of eukaryotic model organisms // Nucleic Acids Res. – 1999. – Vol. 27(15). – p. 3219-3228.

73. Sakharkar M.K., Chow V.T., Kangueane P. Distributions of exons and introns in the human genome // In Silico Biol. – 2004. – Vol. 4(4). – p. 387-393.
74. Keeling P. J., Palmer J. D. Horizontal gene transfer in eukaryotic evolution // Nat Rev Genet. – 2008. – Vol. 9. – p. 605-618.
75. Koonin E. V., Galperin M. Y. Sequence - Evolution - Function. Computational Approaches in Comparative Genomics. – Boston, 2002 – 461 p.
76. Lynch M. Genomics. Gene duplication and evolution // Science. – 2002. – Vol. 297. – p. 945-947.
77. Kondrashov F.A., Koonin E.V. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications // Trends Genet. – 2004. – Vol. 20. – p. 287–290.
78. Kolkman J. A., Stemmer W. P. C. Directed evolution of proteins by exon shuffling. // Nat Biotechnol. – 2001. – Vol. 19(5). – p. 423-428.
79. Rogozin I.B., Carmel L., Csuros M., Koonin E.V. Origin and evolution of spliceosomal introns // Biol Direct. – 2012. – Vol. 16. – p. 7-11.
80. Patthy L. Genome evolution and the evolution of exon-shuffling-a review // Gene. – 1999. – Vol. 238(1). – p. 103-114.
81. Andersson J.O. Lateral gene transfer in eukaryotes // Cell Mol Life Sci. – 2005. – Vol. 62(11). – p. 1182-1197.
82. Nikoh N., Tanaka K., Shibata F., Kondo N., Hizume M., Shimada M., Fukatsu T. Wolbachia genome integrated in an insect chromosome: evolution and fate of laterally transferred endosymbiont genes // Genome Res. – 2008. – Vol. 18(2). – p. 272-280.
83. Kondrashov F. A., Koonin E. V., Morgunov I. G., Finogenova T. V., Kondrashova M. N. Evolution of glyoxylate cycle enzymes in Metazoa: evidence of multiple horizontal transfer events and pseudogene formation // Biology Direct. – 2006. – Vol. 1:31 – 14 p.
84. Jenkins C., Samudrala R., Anderson I., Hedlund B. P., Petroni G., Michailova N., Pinel N., Overbeek R., Rosati G., Staley J. T. Genes for the cytoskeletal

protein tubulin in the bacterial genus Prosthecobacter // Proc Natl Acad Sci. – 2002. – Vol. 99(26). – p. 17049-17054.

85. Richards T. A, Dacks J. B., Jenkinson J. M., Thornton C. R., Talbot N. J. Evolution of filamentous plant pathogens: gene exchange across eukaryotic kingdoms // Curr Biol. – 2006. – Vol. 16(18). – p. 1857-1864.
86. Gelfand M. S. Statistical analysis of mammalian pre-mRNA splicing sites // Nucleic Acids Res. – 1989. – Vol. 17(15). – p. 6369-6382.
87. Parmley J. L., Chamary J. V., Hurst L. D. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. Mol Biol Evol. – 2006. – Vol. 23. – p. 301-309.
88. Zhou T., Weems M., Wilke C. O. Translationally optimal codons associate with structurally sensitive sites in proteins. // Mol Biol Evol. – 2009. - Vol. 26(7). – p. 1571-1580.
89. Hershberg R., Petrov D.A. Selection on Codon Bias // Annu. Rev. Genet. – 2008. – Vol. 42. – p. 287–299.
90. Andreeva A., Murzin A. G. Evolution of protein fold in the presence of functional constraints // Curr Opin Struct Biol. – 2006. – Vol. 16(3). – p. 399-408.
91. Hadley C., Jones D. T. A systematic comparison of protein structure classifications: SCOP, CATH, FSSP // Structure. – 1999. – Vol. 7. – p. 1099-1112.
92. Apic G., Gough J., Teichmann S.A. An insight into domain combinations // Bioinformatics. – 2001. – Vol. 17. Suppl. 1. – p. S83-S89.
93. Ejima Y., Yang L. Trans mobilization of genomic DNA as a mechanism for retrotransposon-mediated exon shuffling // Hum Mol Genet. – 2003. – Vol. 12(11). – p. 1321-1328.
94. Vogel C., Teichmann S.A., Pereira-Leal J. The relationship between domain duplication and recombination // J Mol Biol. – 2005. – Vol. 346(1). – p. 355-365.

95. van Rijk A., Bloemendal H. Molecular mechanisms of exon shuffling: illegitimate recombination // Genetica. – 2003. – Vol. 118(2-3). – p. 245-249.
96. Vibranovski M. D., Sakabe N. J., de Oliveira R. S., de Souza S. J. Signs of Ancient and Modern Exon-Shuffling Are Correlated to the Distribution of Ancient and Modern Domains Along Proteins // J Mol Evol. – 2005. - Vol. 61. – p. 341-350.
97. Liu M., Walch H., Wu Sh., Grigoriev A. Significant expansion of exon-bordering protein domains during animal proteome evolution // Nucleic Acids Res. – 2005. – Vol. 33(1). – p. 95-105.
98. Vivek G., Tan T. W., Ranganathan S. XdomView: protein domain and exon position visualization // Bioinformatics. – 2003. – Vol. 19. – p. 159-160.
99. Bhasi A., Philip P., Manikandan V., Senapathy P. ExDom: an integrated database for comparative analysis of the exon-intron structures of protein domains in eukaryotes // Nucleic Acids Res. – 2009. – Vol. 37(Database issue). – p. D703-D711.
100. Leslin C. M., Abyzov A., Ilyin V. A. Structural exon database, SEDB, mapping exon boundaries on multiple protein structures // Bioinformatics. – 2004. – Vol. 20. – p. 1801-1803.
101. Siddiqui A.S., Dengler U., Barton G.J. 3Dee: a database of protein structural domains // Bioinformatics. – 2001. – Vol. 17. – p. 200-201.
102. Wang Y., Addess K.J., Geer L., Madej T., Marchler-Bauer A., Zimmerman D., Bryant S.H. MMDB: 3D structure data in Entrez // Nucleic Acids Res., 2000. – Vol. 28. – p. 243-245.
103. Sakharkar M., Passetti F., de Souza J.E., Long M., de Souza S.J. ExInt an exon/intron database // Nucleic Acids Res. – 2002. – Vol. 30. – p. 191–194.
104. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D. J. Gapped BLAST and PSIBLAST: a new generation of protein database search programs // Nucleic Acids Res. – 1997. – Vol. 25. – p. 3389–3402.

105. Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J., Rapp B.A., Wheeler D.L. GenBank // Nucleic Acids Res. - 2002. - Vol. 30. – p. 7-20.

106. Shindyalov I.N., Bourne P.E. A database and tools for 3-D protein structure comparison and alignment using the combinatorial extension (CE) algorithm. Nucleic Acids Res. – 2001. – Vol. 29. – p. 228-229.

107. Saxonov S., Daizadeh I., Fedorov A., Gilbert W. EID: the exon–intron database–an exhaustive database of protein-coding intron-containing genes // Nucleic Acids Res. – 2000. – Vol. 28. – p. 185-190.

108. Birney E., Andrews T. D., Bevan P., Caccamo M., Chen Y., Clarke L., Coates G., Cuff J., Curwen V., Cutts T., Down T., Eyras E., Fernandez-Suarez X. M., Gane P., Gibbins B., Gilbert J., Hammond M., Hotz H.-R., Iyer V., Jekosch K., Kahari A., Kasprzyk A., Keefe D., Keenan S., Lehvaslaiho H., McVicker G., Melsopp C., Meidl P., Mongin E., Pettett R., Potter S., Proctor G., Rae M., Searle S., Slater G., Smedley D., Smith J., Spooner W., Stabenau A., Stalker J., Storey R., Ureta-Vidal A., Woodwark K. C., Cameron G., Durbin R., Cox A., Hubbard T., Clamp M.. An Overview of Ensembl // Genome Research. – 2004. – Vol. 14(5). – p. 925-928.

109. Flicek P., Aken B.L., Ballester B., Beal K., Bragin E., Brent S., Chen Y., Clapham P., Coates G., Fairley S., Fitzgerald S., Fernandez-Banet J., Gordon L., Graf S., Haider S., Hammond M., Howe K., Jenkinson A., Johnson N., Kahari A., Keefe D., Keenan S., Kinsella R., Kokocinski F., Koscielny G., Kulesha E., Lawson D., Longden I., Massingham T., McLaren W., Megy K., Overduin B., Pritchard B., Rios D., Ruffier M., Schuster M., Slater G., Smedley D., Spudich G., Tang Y. A., Trevanion S., Vilella A., Vogel J., White S., Wilder S.P., Zadissa A., Birney E., Cunningham F., Dunham I., Durbin R., Fernandez-Suarez X.M., Herrero J., Hubbard T.J.P., Parker A., Proctor G., Smith J., Searle S.M.J. Ensembl's 10th year // Nucleic Acids Research. – 2010. – Vol. 38(Database issue). – p. D557-D562.

110. Hunter S., Apweiler R., Attwood T. K., Bairoch A., Bateman A., Binns D., Bork P., Das U., Daugherty L., Duquenne L., Finn R.D., Gough J., Haft D.,

Hulo N., Kahn D., Kelly E., Laugraud A., Letunic I., Lonsdale D., Lopez R., Madera M., Maslen J., McAnulla C., McDowall J., Mistry J., Mitchell A., Mulder N., Natale D., Orengo C., Quinn A.F., Selengut J.D., Sigrist C.J., Thimma M., Thomas P.D., Valentin F., Wilson D., Wu C.H., Yeats C. InterPro: the integrative protein signature database // Nucleic Acids Res. – 2009. – Vol. 37 (Database Issue). – p. D224-228

111. Schorderet D.F. Using OMIM (On-line Mendelian Inheritance in Man) as an expert system in medical genetics // Am J Med Genet. – 1991. – Vol. 39(3). – p. 278-284.
112. Velculescu V.E., Zhang L., Vogelstein B., Kinzler K.W. Serial Analysis Of Gene Expression // Science. – 1995. – Vol. 270. – p. 484-487.
113. Bairoch A., Boeckmann B., Ferro S., Gasteiger E. Swiss-Prot: juggling between evolution and stability // Brief. Bioinform. – 2004. – Vol. 5. – p. 39-55.
114. Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. Basic local alignment search tool // J. Mol. Biol. – 1990. – Vol. 215. – p. 403-410.
115. Основы биоинформатики. Учебное Пособие. Издание 2-е исправленное – М.: ФГОУ ВПО РГАУ – МСХА им. К.А. Тимирязева, 2013. - 120 с.
116. Mount D.W. Bioinformatics. – Cold Spring Harbor, 2004. – 665 p.
117. Thompson J.D., Higgins D.G., Gibson T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice // Nucleic Acids Res. – 1994. – Vol. 22. – p. 4673-4680.
118. Saitou N., Nei. M. The neighbor-joining method: a new method for reconstructing phylogenetic trees // Mol. Evol. Biol. – 1987. – Vol. 4(4). – p. 406-425
119. Sneath P.H.A., Sokal R.R. Numerical taxonomy - the principles and practice of numerical classification. – San Francisco, 1973 – 573 p.

120. Krissinel E., Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions // Acta Cryst. – 2004. – Vol. D60. – p. 2256-2268.
121. Ivanisenko V.A., Pintus S.S., Grigorovich D.A., Kolchanov N.A. PDBSiteScan: a program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins // Nucleic Acids Res. – 2004. – Vol. 32. – p. W549–W554.
122. Jenkins J.L., Tanner J.J. High-resolution structure of human D-glyceraldehyde-3-phosphate dehydrogenase // Acta Crystallogr D Biol Crystallogr. – 2006. – Vol. 62(Pt 3). – p. 290-301.
123. Sofer W., Martin P.F. Analysis of alcohol dehydrogenase gene expression in Drosophila // Annual Review of Genetics. – 1987. – Vol. 21. – p. 203-25
124. Jairama S., Edenberg H. J. An enhancer-blocking element regulates the cell-specific expression of alcohol dehydrogenase 7 // Gene. – 2014. – http://dx.doi.org/10.1016/j.gene.2014.06.047.
125. Chambers G.K. The Drosophila alcohol dehydrogenase gene-enzyme system // Adv. Genet. - 1988. – Vol. 25. – p. 40-107.
126. Gaston D., Roger A.J. Functional divergence and convergent evolution in the plastid-targeted glyceraldehyde-3-phosphate dehydrogenases of diverse eukaryotic algae // PLoS One. – 2013. – Vol. 8(7). – p. e70396.
127. Fan J., Liu Q., Hao Q., Teng M., Niu L. Crystal structure of uroporphyrinogen decarboxylase from Bacillus subtili // J Bacteriol. – 2007. – Vol. 189(9). – p. 3573-3580.
128. Zhou M., Guo J., Cha J., Chae M., Chen S., Barral J.M., Sachs M.S., Liu Y. Non-optimal codon usage affects expression, structure and function of clock protein FRQ // Nature. – 2013. – Vol. 495(7439). – p. 111-115.
129. Trifonov E.N. Elucidating sequence codes: three codes for evolution // Ann N Y Acad Sci. – 1999. – Vol. 870. – p. 330-338.

130. Trifonov E.N., Volkovich Z., Frenkel Z.M. Multiple levels of meaning in DNA sequences, and one more // Ann N Y Acad Sci. – 2012. – Vol. 1267. – p. 35-38.

Приложения

Приложение 1

Физико-химические характеристики аминокислотных остатков в составе белков, согласно [88]

А.к.	Наличие		число	Диполь/заряд	pK	Яркая тенденция быть в:							
ост.	NH	Cβ	γ			до	спираль			за		клубок	ядро
						α_N	$\|\alpha_N$	α	$\alpha_C\|$	α_C	β		
Gly	√							-			-	+	
Ala	√	√						+				-	
Pro		√	1				+	-	-	-	-	+	
Glu	√	√	1	$COOH \Rightarrow CO_2^-$	4.3	+	+		-	-	-		-
Asp	√	√	1	$COOH \Rightarrow CO_2^-$	3.9	+	+	-	-	-	-	+	-
Gln	√	√	1	$OCNH_2$									-
Asn	√	√	1	$OCNH_2$		+		-		+	-	+	-
Ser	√	√	1	OH		+						+	
His	√	√	1	NH; и N $\Rightarrow$ NH^+	6.5		-		+	+			
Lys	√	√	1	$NH_2 \Rightarrow NH_3^+$	10.5	-	-		+	+	-		-
Arg	√	√	1	$HNC(NH_2)_2^+$	12.5	-	-		+	+	-	+	-
Thr	√	√	2	OH		+					+		
Ile	√	√	2								+	-	+
Val	√	√	2								+	-	+
Leu	√	√	1					+			+	-	+
Met	√	√	1					+			+	-	+
Phe	√	√	1								+	-	+
Tyr	√	√	1	$OH \Rightarrow O^-$	10.1			-			+		+
Cys	√	√	1	$SH \Rightarrow S^-$	9.2			-			+		+
Trp	√	√	1	NH							+		+

Приложение 2

Список ключевых, по которым проводилась автоматическая классификация белков и классификация функциональных сайтов, содержащихся в БД SitEx.

Таблица 1. Ключевые слова для классификации белков.

Наименование	Ключевые слова
Белки мышц	actin, interacting, kinesin, myosin, dynein, muscle
Белки крови	anticoagulant, blood, erythrocyte, hemoglobin
Белки клеточного цикла	cyclin,cell cycle,apoptosis,cell death,cell,apoptotic
Ферменты (с EC-номером, киназы, синтазы)	EC,kinase,synthetase,synthase
Белки имунной системы	antigen, T-cell, B-cell, ig, immunoglobulin
Мембранные белки	membrane, channel
Рецепторы	receptor
Белки, участвующие в репликации, транскрипции, трансляции	transcription, translation, replication
Белки теплового шока	heat shock, chaperon
Транспортные белки	transport, transfer
Белки опухолей	cancer, melanoma, tumor, onco
Цинковые пальцы, RING пальцы	zinc finger, finger
Другие белки и предшественники	precursor

Таблица 2. Ключевые слова для классификации функциональных сайтов.

Наименование	Ключевые слова
Неизвестный лиганд	unknown
Ионы металлов	ion, ALF, MOO, NA
Анионы кислот	ACT, BCT, BR, CAC, F, IOD, SO3, SO4, PO3, PO4, AZI, NO3, CL, SCN, 2HP, CO3
Органические кислоты	Acid, CHT, MLI, MLT, OXL, TMA, OAA
Нуклеотидфосфаты	adenosine, guanosine, cytidine, uridine
Фосфосахара	glucose, mannose, galactose, fructose, fucose, maltose, pyranose, amylose, sucrose
Белки	protein, factor, ase
Аминокислоты и их соединения	alanine, cystein, glycine, histidine, leucine, lysine, methionine, asparagine, proline, glutamine, arginine, serine, threonine, valine, tryptophan, tyrosine, glutamate, asparate
Коферменты	enzyme
Спирты и их производные	ol
Атомы и неорганические соединения	atom, molecule
Амины и амиды	amine,amide
Порфирины	porphyrin
Более мелкие классы	-

Приложение 3

Таблица отражает частоты встречаемости кодонов в различных позициях и значения χ^2 при их сравнении. Так, наименования столбцов соответствуют следующим названиям:

A- однобуквенный код аминокислоты, которая кодируется кодоном из столбца **C**. Сумма частот встречаемости кодонов, кодирующих одну аминокислоту равна 1.

1 – участок последовательности между концами экзона, ограниченными пятью кодонами

2 – участок последовательности на 5'-конце, ограниченный пятью кодонами

3 – участок последовательности на 3'-конце, ограниченный пятью кодонами

4 – часть сайта, кодируемая между концами экзона, ограниченными пятью кодонами

5 – часть сайта, кодируемая на 5'-конце, ограниченном пятью кодонами

6 – часть сайта, кодируемая на 3'-конце, ограниченном пятью кодонами

В оставшихся столбцах приведены значения χ^2 при сравнении частот, которые указаны нижними индексами. Учитывалась поправка Бонферрони. При этом:

Значимые значения выделены жирным шрифтом.

Темно-серый цвет соответствует значениям с вероятностью отличия меньше 1%.

Светло-серый цвет соответствует значениям с вероятностью отличия меньше 5%.

Без цвета соответствует значениям с вероятностью отличия меньше 10%.

A	C	1	2	3	4	5	6	χ^2_{12}	χ^2_{25}	χ^2_{45}	χ^2_{13}	χ^2_{36}	χ^2_{46}
A	GCA	0.2342	0.1736	0.2695	0.2584	0.25	0.2183	**11.1**	**13.7**	0.99	0.46	2.55	0
A	GCC	0.3997	0.3131	0.4473	0.4091	0.3476	0.3604						
A	GCG	0.091	0.3067	0.0547	0.0922	0.0793	0.0711						
A	GCT	0.2751	0.2066	0.2285	0.2403	0.3232	0.3503						
C	TGC	0.5442	0.68	0.5873	0.508	0.4098	0.538	**2.82**	**12.6**	1.23	0.1	0.17	0
C	TGT	0.4558	0.32	0.4127	0.492	0.5902	0.462						
D	GAC	0.5303	0.612	0.566	0.5757	0.531	0.53	**3.52**	**3.48**	0.2	0	0	0
D	GAT	0.4697	0.388	0.434	0.4243	0.469	0.47						
E	GAA	0.434	0.3057	0.463	0.4064	0.4807	0.417	2.51	**5.03**	0.6	0	0	0
E	GAG	0.566	0.6943	0.537	0.5936	0.5193	0.583						
F	TTC	0.5225	0.5612	0.5938	0.5458	0.435	0.5088	0.07	2.25	1.65	0	0	0
F	TTT	0.4775	0.4388	0.4063	0.4542	0.565	0.4912						
G	GGA	0.2609	0.2078	0.2749	0.2712	0.2507	0.3112	0.69	**9.78**	2.94	0	3.29	1.39
G	GGC	0.3363	0.3627	0.3578	0.3534	0.3207	0.27						
G	GGG	0.2346	0.298	0.2322	0.2026	0.1341	0.1739						
G	GGT	0.1682	0.1314	0.1351	0.1727	0.2945	0.2449						
H	CAC	0.5703	0.6226	0.5848	0.5616	0.4848	0.4764	0.22	**2.8**	0.65	0.01	1.57	0.85
H	CAT	0.4297	0.3774	0.4152	0.4384	0.5152	0.5236						
I	ATA	0.1597	0.2118	0.1633	0.1779	0.1925	0.1802	0.32	0.32	1.57	0	0	0
I	ATC	0.4689	0.4294	0.4728	0.4922	0.3851	0.4826						
I	ATT	0.3714	0.3588	0.3639	0.33	0.4224	0.3372						
K	AAA	0.4284	0.3581	0.4568	0.3755	0.487	0.4048	0.52	2.42	0	0	0.21	0
K	AAG	0.5716	0.6419	0.5432	0.6245	0.513	0.5952						

L	CTA	0.0732	0.0595	0.0699	0.0602	0.04	0.0905	0.39	4.2	1.81	1	2.34	4.53
L	CTC	0.1851	0.2194	0.2389	0.2069	0.1644	0.2672						
L	CTG	0.3949	0.4541	0.3329	0.4063	0.3689	0.2457						
L	CTT	0.1359	0.0986	0.136	0.1209	0.0889	0.1293						
L	TTA	0.0777	0.0561	0.0915	0.0705	0.12	0.0517						
L	TTG	0.1331	0.1122	0.1309	0.1352	0.2178	0.2155						
M	ATG	1	1	1	1	1	1						
N	AAC	0.524	0.5663	0.5728	0.5556	0.4015	0.6256	0.1	**4.2**	**3.6**	0.17	0.22	0
N	AAT	0.476	0.4337	0.4272	0.4444	0.5985	0.3744						
P	CCA	0.2828	0.2095	0.2542	0.2691	0.5	0.2677	5.7	**16**	**11.5**	0.26	4.62	5.23
P	CCC	0.3191	0.3261	0.3626	0.3745	0.1977	0.2126						
P	CCG	0.1035	0.2484	0.114	0.0749	0.1163	0.126						
P	CCT	0.2946	0.216	0.2692	0.2816	0.186	0.3937						
Q	CAA	0.2686	0.1949	0.2758	0.2511	0.4022	0.1365	0.81	**8.38**	**3.9**	0	**4.35**	**2.86**
Q	CAG	0.7314	0.8051	0.7242	0.7489	0.5978	0.8635						
R	AGA	0.2131	0.1792	0.2515	0.1887	0.2036	0.3948	0.52	7.51	2.28	0.15	4.14	13.1
R	AGG	0.1982	0.2542	0.2206	0.1658	0.1127	0.1844						
R	CGA	0.1185	0.092	0.1381	0.1226	0.1964	0.1354						
R	CGC	0.1792	0.1671	0.1546	0.2256	0.1418	0.0778						
R	CGG	0.2041	0.247	0.1773	0.2056	0.2218	0.1816						
R	CGT	0.0869	0.0605	0.0577	0.0917	0.1236	0.0259						
S	AGC	0.2377	0.2473	0.259	0.2308	0.1697	0.254	2.37	9.95	3.09	0.08	0.18	0.61
S	AGT	0.1575	0.1126	0.1364	0.1726	0.2202	0.1415						
S	TCA	0.1532	0.1071	0.1377	0.1465	0.1661	0.1158						
S	TCC	0.2119	0.2321	0.2397	0.2255	0.1444	0.2154						
S	TCG	0.0522	0.1305	0.0262	0.0567	0.0325	0.0418						
S	TCT	0.1876	0.1703	0.2011	0.1678	0.2671	0.2315						
T	ACA	0.2867	0.2514	0.2776	0.2644	0.2489	0.3717	0	5.19	1.5	0	1.44	2.52
T	ACC	0.3513	0.4086	0.3435	0.3477	0.2574	0.267						
T	ACG	0.1107	0.1429	0.0965	0.1372	0.1603	0.0785						
T	ACT	0.2514	0.1971	0.2824	0.2506	0.3333	0.2827						
V	GTA	0.1164	0.078	0.1329	0.1051	0.1237	0.1111	1.5	5.4	0	0	**11.2**	8.71
V	GTC	0.234	0.2168	0.2549	0.2624	0.2062	0.3519						
V	GTG	0.4608	0.5751	0.4314	0.4178	0.4124	0.2037						
V	GTT	0.1888	0.1301	0.1808	0.2147	0.2577	0.3333						
W	TGG	1	1	1	1	1	1						
Y	TAC	0.5583	0.561	0.6389	0.5589	0.4381	0.528	0	2.12	2.03	0.76	1.7	0
Y	TAT	0.4417	0.439	0.3611	0.4411	0.5619	0.472						

Приложение 4

Таблица отражает встречаемость либо вблизи 5’-конца экзона (расстояние до 5 кодонов включительно), либо встречаемость кодонов, кодирующих аминокислоты функционального сайта. Цветом отмечены те значения отношения встречаемостей, которые показывают значения больше, нежели встречаемость данного кодона в сайте. Встречаемость кодонов на участке последовательности рассчитывалась как отношение использования кодонов во фрагментах ДНК, кодирующих функциональные сайты, по отношению к встречаемости их в рассматриваемых участках последовательностей.

Светло-серым цветом отмечены те кодоны из них, которые AT-богаты. Темно-серым – AG-богатые.

кодон	встречаемость вблизи 5-конца	встречаемость в сайте	отношение
TGT	8.97	2.15	4.17
TAT	6.04	1.73	3.5
CAT	5.93	2.4	2.47
AAT	4.46	1.41	3.17
TAC	3.68	1.66	2.21
GAT	3.58	1.46	2.45
CAC	3.38	2.19	1.54
GGT	3.01	1.27	2.37
TGC	2.93	1.84	1.59
CGA	2.83	1.72	1.65
CGT	2.71	1.6	1.7
TTT	2.62	1.24	2.11
GAC	2.57	1.68	1.53
AAC	2.29	1.47	1.56
ACT	2.28	0.96	2.38
ATT	2.22	0.64	3.45
GTT	2.22	0.71	3.13
GTA	1.77	0.52	3.37
ATA	1.72	0.73	2.34
AAA	1.72	0.87	1.97
ATC	1.69	0.68	2.5
TTA	1.63	0.58	2.81
GGA	1.62	1.16	1.39
CAA	1.6	0.73	2.19
GAA	1.59	0.82	1.93
TTC	1.58	1.26	1.25
ACG	1.52	1.11	1.37
AGA	1.51	1.47	1.03
TTG	1.48	0.66	2.23
AGT	1.48	0.96	1.55
ACA	1.34	0.86	1.55
GGC	1.19	1.12	1.06
CGG	1.19	1.55	0.77

TCT	1.19	0.86	1.39
TCA	1.18	0.8	1.48
CGC	1.13	1.8	0.63
GTC	1.06	0.64	1.65
AAG	1.01	1.04	0.97
CCA	0.88	0.39	2.26
ACC	0.85	0.86	0.98
GTG	0.8	0.5	1.6
GAG	0.76	0.91	0.83
CTT	0.69	0.52	1.32
CTG	0.62	0.59	1.04
GGG	0.6	0.9	0.67
AGG	0.59	1.27	0.46
CAG	0.58	0.84	0.69
CTC	0.57	0.65	0.88
GCT	0.54	0.56	0.96
AGC	0.52	0.83	0.63
CTA	0.51	0.51	1
GCA	0.5	0.65	0.77
TCC	0.47	0.88	0.53
GCC	0.39	0.6	0.65
CCT	0.32	0.38	0.84
CCC	0.22	0.42	0.53
TCG	0.19	0.86	0.22
CCG	0.17	0.31	0.54
GCG	0.09	0.58	0.16

Приложение 5

Частота фаз экзонов, которые кодируют на 5'-конце аминокислоту функционального сайта, (сверху) и среди остальных (снизу) в 11 организмах.

Первая колонка – фазы.

Вторая колонка - абсолютное количество экзонов.

Третья колонка - относительное количество экзонов.

Bos taurus			Canis familiaris			Cavia porcellus		
0	39	0.433333	0	0	0	0	1	0.333333
1	30	0.333333	1	3	0.75	1	2	0.666667
2	21	0.233333	2	1	0.25	2	0	0
0	594	0.465153	0	54	0.545455	0	2	0.222222
1	439	0.343774	1	23	0.232323	1	4	0.444444
2	244	0.191073	2	22	0.222222	2	3	0.333333

Danio rerio			Equus caballus			Gallus gallus		
0	3	0.428571	0	2	0.666667	0	4	0.181818
1	2	0.285714	1	0	0	1	13	0.590909
2	2	0.285714	2	1	0.333333	2	5	0.227273
0	61	0.525862	0	11	0.458333	0	224	0.532067
1	30	0.258621	1	8	0.333333	1	129	0.306413
2	25	0.215517	2	5	0.208333	2	68	0.16152

Homo sapiens			Mus musculus			Oryctolagus cuniculus		
0	554	0.388499	0	54	0.377622	0	13	0.590909
1	493	0.345722	1	49	0.342657	1	7	0.318182
2	379	0.265778	2	40	0.27972	2	2	0.090909
0	20840	0.445271	0	3484	0.470684	0	192	0.520325
1	16287	0.347991	1	2407	0.325182	1	120	0.325203
2	9676	0.206739	2	1511	0.204134	2	57	0.154472

Rattus norvegicus			Sus scrofa		
0	63	0.456522	0	10	0.5
1	47	0.34058	1	7	0.35
2	28	0.202899	2	3	0.15
0	1530	0.484791	0	184	0.511111
1	1006	0.318758	1	106	0.294444
2	620	0.196451	2	70	0.194444