

На правах рукописи

Нуртдинов Рамиль Наилевич

ЭВОЛЮЦИЯ АЛЬТЕРНАТИВНОГО СПЛАЙСИНГА ГЕНОВ
МЛЕКОПИТАЮЩИХ

03.00.28 Биоинформатика

Автореферат диссертации на соискание ученой степени
кандидата физико-математических наук

Москва – 2008

Работа выполнена на Факультете биоинженерии и биоинформатики Московского Государственного Университета имени М.В. Ломоносова

Научный руководитель:

доктор биологических наук, кандидат физико-математических наук,
профессор Миронов Андрей Александрович

Официальные оппоненты:

доктор биологических наук Карягина Анна Станиславовна
ГУ НИИ эпидемиологии и микробиологии имени Н.Ф. Гамалеи

кандидат физико-математических наук Макеев Всеволод Юрьевич
ФГУП ГНЦ РФ «ГосНИИгенетика»

Ведущая организация:

Учреждение Российской Академии Наук Институт общей генетики им.
Н.И. Вавилова РАН

Защита диссертации состоится ____ ноября 2008 года в ____ часов на заседании диссертационного совета Д002.077.02 при Учреждении Российской академии наук Институт проблем передачи информации им. А.А. Харкевича РАН по адресу: 127994, Москва, ГСП-4, Большой Каретный переулок, д. 19, стр. 1.

С диссертацией можно ознакомиться в библиотеке Учреждения Российской академии наук Институт проблем передачи информации им. А.А. Харкевича РАН.

Автореферат разослан _____ 2008 года.

Ученый секретарь диссертационного совета

доктор биологических наук, профессор

Рожкова Г.И.

Общая характеристика работы

Актуальность темы

Характерной особенностью эукариотических генов является существование механизма вырезания из первичного транскрипта пре-мРНК протяженных участков, называемых интронами. Оставшиеся участки, экзоны, сшиваются, и получаемая мРНК впоследствии используется как матрица для синтеза белка. Процесс вырезания интрона и сшивки экзонов называется сплайсинг. Каждое событие сплайсинга вырезает один интрон и, как правило, сплайсинг интрона происходит независимо от сплайсинга остальных интронов. Для многих генов характерен альтернативный сплайсинг – процесс неоднозначного вырезания интронов. В результате один ген может кодировать несколько разных белков, которые могут иметь как сходные функции, так и сильно отличаться друг от друга. Согласно современным представлениям, более половины генов млекопитающих подвержены альтернативному сплайсингу.

Для предсказания экзон-интронной структуры гена и возможного альтернативного сплайсинга используются два класса методов: статистические методы и методы, основанные на выравнивании. Статистические методы используют статистики длин экзонов и интронов, их аминокислотный (после формальной трансляции) и нуклеотидный состав, частоты нуклеотидов в сайтах сплайсинга. Недостатком статистических методов является недостаточная точность предсказания и предсказание сильно ограниченного множества из возможных схем альтернативного сплайсинга.

Другим классом методов является выравнивание нуклеотидной последовательности уже прошедших сплайсинг мРНК с последовательностью геномной ДНК. Для предсказания альтернативного сплайсинга необходим очень большой объем таких данных. Обычно ген имеет одну, базовую, схему вырезания интронов. Альтернативные варианты белка порождаются реже, или, в случае регулируемого альтернативного сплайсинга,

достигают уровня экспрессии базового варианта в ограниченном наборе тканей или в течение ограниченного промежутка времени.

Уровень современных технологий секвенирования мРНК не позволяет получать в большом количестве мРНК для конкретных генов, поскольку для этого требуется большой объем ручного труда. Сейчас разработаны методики массового секвенирования, позволяющие в автоматическом режиме быстрое, качественное и достаточно полное секвенирование всей клеточной мРНК. Это позволяет за счет объема данных получить адекватную информацию об экспрессии большинства генов. Получаемые нуклеотидные последовательности мРНК были названы EST (Expressed Sequence Tag). Преимуществом EST является их массовость, дающая возможность получать данные об экспрессии генов в разных тканях и на разных стадиях эмбрионального и постэмбрионального развития. Массовость EST одновременно является и их недостатком, поскольку вероятность секвенирования мРНК для определенного гена зависит от уровня его экспрессии, что дискриминирует гены со слабым уровнем экспрессии.

Как и любой другой клеточный процесс, сплайсинг подвержен ошибкам. В общем случае нет четкой границы между ошибками сплайсинга и альтернативным сплайсингом. Однако большинство ошибок сплайсинга приводят к образованию мРНК, кодирующих нефункциональные короткие варианты белка из-за нарушения рамки считывания вследствие ее сдвига или вставки в мРНК последовательностей содержащих терминирующие кодоны.

Использование данных EST является основным методом изучения альтернативного сплайсинга. Анализ большого количества последовательностей (около восьми миллионов EST человека) невозможен без использования современных компьютерных технологий, начиная с выравнивания последовательностей и хранения полученных экзон-интронных структур в базе данных, и заканчивая алгоритмами анализа графов для выявления альтернативного сплайсинга.

Цели и задачи исследования

Целью данной работы является исследование альтернативного сплайсинга и его эволюции, с использованием биоинформатических методов анализа EST данных.

При этом решались следующие **задачи**:

- Предварительная оценка консервативности альтернативного сплайсинга.
- Создание базы данных для альтернативно сплайсируемых генов.
- Оценка достоверности наблюдаемых альтернатив.
- Определение свойств альтернатив, таких как частота встречаемости, влияние на аминокислотную последовательность белка, функциональность белка, кодируемого альтернативным вариантом мРНК.
- Определение консервативности альтернатив в зависимости от их свойств.
- Выявление закономерностей эволюции альтернативного сплайсинга.

Научная новизна и практическое значение

- Впервые оценена доля неконсервативного альтернативного сплайсинга генов млекопитающих.
- Создана база данных альтернативно сплайсируемых генов EDAS.
- Разработан метод анализа выделения сложных и элементарных альтернатив с использованием алгоритмов анализа графов сплайсинга.
- Реализован метод оценки достоверности наблюдаемых альтернатив.
- Проведен анализ консервативности элементарных событий альтернативного сплайсинга человека в геномах мыши и собаки.
- Проведен анализ консервативности элементарных событий альтернативного сплайсинга мыши в геномах человека, собаки и крысы.
- Оценена доля эволюционных новоприобретений в ходе эволюции грызунов для групп генов с разными скоростями молекулярной эволюции.

Апробация работы

Результаты работы были представлены на международных конференциях:

- ✓ Third International Conference on Bioinformatics of Genome Regulation and Structure. (BGRS'2002), Новосибирск, 2002;
- ✓ First Moscow Conference on Computational Molecular Biology (MCCMB'03, Москва, 2003).
- ✓ Second Moscow Conference on Computational Molecular Biology (MCCMB'05, Москва, 2005);
- ✓ Meeting of HHMI International Research Scholars (Ashburn, Virginia, USA, 2006);
- ✓ а также на конференции Информационные технологии и системы (ИТиС'07, Звенигород, 2007).

По материалам диссертации опубликовано 9 работ в рецензируемых российских и международных журналах.

Объем и структура диссертации

Диссертация изложена на 104 страницах и состоит из введения и 3 глав. Глава 1 содержит обзор литературы, в главах 2-3 представлены оригинальные результаты. Список литературы, приведенный в конце диссертации, содержит 127 наименований. Работа содержит 22 рисунка и 6 таблиц.

Содержание работы

Глава 1. Предварительная оценка консервативности альтернативного сплайсинга

Первоначально была исследована консервативность альтернативно сплайсируемых участков человека в ортологичных генах мыши и, наоборот, консервативность альтернативно сплайсируемых участков мыши в ортологичных генах человека. Работа была выполнена на небольшой группе ортологичных генов (166 пар). Альтернативный сплайсинг для исследуемых генов определялся выравниванием последовательностей мРНК и EST из GenBank (Wheeler DL at al 2003) и двух баз данных альтернативно сплайсированных генов AsMamDB (Ji H at al 2001) и HASDB (Modrek B at al 2001). Полученные выравнивания просматривались вручную, выделялись и классифицировались элементарные альтернативы. Для 84 пар генов альтернативный сплайсинг был известен у обоих организмов, для 42 пар генов альтернативный сплайсинг был найден только у гена человека, для 40 пар генов только у мыши.

Таблица 1. Консервативность альтернативного сплайсинга по данным полноразмерных мРНК и данным EST. **К** – консервативные альтернативы, **НК** – неконсервативные альтернативы.

Тип альтернативы	Человек				Мышь			
	мРНК		EST		мРНК		EST	
	К	НК	К	НК	К	НК	К	НК
Кассетный экзон	56	25	74	26	70	5	39	9
Альт. донорный сайт	18	7	16	10	24	6	17	6
Альт. акцепторный сайт	13	5	19	15	15	6	16	9
Удерживаемый интрон	4	3	5	0	8	7	10	4
Всего генов	45	28	41	44	68	22	30	26

Консервативность исследуемых элементарных альтернатив приведена в Таблице 1. Альтернатива человека считалась неконсервативной, если соответствующий участок отсутствует в геноме мыши. Анализ консервативности показал, что приблизительно 30% альтернатив человека не консервативны в мыши, 20% альтернатив мыши не консервативны в человеке. Около 50% генов человека имеют неконсервативные в геноме

мышь альтернативы, и около 40% генов мышь имеют неконсервативные в геноме человека альтернативы. Недостаточно большая выборка не позволила нам учесть в анализе несколько важных свойств элементарных альтернатив, таких как частота наблюдения альтернативного варианта мРНК и способность кодирования этим вариантом функционального варианта белка. Дальнейшие исследования альтернативного сплайсинга генов человека и мышь, с использованием всей доступной мРНК, EST и белковой информации, потребовали существенной автоматизации процесса выделения, хранения и анализа событий альтернативного сплайсинга.

Глава 2. Создание базы данных альтернативно сплайсируемых генов человека EDAS

Был создан набор программных средств, позволяющий с минимальным участием пользователя производить отбор EST, выравнивание их с соответствующим участком геномной последовательности, оценку качества полученных выравниваний и последующую обработку для получения информации о сплайсинге. Полученная информация о 20809 генах человека и 31811 генах мышь собрана в базу данных EDAS.

Для каждого гена на основе набора всех наблюдаемых экзонов и интронов был построен граф сплайсинга. Вершинами этого графа являются сайты сплайсинга, а ребрами – экзоны и интроны. Анализ графа сплайсинга позволяет выделить альтернативный сплайсинг. Среди всех возможных схем альтернативного сплайсинга выделялись четыре базовых схем альтернативного сплайсинга (Рисунок 1), т.е. элементарных альтернатив.

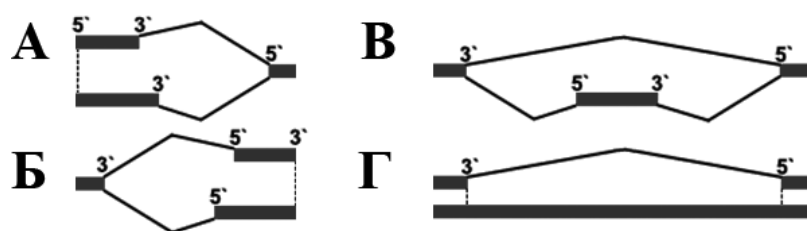


Рисунок 1. Элементарные альтернативы. Альтернативный донорный А и акцепторный Б сайты сплайсинга, кассетный экзон В и удерживаемый интрон Г.

Для каждой элементарной альтернативы определялся базовый вариант – путь, наблюдаемый в известной последовательности белка. Если оба варианта наблюдаются в белке, то базовым является путь, покрытый наибольшим числом EST, и, соответственно, второй путь объявлялся альтернативным. Если ни один из путей не наблюдается в белке, то альтернатива считалась не кодирующей и исключалась из дальнейшего рассмотрения.

В рамках общепризнанной модели ошибочного сплайсинга предполагается, что вероятность порождения ошибочного варианта мРНК имеет биномиальное распределение. Считается, что в рамках этой модели вероятностный параметр, биномиальная вероятность, может быть оценен значением 0.01 (Neverov AD et al 2005). В настоящем исследовании для каждой альтернативы оценивалась интегральная вероятность порождения ошибочной альтернативы, имеющей сходное EST-покрытие. Исследуемая альтернатива считалась ошибочной в случае превышения данной вероятностью значения 0.05.

С точки зрения последовательности мРНК, все альтернативные схемы вырезания интронов относительно базовой схемы можно разделить на две группы: альтернативы, приводящие к удалению участка мРНК, и альтернативы, приводящие к добавлению участка мРНК. Шкала, основанная только на частоте использования альтернативного варианта мРНК, не учитывает данное различие, поэтому она была видоизменена. За основу была взята частота более длинного (по длине нуклеотидной последовательности) варианта. Таким образом, по определению событие вставки нуклеотидов (вставка экзона или использование сайтов сплайсинга, находящихся вне экзона) имеет частоту в интервале 0-0.5. Событие удаления нуклеотидов (пропуск экзона или использование сайтов сплайсинга, находящихся внутри экзона) имеет частоту в интервале 0.5-1.

Глава 3. Статистический анализ альтернативных сайтов сплайсинга

Всего нами было наблюждено 6183 альтернативных акцепторных и 4468 альтернативных донорных сайтов. Был проведен статистический анализ появления альтернативного сайта сплайсинга вне или внутри экзона в зависимости от длины окружающих его экзона и интрона. Для анализа были отобраны 3902 альтернативных акцепторных и 2975 альтернативных донорных сайтов, находящихся на расстоянии 9 и более нуклеотидов от базового сайта. Для разных интервалов длин интрона исследовалось распределение доли удлинняющих или укорачивающих его альтернативных сайтов, по сравнению с интроном, получаемым в результате реализации базовой схемы сплайсинга (Рисунок 2).

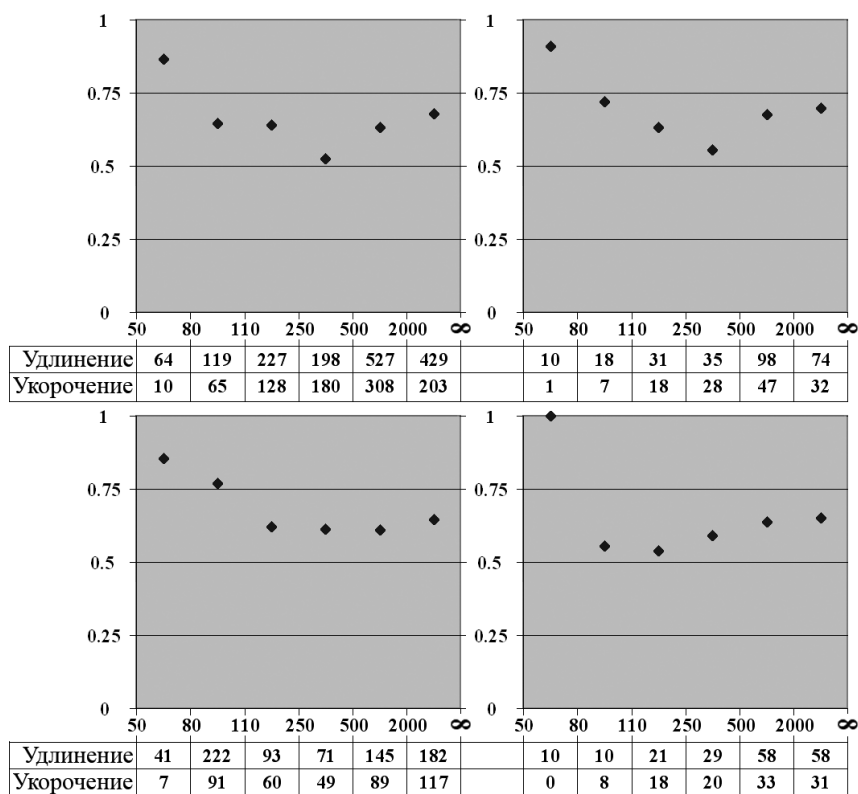


Рисунок 2. Доля удлинняющих интрон альтернативных акцепторных (вверху) и альтернативных донорных (внизу) сайтов сплайсинга генов человека, в зависимости от длины интрона. Слева редкие или нарушающие рамку считывания, справа частые и сохраняющие рамку считывания альтернативные сайты сплайсинга.

Как следует из Рисунка 2, короткие интроны преимущественно удлиняются при использовании альтернативных сайтов сплайсинга. Соотношение укорачивающих и удлиняющих интрон альтернативных акцепторных и донорных сайтов стабилизируется с ростом длины интрона и составляет приблизительно 60%.

Распределение альтернативных сайтов сплайсинга согласуется с моделью случайной фиксации сайтов. Для разных интервалов длин экзонов исследовалось распределение числа альтернативных акцепторных и альтернативных донорных сайтов, удлиняющих или укорачивающих его, по сравнению с экзоном, получаемым в результате реализации базовой схемы сплайсинга (Рисунок 3).

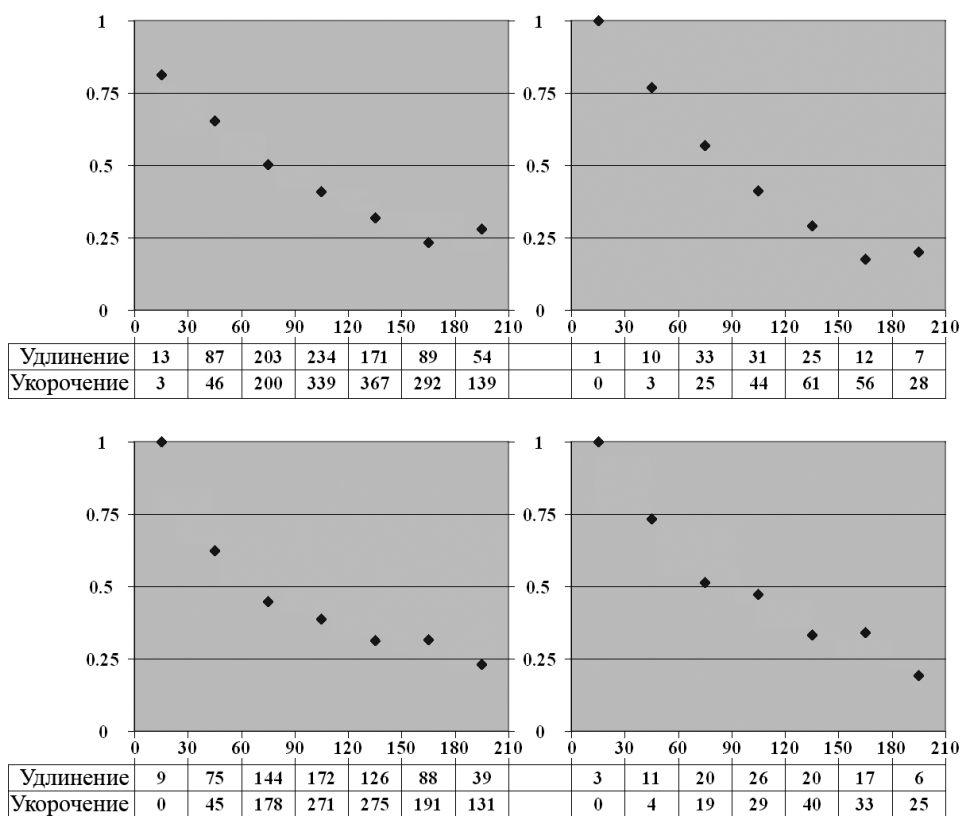


Рисунок 3. Доля удлиняющих экзон альтернативных акцепторных (вверху) и альтернативных донорных (внизу) сайтов сплайсинга генов человека, в зависимости от длины экзона. Слева редкие или нарушающие рамку считывания, справа частые и сохраняющие рамку считывания альтернативные сайты сплайсинга.

Для каждой группы альтернатив наблюдается сильная корреляция между преимущественным положением сайтов сплайсинга и длиной экзона. Альтернативные

сайты предпочитают удлинять короткие экзоны и укорачивать длинные. Для всех групп альтернативных сайтов равновесие между укорочением и удлинением экзона достигается приблизительно в районе 90 нуклеотидов, что соответствует моде распределения длин экзонов.

Такое поведение альтернативных сайтов сплайсинга хорошо согласуется с предположением о фиксации случайных сайтов. В самом деле, вероятность появления случайного сайта внутри экзона возрастает с ростом его длины, а укорочение очень короткого экзона маловероятно. Для проверки данной гипотезы была рассмотрена модель фиксации случайных сайтов. Мы предположили, что альтернативный сайт сплайсинга может с одинаковой вероятностью порождаться вследствие случайной мутации в любом месте относительно экзона, при условии, что его использование не приведет к нарушению рамки считывания. Тем самым, вероятность укорочения экзона путем появления внутри альтернативного акцепторного или донорного сайтов пропорциональна его длине, а вероятность удлинения экзона пропорциональна расстоянию до ближайших находящихся в рамке считывания терминирующих кодонов (Рисунок 4).



Рисунок 4. Интервал, в котором возможна фиксация случайного сайта сплайсинга.

Равновесие достигается в случае, если вероятности удлинения и укорочения станут равными между собой, что эквивалентно равенству длин экзона и длин сегментов слева и справа от него. Поэтому среднее расстояние до терминирующих кодонов должно быть равно средней длине экзона. Было смоделировано распределение расстояния от случайной точки в интроне до 5' и 3'терминирующих кодонов. Суммарное среднее расстояние до терминирующих кодонов составляло 138 нуклеотидов, что близко к средней длине экзона в 130 нуклеотидов.

Глава 4. Консервативность альтернативного сплайсинга

Для 10961 генов человека были получены ортологичные гены мыши и собаки из работы (Lindblad-Toh K et al 2005). Первоначально последовательности ДНК ортологичных генов были выравнены программой Blat (Kent WJ et al 2002). Результатом этого выравнивания является цепочка консервативных на аминокислотном уровне участков, формирующая набор пар сегментов ДНК. Границами этих сегментов служат хорошо выравнивающиеся кодирующие белок экзоны. Для каждого такого сегмента исследовалась возможность реализации всех содержащихся внутри него интронов человека в исследуемом гене мыши или собаки при помощи программы ProGene (Novichkov PS et al 2001). Результатом работы программы ProGene является набор пар ортологичных акцепторных и донорных сайтов. Внутренний кодирующий белок считался консервативным, если были определены ортологи для его акцепторного и донорного сайтов сплайсинга. Экзон, содержащий внутри себя терминирующий кодон, считался консервативным, если был найден ортолог его акцепторного сайта.

Таблица 2. Консервативность константных экзонов.

EST покрытие экзона	Всего экзонов	Консервативность (%)		Консервативность (штуки экзонов)			
		Мышь	Собака	Везде	Только в мыши	Только в собаке	Нигде
0 EST	3208	96,9%	97,0%	3042	65	69	32
1-10 EST	15482	97,0%	97,8%	14804	212	337	129
10-50 EST	2298	98,4%	98,7%	2239	23	28	8
Более 50 EST	719	99,4%	99,3%	711	4	3	1

Консервативность константных (не подверженных альтернативному сплайсингу) экзонов представлена в Таблице 2. Консервативность константных экзонов очень высока, доля неконсервативных экзонов не превышает 3%. Хотя мышь находится на эволюционно более близком расстоянии к человеку, чем собака, консервативность константных экзонов человека в геноме собаки немного выше, чем в геноме мыши. Это согласуется с известным наблюдением о повышенной скорости молекулярной эволюции геномов грызунов.

Консервативность константных экзонов повышается с ростом EST покрытия экзона. Это естественно, поскольку EST покрытие отражает уровень экспрессии генов, а высокоэкспрессирующиеся гены, как правило, эволюционируют медленней.

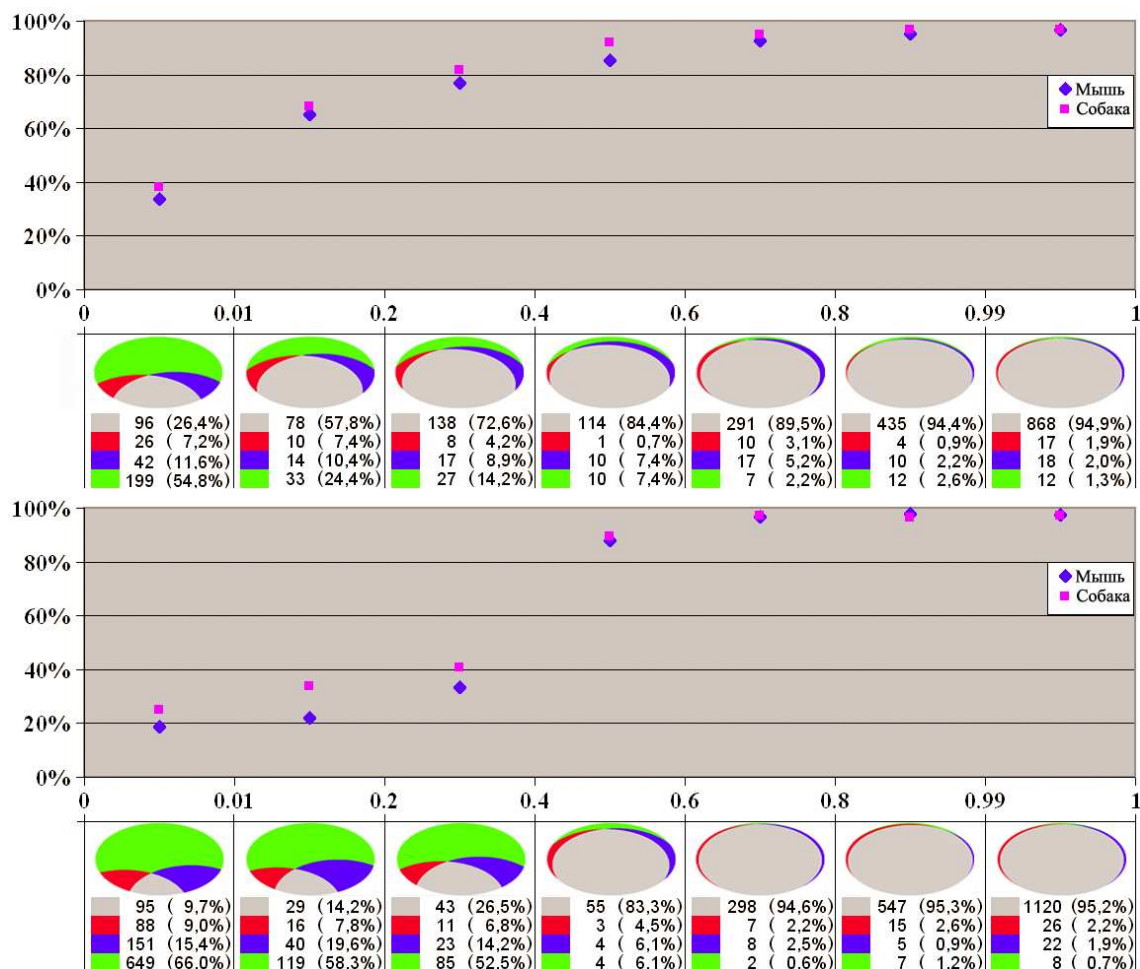


Рисунок 5. Консервативность сохраняющих (вверху) и нарушающих (внизу) рамку считывания кассетных экзонов в ортологичных генах человека, мыши и собаки. На диаграммах зеленым цветом показана доля не консервативных в мыши и в собаке альтернатив. Серым цветом показана доля полностью консервативных альтернатив. Красным цветом показана доля альтернатив консервативных только в мыши, а синим цветом – только в собаке.

Консервативность сохраняющих и нарушающих рамку считывания кассетных экзонов человека в геномах мыши и собаки показана на Рисунке 5. Консервативность кассетных экзонов повышается с ростом частоты включения экзона. Консервативность кассетных экзонов, нарушающих рамку считывания, в целом меньше, чем консервативность экзонов,

сохраняющих рамку считывания. Однако для экзонов, имеющих высокую частоту включения, разница в консервативности между сохраняющими и нарушающими рамку считывания экзонами крайне мала. Наблюдается достаточно высокая (35%) консервативность сохраняющих рамку считывания кассетных экзонов, имеющих крайне низкую частоту включения. Консервативность редко включаемых и при этом нарушающих рамку считывания кассетных экзонов (результата потенциальных ошибок или артефактов) менее высокая (24%), но не нулевая.

Для альтернативных сайтов сплайсинга существует некоторая эволюционная асимметрия между сдвигом сайта сплайсинга в экзон и сдвигом в интрон. Даже если внутренний сайт сплайсинга не используется для создания альтернативного варианта белка, он может быть консервативен из-за требования консервативности аминокислотной последовательности экзона. Эта асимметрия подтверждается анализом консервативности альтернативных сайтов сплайсинга. Внутренние альтернативные сайты сплайсинга существенно более консервативны, чем внешние. В остальном консервативность альтернативных сайтов сплайсинга сходна с консервативностью кассетных экзонов.

Альтернативно сплайсируемые участки генов человека, консервативные в только в геноме мыши, и альтернативно сплайсируемые участки, консервативные только в геноме собаки, являются эволюционными потерями в ходе эволюции соответственно собаки и мыши. Альтернативно сплайсируемые участки генов человека, неконсервативные ни в геноме мыши, ни в геноме собаки могут являться либо эволюционными приобретениями в ходе эволюции человека, либо ошибками сплайсинга, попавшими в EST и ошибочно трактованными как альтернативные варианты мРНК. Одна из возможных оценок доли эволюционных приобретений среди неконсервативных альтернативно сплайсируемых участков будет дана далее. Косвенным доказательством значительной доли таких альтернатив является заметная доля неконсервативных альтернатив, имеющих высокое EST-покрытие, а косвенным доказательством функциональности альтернатив, имеющих

крайне низкую, на уровне ошибки сплайсинга, частоту использования является их консервативность.

Аналогичное исследование было проведено для альтернативного сплайсинга генов мыши. Важность такого исследования заключается в том, что доступность родственного генома крысы позволяет оценить долю “молодых” (таксон-специфичных) элементарных альтернатив. При этом были отдельно рассмотрены гены, эволюционирующие с разной скоростью. В качестве выборки ортологичных генов были использованы данные Homologene (Wheeler DL et al 2003). Для 11014 генов мыши из EDAS были получены ортологичные гены человека, собаки и крысы.

Для определения скоростей молекулярной эволюции генов в каждом гене были выделены ортологичные экзоны и оценено их сходство. Для определения скорости молекулярной эволюции использовались только последовательности белков человека, собаки и мыши, а скорость молекулярной эволюции генов крысы принималась равной скорости эволюции генов мыши. Набор экзонов был получен выравниванием белковых последовательностей мыши, человека и собаки с соответствующей последовательностью генома. Было проведено попарное выравнивание кодируемых каждым экзоном аминокислотных последовательностей между разными видами (человек-мышь, человек-собака и собака-мышь). Ортологичные экзоны выявлялись при помощи метода **ВВН** (best bidirectional hit или лучшее двустороннее сопоставление), распространенного на тройки выравниваемых объектов. Для каждого экзона был найден максимально сходный аналог в двух других организмах, экзоны считались ортологичными, если такое сопоставление является самосогласованным и единственными для всех трех экзонов из разных организмов. Далее было отсеяно 911 кластера, для которых более четверти аминокислот мыши и человека не попали в набор ортологичных экзонов. Аминокислотные последовательности ортологичных экзонов были объединены в рамках соответствующих видов и выравнены друг с другом с использованием матрицы сходства **Blosum62**.

Полученное распределение уровня сходства белковых последовательностей показано на Рисунке 6.

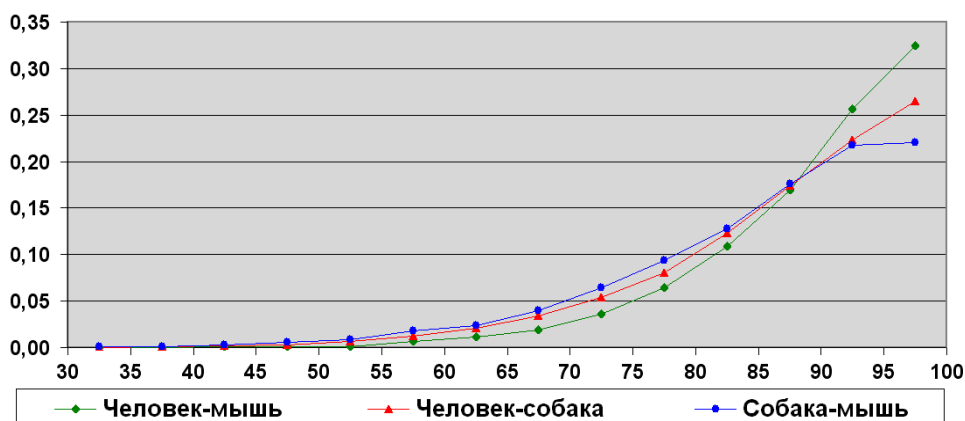


Рисунок 6. Распределение сходства белков человека собаки и мыши.

Расхождение видов мыши и крысы произошло относительно недавно, поэтому консервативность в геноме крысы элементарных альтернатив мыши может быть объяснена сходством ДНК последовательностей, не накопивших достаточного количества мутаций, а не функциональной значимостью альтернативы, защищающей ее от случайных мутаций. При исследовании кассетных экзонов для учета такой возможности мы провели анализ консервативности случайно отобранных участков интронов мыши моделирующих кассетные экзоны, 14448 участков без терминирующих кодонов внутри, и 10799 участков, содержащих терминирующие кодоны. Консервативность псевдоэкзонов зависит от их длины и наличия внутреннего терминирующего кодона, поэтому размер выборки псевдоэкзонов был подобран так, чтобы обеспечить покрытие в 100 и более штук на каждый 12 нуклеотидный интервал длины. Мы оценивали вероятность случайного сохранения в геноме крысы кассетного экзона как среднюю консервативность псевдоэкзонов, имеющих сходную длину. Количество случайно сохраненных экзонов оценивали как сумму соответствующих вероятностей для каждого не консервативного в геномах человека и собаки экзона. Доля “истинных” эволюционных новоприобретений вычислялась вычитанием полученной доли случайно консервативных экзонов из всех консервативных только в геноме крысы кассетных экзонов.

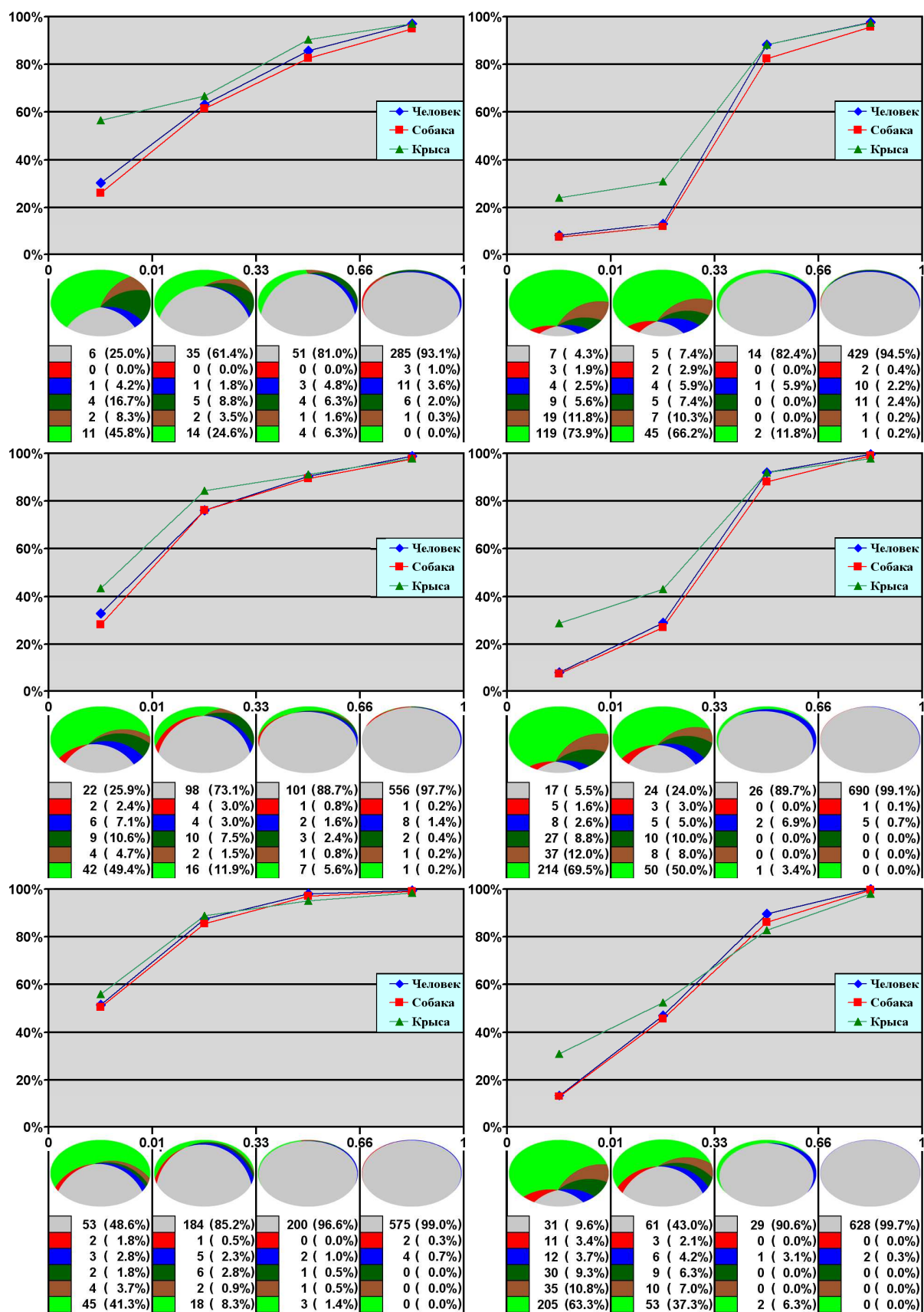


Рисунок 7. Консервативность сохраняющих (слева) и нарушающих (справа) рамку считывания кассетных экзонов мыши в генах человека, собаки и крысы для разных интервалов частоты включения кассетного экзона. (Продолжение на следующей странице).

Консервативность подсчитана отдельно для трех групп генов: сверху быстро эволюционирующие гены (2717 генов, уровень сходства белковых последовательностей менее 80%), посередине 3939 гена с уровнем сходства между 80% и 92%, внизу крайне консервативные гены (3447 генов, уровень сходства более 92%). На диаграммах серым цветом показана доля полностью консервативных альтернатив. Красным цветом показана доля альтернатив консервативных только в генах собаки, а синим цветом только в генах человека. Светло-зеленым цветом показаны неконсервативные кассетные экзоны. Экзоны мышцы, консервативные только в генах крысы, разделены на две подгруппы, случайно консервативные экзоны – коричневые, и экзоны, появившиеся в ходе эволюции грызунов (“молодые экзоны”) – темно-зеленые.

Консервативность сохраняющих и нарушающих рамку считывания кассетных экзонов мышцы показана на Рисунке 7. Консервативность была подсчитана отдельно для трех групп генов с разной скоростью молекулярной эволюции белковой последовательности. Консервативность кассетных экзонов мышцы сходна с консервативностью кассетных экзонов человека. Существует положительная корреляция между скоростью молекулярной эволюции генов (скоростью накопления и фиксации в геномной последовательности мутаций) и скоростью эволюции альтернативного сплайсинга, кассетные экзоны в быстро эволюционирующих генах существенно менее консервативны, чем кассетные экзоны в высоко консервативных генах. Доля экзонов, “случайно консервативных” в геноме крысы, составляет около половины от всех консервативных только в крысе экзонов, а оставшаяся половина кассетных экзонов является эволюционными новоприобретениями в ходе эволюции грызунов. Однако, в отличие от кассетных экзонов мышцы, консервативных только в человеке или собаке, трактуемых как эволюционные потери в собаке или человеке соответственно, или неконсервативных кассетных экзонов мышцы, мы можем лишь оценить лишь долю “молодых экзонов”, в то время как для остальных экзонов можно конкретно указать к какой группе они принадлежат. Доля “молодых экзонов” уменьшается с ростом частоты его включения в мРНК, что подтверждает теорию о том, что новые экзоны первоначально порождаются как кассетные экзоны с крайне малой частотой

включения в мРНК, и, постепенно, увеличивают ее в ходе эволюции. Доля “молодых экзонов” также существенно выше в быстро эволюционирующих генах по сравнению с высоко консервативными генами.

Глава 5. Альтернативный сплайсинг в дублицированных генах

Крайне интересным является вопрос об эволюции альтернативного сплайсинга в дублицированных генах. Обычно дублицированная копия гена является функционально избыточной и деградирует в результате накопления множества мутаций, приводящих к появлению терминирующих кодонов или к сдвигу рамки считывания и досрочному прекращению трансляции белковой последовательности. Для предотвращения уничтожения дублицированная копия должна частично или полностью изменить функцию кодируемого ей белка, а получившийся в результате новый ген должен давать преимущество, достаточное для его фиксации в популяции в ходе эволюции. Альтернативный сплайсинг является хорошим тестовым полигоном для тестирования и отбора новых экзонов и приносимых ими новых свойств белков, поэтому мы предположили, что в дублицированных генах должна наблюдаться повышенная доля новых кассетных экзонов, в частности “молодых экзонов”.

К исследуемой выборке дублицированных генов мы предъявляли следующие требования:

1. Дупликация произошла в ходе эволюции грызунов.
2. Дупликация произошла до расхождения предковых видов мыши и крысы.
3. В ходе эволюции были сохранены исходный вариант гена в геноме человека, собаки, мыши и крысы и дублицированный вариант гена в геноме мыши и крысы.

В исследуемых генах мы анализировали консервативность кассетных экзонов мыши в генах человека, собаки и крысы.

Для поиска генов, дублицированных в ходе эволюции грызунов мы выравнивали аминокислотные последовательности всех доступных в EDAS генов, 21791 генов мыши и 19718 генов человека при помощи программы Blat. Для каждого гена мыши был определен наиболее близкий, сходный по аминокислотной последовательности кодируемого белка, ген человека, гомолог. Гены мыши, имеющие гомологом один и тот же ген человека, отбирались для дальнейшего анализа. Для каждого гена мыши исследовалось наличие его ортолога в геноме крысы. Все дублицированные гены мыши, имеющие ортолога в крысе, были разделены на две группы: сохранившие исходную функцию гены (СФ гены) определялись как наиболее близкие к соответствующему гомологу человека дублицированные гены мыши, а в группу изменивших исходную функцию гены (ИФ гены) были отнесены все остальные дублицированные гены. Всего было получено 250 СФ генов и 387 ИФ генов.

Таблица 3. Суммарная информация о консервативности кассетных экзонов мыши.

	дублицированные гены		ортологичные гены эволюционирующие:			
	ИФ	СФ	быстро	умеренно	медленно	все
Консервативные экзоны	29 (62%)	43 (72%)	877 (76%)	1591 (78%)	1813 (81%)	4281 (79%)
“Молодые экзоны”	11 (23%)	7 (12%)	77 (7%)	114 (6%)	100 (4%)	291 (5%)
Оценка реального числа “молодых экзонов”	6 (13%)	4 (7%)	44 (4%)	61 (3%)	48 (2%)	153 (3%)
Неконсервативные экзоны	7 (15%)	10 (17%)	196 (17%)	331 (16%)	327 (15%)	854 (16%)

Анализ данных EDAS выявил наличие 47 кассетных экзонов в ИФ генах и 60 экзонов в СФ генах, что недостаточно для детального анализа консервативности наблюдаемых экзонов, поэтому мы суммировали информацию о консервативности кассетных экзонов для всех исследуемых групп генов в Таблице 3. Доля неконсервативных кассетных экзонов мыши совпадает для всех исследуемых групп генов. Доля “молодых экзонов” в дублицированных генах существенно больше, чем в не дублицированных, а среди дублицированных генов доля “молодых экзонов” в ИФ генах в два раза больше, чем в СФ генах.

Основные результаты и выводы

1. Создана база данных альтернативно сплайсируемых генов EDAS, в которой собрана информация об альтернативном сплайсинге генов человека и мыши.
2. Разработан алгоритм анализа выделения сложных и элементарных альтернатив с использованием алгоритмов анализа графов сплайсинга. Реализован алгоритм оценки достоверности наблюдаемых альтернатив.
3. Проведен анализ консервативности элементарных событий альтернативного сплайсинга генов человека в геномах мыши и собаки, а также анализ консервативности элементарных событий альтернативного сплайсинга генов мыши в геномах человека, собаки и крысы. Впервые показана существенно меньшая консервативность альтернативных экзонов и сайтов сплайсинга по сравнению с константными.
4. Показано, что нарушающие рамку считывания альтернативы существенно менее консервативны, чем сохраняющие рамку считывания. Показано, что существенная доля редких нарушающих рамку считывания альтернатив функциональны.
5. Показано, что распределение альтернативных сайтов сплайсинга согласуется с моделью случайной фиксации сайтов. Альтернативный вариант акцепторного и донорного сайтов преимущественно приближает длину экзона к стандартной, укорачивая длинные экзоны и удлиняя короткие.
6. Установлено, что максимальная консервативность кассетных экзонов наблюдается в генах с наименьшей скоростью молекулярной эволюции, в более быстро эволюционирующих генах возрастает доля неконсервативных альтернатив.
7. Показано, что гены с повышенной скоростью молекулярной эволюции существенно более часто приобретают новые кассетные экзоны в ходе эволюции. Также такое поведение характерно для дублицированных генов, особенно для изменивших в ходе эволюции функцию кодируемого белка дублицированных копий генов.

Список работ опубликованных по теме диссертации

1. Нуртдинов РН, Миронов АА и Гельфанд МС. Консервативен ли альтернативный сплайсинг млекопитающих? // Биофизика, 2002, Т. 47. № 4. С. 587-594.
2. Nurtadinov RN, Artamonova II, Mironov AA and Gelfand MS. Low conservation of alternative splicing patterns in the human and mouse genomes. // *Hum. Mol. Genet.*, 2003, V. 12 P. 1313-1320.
3. Offman MN, Nurtadinov RN, Gelfand MS and Frishman D. No statistical support for correlation between the positions of protein interaction sites and alternatively spliced regions. // *BMC Bioinformatics*, 2004, 5: 41.
4. Neverov AD, Artamonova II, Nurtadinov RN, Frishman D, Gelfand MS and Mironov AA: Alternative splicing and protein function. // *BMC Bioinformatics*. 2005, 6: 266.
5. Ermakova EO, Nurtadinov RN and Gelfand MS. Fast rate of evolution in alternatively spliced coding regions of mammalian genes. // *BMC Genomics*, 2006, 7: 84.
6. Нуртдинов РН, Неверов АД, Малько ДБ, Космодемьянский ИА, Ермакова ЕО, Раменский ВЕ, Миронов АА и Гельфанд МС. EDAS, база данных альтернативно сплайсируемых генов человека. // *Биофизика*, 2006, 51(4), 589-592.
7. Ermakova EO, Nurtadinov RN, Gelfand MS. Overlapping alternative donor splice sites in the human genome. // *J Bioinform Comput Biol*. 2007, 5(5): 991-1004.
8. Nurtadinov RN, Neverov AD, Favorov AV, Mironov AA, Gelfand MS. Conserved and species-specific alternative splicing in mammalian genomes. // *BMC Evol Biol*. 2007, 7: 249.
9. Ramensky VE, Nurtadinov RN, Neverov AD, Mironov AA, Gelfand MS. Positive selection in alternatively spliced exons of human genes. // *Am J Hum Genet*. 2008, 1: 94-98.

Эволюция альтернативного сплайсинга генов млекопитающих

Была исследована эволюция альтернативного сплайсинга. Альтернативный сплайсинг является одним из важнейших механизмов образования белкового многообразия млекопитающих. Наш предварительный анализ впервые показал, что доля видоспецифичных альтернатив по разным методам оценки составляет от одной четверти до одной трети от всех альтернатив.

Была создана база данных альтернативно сплайсируемых генов EDAS. Эта база данных содержит информацию о конститутивном и альтернативном сплайсинге 20809 генов человека и 31818 генов мыши. Анализ консервативности альтернативно сплайсированных участков генов человека показал, что консервативность кассетных экзонов зависит от частоты их использования и способности сохранять рамку считывания при трансляции. Было показано, что распределение альтернативных сайтов сплайсинга подчиняется модели фиксации случайных сайтов: альтернативные сайты сплайсинга преимущественно удлиняют короткие экзоны и укорачивают длинные.

Далее был проведен анализ специфичного для грызунов альтернативного сплайсинга. Для этого были сравнены геномы мыши и крысы, а геномы человека и собаки были использованы в качестве внешних объектов. Данная схема позволяет отделить возможные ошибочные варианты мРНК от настоящих кассетных экзонов, появившихся в ходе эволюции линии грызунов. В результате было показано, что гены с повышенной скоростью молекулярной эволюции и дублицированные копии генов существенно чаще (по сравнению с общей выборкой) приобретают новые кассетные экзоны в ходе эволюции.

Evolution of mammalian alternative splicing

We addressed the evolutionary aspect of alternative splicing. Alternative splicing is one of the main mechanisms for generating functional and evolutionary diversity of proteins in mammals. Our initial comparative-genomic analyses of alternative splicing for the first time demonstrated that the fraction of genome-specific alternative splicing may be as high as 1/4 to 1/3 of all observed alternatives.

We developed a database of alternatively spliced genes, EDAS. This database now contains information about splicing and alternative splicing of 20809 human and 31811 mouse genes. In a study of conservation of human alternatively spliced genes in the mouse and dog genomes, we demonstrated that conservation of cassette exons depends on their expression level and their frame-preservation ability. The distribution of alternative sites is consistent with a model of random fixation: alternative splice sites tend to extend short exons, truncate long exons, and extend very short introns.

We then extended this analysis to rodent-specific alternative splicing by comparing mouse and rat genomes with human and dog as outgroups. This study design allowed us to distinguish between splicing noise and bona fide lineage-specific alternative splicing and to estimate the rate of the latter. We also demonstrated that rodent-specific alternative exons are relatively more frequent in rapidly evolving genes and in paralogs.

Автор выражает глубокую благодарность научному руководителю А.А.Миронову, а также М.С.Гельфанду, соавторам многих совместных работ и сотрудникам Национальной Лаборатории Лоуренса в городе Беркли, США за предоставленные вычислительные ресурсы.

В разное время научная работа была поддержана грантами: РФФИ (00-15-99362), РФФИ (04-04-49440), INTAS (99-1476), INTAS (05-1000008-8028), ННМИ (55000309), ННМИ (55001056), LICR/CRDF (RB0-1268) и программой "Молекулярная и клеточная биология" РАН России.