На правах рукописи

Ермакова Екатерина Олеговна

# ОСОБЕННОСТИ ЭВОЛЮЦИИ РАЗЛИЧНЫХ ФУНКЦИОНАЛЬНЫХ ОБЛАСТЕЙ АЛЬТЕРНАТИВНО СПЛАЙСИРУЕМЫХ ГЕНОВ ЭУКАРИОТ

03.00.28 – биоинформатика

Автореферат

диссертации на соискание ученой степени

кандидата биологических наук

Москва – 2008

Работа выполнена на факультете биоинженерии и биоинформатики Московского государственного университета имени М.В. Ломоносова и в Учебно-научном центре „Биоинформатика“ Учреждения Российской академии наук Института проблем передачи информации им. А.А. Харкевича РАН.

Научный руководитель: кандидат физико-математических наук, доктор биологических наук, профессор Гельфанд Михаил Сергеевич

Официальные оппоненты: доктор биологических наук Алёшин Владимир Вениаминович Институт физико-химической биологии им. А.Н. Белозерского МГУ

доктор биологических наук, профессор Евгеньев Михаил Борисович Учреждение Российской академии наук Институт молекулярной биологии им.В.А. Энгельгардта РАН

Ведущая организация: Федеральное государственное унитарное предприятие Государственный научно-исследовательский институт генетики и селекции промышленных микроорганизмов

Защита диссертации состоится __ _____ 2008 года в ___ часов на заседании диссертационного совета Д.002.077.02 при Учреждении Российской академии наук Институте проблем передачи информации им. А.А. Харкевича РАН по адресу: 127994, г. Москва, ГСП-4, Большой Каретный переулок, д. 19, стр. 1.

С диссертацией можно ознакомиться в библиотеке Учреждения Российской академии наук Института 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

На данный момент секвенировано более 150 геномов эукариот и 1200 геномов прокариот, ведутся работы по секвенированию ещё около 200 геномов эукариот и 600 геномов прокариот. Темпы секвенирования значительно опережают темпы экспериментального анализа геномов, и изучение структуры и функции ДНК, РНК и белков на всех этапах включает использование специальных вычислительных средств. Наличие большого количества геномов сделало возможным изучение эволюции биологических последовательностей биоинформатическими методами. Задача восстановления профилей экспрессии и эволюционной истории генов вычислительными методами на основе данных о нуклеотидных последовательностях ДНК и мРНК и аминокислотных последовательностях белков особенно сложна и интересна для генов многоклеточных эукариот, так как они имеют наиболее сложную структуру и считываемая с них пре-мРНК часто альтернативно сплайсируется. Эволюция сайтов сплайсинга и альтернативно сплайсируемых участков генома и составляет предмет данной работы.

У многоклеточных эукариот альтернативный сплайсинг — один из основных механизмов создания разнообразия белковых последовательностей. Альтернативный сплайсинг в кодирующей области может внести слабые изменения в структуру и функцию белка, может резко изменить их, может привести к образованию нетранслируемой изоформы. Альтернативный сплайсинг является объектом сложной регуляции, но и сам может выступать в роли регуляторного механизма. Хотя в конце 1990х годов уже было описано достаточное количество отдельных важных случаев альтернативного сплайсинга, а также мутаций, нарушающих механизм альтернативного сплайсинга в отдельных генах и являющихся причиной врождённых заболеваний, альтернативный сплайсинг казался редким явлением: считалось, что альтернативно сплайсируются примерно 5% генов человека. Только недавние проекты по массовому секвенированию EST-маркеров, результатом которых стало накопление большого объёма нуклеотидных последовательностей фрагментов мРНК человека, породили достаточно данных для реальных оценок распространённости альтернативного сплайсинга. Выравнивание нуклеотидных последовательностей EST-маркеров с последовательностями хромосомной ДНК и полноразмерных мРНК показало, что альтернативно сплайсируется по меньшей мере треть генов человека (Mironov et al 1999). Последующее накопление данных и усовершенствование биоинформатических алгоритмов только увеличило эту оценку.

Есть все основания считать, что альтернативно сплайсируемые участки генов служат „экспериментальной площадкой“ молекулярной эволюции. Многие исследования подтверждают эту точку зрения. Так, альтернативные изоформы часто эволюционно молоды как в генах млекопитающих, так и в генах насекомых. Плотность несинонимичных нуклеотидных замен ($d_N$) в альтернативных областях генов выше, чем в постоянных областях. Постоянные экзоны в генах с геномспецифичным альтернативным сплайсингом эволюционируют быстрее, чем постоянные участки генов с консервативной структурой. Многие молодые (специфичные для грызунов, и отсутствующие в ортологичных генах человека и свиньи) экзоны альтернативно сплайсируются и при сравнении нуклеотидной последовательности мыши и крысы обнаруживают $d_N/d_S>1$. Частота несинонимичных однонуклеотидных полиморфизмов в генах человека выше в альтернативных областях, чем в постоянных.

Существенную роль в эволюции кодирующих, в том числе, альтернативно сплайсируемых, последовательностей играют точечные нуклеотидные замены, т. е. зафиксировавшиеся в популяции точечные мутации. Литературные данные о фиксации нуклеотидных замен в альтернативно сплайсируемых генах были противоречивы и нуждались в повторном анализе. В данной работе изучено распределение точечных нуклеотидных замен в альтернативных кодирующих областях генов млекопитающих, на материале полных

геномов человека и мыши, и насекомых, на примере полных геномов двух видов плодовой мушки. Отдельно исследовано поведение нуклеотидных замен в концевых и внутренних участках гена. Рассмотрены как синонимичные замены, так и замены, изменяющие последовательность кодирумого белка. Установлено их взаимное распределение на геномном уровне, что позволило промоделировать действие отрицательного и положительного отбора на кодирующие области альтернативно сплайсируемых генов.

Недавно были исследованы перекрывающиеся сайты сплайсинга со сдвигом сайта на три нуклеотида: акцепторных, с консенсусом NAGNAG, и донорных, с консенсусом GYNGYN. При выборе альтернативы в таком сайте не происходит сдвига рамки считывания, однако мотив GYNGYN далёк от консенсусной последовательности донорного сайта, и левый (5′) сайт оказывается нарушенным. Поэтому возникла необходимость рассмотрения перекрывающихся донорных сайтов и других типов.

## Цель и задачи исследования

Целью данной работы было изучение экспрессии и эволюции альтернативно сплайсируемых генов эукариот методами сравнительной геномики. Были поставлены и решены следующие задачи:

- поиск потенциальных донорных сайтов сплайсинга, перекрывающихся с активными донорными сайтами сплайсинга;
- оценка возможности порождения транслируемой изоформы потенциальными сайтами сплайсинга, а также сайтами сплайсинга, подтверждёнными только фрагментами мРНК (EST-маркерами);
- изучение консервативности потенциальных и активных перекрывающихся донорных сайтов человека в геномах мыши и собаки;
- установление корреляции между взаимным расположением перекрывающихся донорных сайтов, их весами, предпочтениями при экспрессии, транслируемостью порождаемых ими изоформ и их сохранением в процессе эволюции;
- реализация метода Ины оценки числа синонимичных и несинонимичных нуклеотидных замен;
- сравнение скорости фиксации точечных мутаций в постоянных и альтернативных кодирующих участках генов млекопитающих и насекомых;
- сравнение скорости фиксации точечных мутаций в различных классах альтернативных кодирующих участков;
- реконструкция действия естественного отбора на кодирующие области альтернативно сплайсируемых генов.

## Новизна работы

В работе впервые на геномном уровне изучены перекрывающиеся альтернативные донорные сайты сплайсинга, переключающие рамку считывания. Впервые получены данные о молекулярной эволюции альтернативно сплайсируемых участков генов насекомых, а также выявлены особенности эволюции концевых альтернативных участков генов млекопитающих и насекомых. Полученные данные о фиксации синонимичных и несинонимичных нуклеотидных мутаций в кодирующих областях генов млекопитающих и насекомых позволяют уточнить действие сил отбора на альтернативных участках генов.

## Практическая ценность

Реализованы алгоритмы построения матрицы позиционных весов и последующего вычисления веса сайта. Построенная весовая матрица для донорного сайта сплайсинга человека может применяться для оценки активности потенциальных донорных сайтов сплайсинга и интенсивности экспрессии альтернативных изоформ.

Разработана программная реализация метода Ины оценки числа синонимичных и несинонимичных нуклеотидных замен. способная производить оценку эволюционных параметров для очень длинных выравниваний (~$10^6$ п. н.).

Полученные данные о функционировании перекрывающихся донорных сайтов могут быть использованы в биоинженерии.

### Апробация работы

Материалы исследований по теме диссертации были представлены на международных конференциях: XII Международной конференции студентов, аспирантов и молодых учёных „Ломоносов“ (Москва, апрель 2005), 2nd Int. Moscow Conference on Computational Molecular Biology MCCMB'05 (Москва, июль 2005), школе „Биоинформатика, геномика, протеомика“ (Алма-Ата, Казахстан, апрель 2006), Human Genome Meeting HGM2006 (Хельсинки, Финляндия, июнь 2006), 4th Special Interest Group Meeting on Alternative Splicing AS-SIG 2007 (Вена, Австрия, июль 2007), 15th Annu. Int. Conf. on Intelligent Systems for Molecular Biology and 6th European Conf. on Computational Biology ISMB/ECCB'07 (Вена, Австрия, июль 2007), 3rd Int. Moscow Conference on Computational Molecular Biology MCCMB'07 (Москва, июль 2007), а также на 30-й конференции молодых ученых и специалистов ИППИ РАН ИТиС'07 (Звенигород, сентябрь 2007) и на научных семинарах на факультете биоинженерии и биоинформатики МГУ и в ИППИ РАН.

### Объём и структура диссертации

Диссертационная работа изложена на ___ страницах и состоит из введения, четырёх глав, выводов и списка цитированной литературы. Глава 1 содержит обзор литературы по теме диссертации. Глава 2 содержит описание использованных данных, а также программного обеспечения (в том числе авторского) и алгоритмов, применявшихся для решения задач, поставленных в диссертации. Главы 3 и 4 содержат описание новых результатов и их обсуждение в контексте литературных данных. Список литературы включает ___ наименований. Работа содержит ___ рисунков и ___ таблиц.

## Содержание работы

### Глава 1. Обзор литературы

Содержит мотивировку поставленных задач, а также аналитический обзор современной литературы по проблемам, рассмотренным в диссертации.

### Глава 2. Материалы и методы

Данные об сплайсинге в генах человека (разметка альтернативных и постоянных сайтов сплайсинга на геномной последовательности) были взяты из базы EDAS (EST-Derived Alternative Splicing database, Neverov et al 2005, http://www.genebee.msu.su/edas). Данные о сплайсинге в генах плодовой мушки *Drosophila melanogaster* были взяты из базы данных FlyBase (Misra et al 2002, Grumbling et al 2006, flybase.bio.indiana.edu), 3 версия аннотации.

Поиск ортологов и выравнивание геномных последовательностей ортологичных генов для человека и мыши было проведено как в (Jordan et al 2001), для *Drosophila melanogaster* и *Drosophila pseudoobscura* — как в (Malko et al 2006). Тройки ортологичных генов человека, мыши и собаки были взяты из (Linblad-Toh et al 2005).

Для оценки количества несинонимичных замен на несинонимичную позицию $d_N$ и синонимичных замен на синонимичную позицию $d_S$ был использован метод Ины (Ina 1995), реализованный в виде специально написанной программы на языке Perl. Время обработки программой одного выравнивания длиной 5904081 п. н. составляет 28 секунд (AMD Sempron 3100+, ОЗУ 512 МБ). Необходимость в разработке собственной программы была вызвана тем,

что известные автору программные реализации данного метода были рассчитаны на исследование отдельных генов и не могли обрабатывать выравнивания длины, сравнимой по порядку с полным эукариотическим геномом. Кроме того, была существенна возможность запуска программы из командной строки: это позволило автоматизировать запуск программы (всего в процессе исследования было обработано несколько десятков тысяч выравниваний).

Для определения точности оценки эволюционных параметров на конкатенированных выравниваниях был применён метод бутстреппинга. Для каждого выравнивания было построено 2000 выравниваний, составленных случайным образом из столбцов исходного выравнивания с возвращением.

Для построения матрицы позиционных весов была использована выборка из 85798 постоянных донорных сайтов, подтвержденных полноразмерной мРНК илиEST-маркерами из как минимум двух независимых клонотек. Использовались позиции сайта с –3 по +6. Позиционные веса нуклеотидов вычислялись как в (Gelfand et al 2000):

$$W(b;m)=\log[N(b;m)+0{,}5]-0{,}25\cdot\sum_{i=A,C,G,T}\log[N(i;m)+0{,}5]$$

где $N(b;m)$ — количество сайтов выборки, содержащих нуклеотид $b$ в позиции $m$. Матрица $W(b;m)$ приведена в таблице 1. Вес сайта ($b_{-3},...,b_9$), состоящего из нуклеотидов $b_j$, вычислялся как сумма позиционных весов:

$$w(b_{-3},...,b_9)=W(b_{-3},-3)+...+W(b_9,9)$$

Таблица 1. Весовая матрица для донорного сайта сплайсинга

| | –3 | –2 | –1 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|---|---|
| A | 0,3945 | 1,2554 | –0,1238 | –1,0455 | –2,5929 | 1,6810 | 1,4464 | –0,3671 | –0,2059 |
| C | 0,4488 | –0,5347 | –1,3751 | –1,7430 | 0,4388 | –1,3981 | –0,7729 | –0,8412 | –0,3878 |
| G | –0,2227 | –0,4793 | 1,9448 | 5,5628 | –2,9786 | 1,0544 | –0,3427 | 1,7347 | –0,1666 |
| T | –0,6207 | –0,2414 | –0,4459 | –2,7743 | 5,1327 | –1,3372 | –0,3307 | –0,5264 | 0,7603 |

Отождествление ортологичных сайтов сплайсинга в генах человека, мыши и собаки проводилось при помощи программ BLAT (Kent 2002) и Pro-Gen (Novichkov et al 2001). Программа IsoformCounter (Neverov et al 2005) использовалась для предсказания транслируемых и нетранслируемых изоформ зрелых мРНК альтернативно сплайсируемых генов. Лого-диаграммы были построены при помощи программы WebLogo [weblogo.berkeley.edu]. Статистические проверки значимости были проведены при помощи статистического пакета R [http://www.r-project.org].

Для обработки и статистического анализа данных применялись программы, написанные автором на языках Perl и Java. В частности, были реализованы алгоритмы построения матрицы позиционных весов и последующего вычисления веса сайта, а также метод Ины оценки числа синонимичных и несинонимичных нуклеотидных замен.

## Глава 3. Перекрывающиеся донорные сайты сплайсинга в геноме человека

Более половины всех донорных сайтов сплайсинга в геноме человека имеют потенциальный перекрывающийся альтернативный донорный сайт на расстоянии 3-6 нуклеотидов. Сохранность этих потенциальных сайтов в ортологичных генах определяется требованиями консенсуса и положением сайта в экзоне либо в интроне относительно активного сайта. Несколько сотен пар перекрывающихся сайтов альтернативно сплайсируются, активность каждого из них может быть подтверждена белком, полноразмерной мРНК или EST-маркерами из двух независимых клонотек. Стремление обоих перекрывающихся сайтов к консенсусу может предъявлять противоречивые требования к области их перекрывания. Альтернативно

сплайсируемые пары донорных сайтов со сдвигом на четыре нуклеотида наиболее распространены, несмотря на порождаемый ими сдвиг рамки считывания в области гена, кодирующей белок. Частота использования двух перекрывающихся сайтов, как правило, сильно отличается, причём основной сайт чаще сохраняется в ортологичных генах других млекопитающих. Функции перекрывающихся донорных и акцепторных сайтов сплайсинга, видимо, отличаются: выбор альтернативного акцепторного сайта обычно приводит лишь к небольшим вариациям последовательности белка, тогда как выбор альтернативного донорного сайта часто порождает сдвиг рамки считывания, что приводит к значительным изменениям белковой последовательности либо к синтезу нетранслируемых изоформ, способных играть регуляторную роль, например, становясь мишенью для механизмов деградации мРНК.

**Определения.** Консенсус донорного сайта сплайсинга человека — MAG|GTRAGT (Gelfand 1989). Будем считать, что донорный сайт сплайсинга состоит из 9 нуклеотидов, пронумерованных, как показано на рисунке 1. Пара нуклеотидов (+1, +2) образует *ядро донорного сайта сплайсинга*. Именно ядро — наиболее строго регламентированный и консервативный участок сайта сплайсинга.

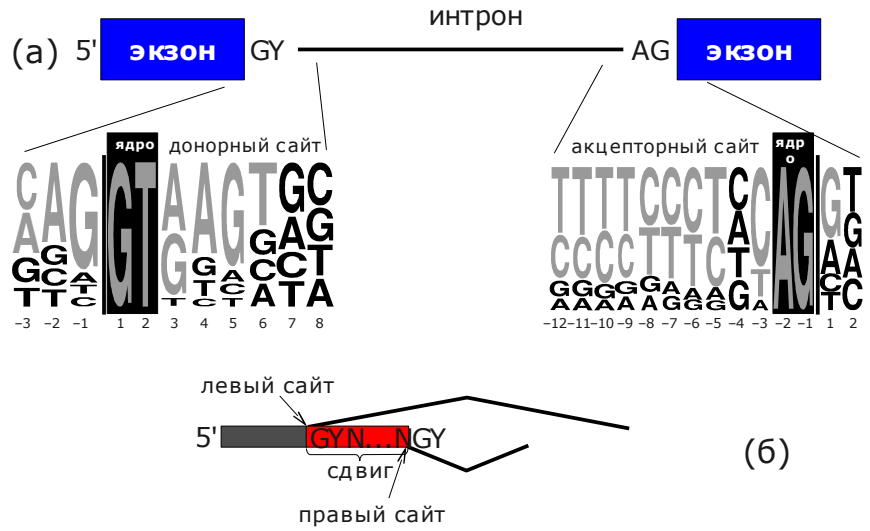



Рисунок 1. (а) донорные и акцепторные сайты сплайсинга
(б) альтернативные донорные сайты сплайсинга

Консенсус акцепторного сайта сплайсинга человека $(Y_n)$NYAG|G. Нумерация нуклеотидов показана на рисунке 1. Пара нуклеотидов (−2, −1) образует *ядро акцепторного сайта сплайсинга*.

Далее *потенциальным донорным сайтом сплайсинга* будет считаться мотив из 9 нуклеотидов, пронумерованных (−3, −2, −1, +1, +2, +3, +4, +5, +6), имеющий динуклеотид GT в позиции (+1, +2).

Расстояние (в нуклеотидах) между потенциальными экзон-интронными границами пары потенциальных донорных или акцепторных сайтов сплайсинга будем называть *сдвигом сайта*.

Поскольку традиционно последовательность ДНК записывают слева направо от 5′ к 3′ концу, из пары потенциальных сайтов сплайсинга ближайший к 5′ концу гена далее будем называть левым, а ближайший к 3′ концу — правым. Как левый, так и правый сайт в паре может быть активным либо спящим (спящим будем называть сайт сплайсинга, активность которого не подтверждена текущими данными о белках, мРНК и EST-маркерах).

Будем говорить, что потенциальный сайт сплайсинга находится в положении *лX* относительно активного сайта сплайсинга, если первый находится ближе второго к 5′ концу

гена на X нуклеотидов, и в положении *nX*, если первый находится ближе второго к 3′ концу гена на X нуклеотидов. В данной работе рассматриваются потенциальные сайты, находящиеся в положениях л6, л5, л4, л3 и п3, п4, п5, п6 относительно активного сайта, т. е. со сдвигом X от 3 до 6 нуклеотидов.

Сайт в альтернативно сплайсируемой паре будем называть *основным*, если он используется не менее чем в двух третях случаев (по данным EST-маркеров) и *минорным*, если он используется менее чем одной трети случаев (в некоторых парах оба сайта используются со сравнимой частотой).

Далее запись вида „$GTN_kGT$“ означает, что имеются в виду как альтернативно сплайсируемые пары вида $|GTN_k|GT$, так и потенциальные пары, в которых активен только левый ($|GTN_kGT$) или только правый ($GTN_k|GT$) сайт.

Активными считались только донорные сайты, подтверждённые белком, полноразмерной мРНК или EST-маркерами из двух или более независимых лабораторий (по данным базы EDAS). Мы рассматривали только канонические сайты с ядром GT.

**Результаты.** Мы рассмотрели 187725 донорных сайтов сплайсинга человека. 96968 (52%) из них имели GT в позиции л6, л5, л4, л3, п3, п4, п5 или п6 (таблица 2). Потенциальные сайты типа п3 оказались самыми редкими (0,6%), тогда как потенциальные сайты типа п4 — самыми частыми (39,4%), т. к. GT — консенсус позиций (+5, +6) донорного сайта сплайсинга человека.

Таблица 2. Статистика потенциальных донорных сайтов на расстоянии 3-6 нуклеотидов от активного донорного сайта сплайсинга

| позиция потенциального сайта | Л6 | Л5 | Л4 | Л3 | П3 | П4 | П5 | П6 |
|---|---|---|---|---|---|---|---|---|
| количество | 8841 | 5555 | 3379 | 3895 | 1182 | 74019 | 7181 | 12034 |
| частота | 4,7% | 3,0% | 1,8% | 2,1% | 0,6% | 39,4% | 3,8% | 6,4% |

385 пар донорных сайтов со сдвигом от 3 до 6 нуклеотидов, для которых был подтверждён альтернативный сплайсинг, удалось картировать на тройки ортологичных генов человека, мыши и собаки. Альтернативно сплайсируемые пары со сдвигом на 4 нуклеотида встречались наиболее часто.

Веса левого ($w_л$) и правого ($w_п$) сайтов в альтернативно сплайсируемых парах вычислялись, как описано в главе 2. Совместное распределение $w_л$ и $w_п$ для альтернативно сплайсируемых пар с основным левым сайтом, с основным правым сайтом и без выраженного основного сайта показано на рисунке 2. В парах вида $|GTN|GT$ два сайта не могли оказаться сильными одновременно, т. к. перекрытие сайтов создаёт конфликт консенсусов.

В парах вида $|GTN_2|GT$ левые сайты сильнее и чаще предпочитаются правым. Для пар вида $|GTN_3|GT$ и $|GTN_4|GT$ распределения весов $w_л$ и $w_п$ отличаются мало. Сила сайта как правило (но не всегда!) определяет, будет ли сайт основным или минорным при любых расстояниях между альтернативными сайтами.

Поскольку по определению для подтверждения альтернативного варианта было достаточно только EST-маркеров (т. е. в базе данных могло не присутствовать полноразмерной изоформы, содержащей данный вариант), мы использовали алгоритм IsoformCounter (Neverov et al 2005) для предсказания транслируемых изоформ (таблица 3).

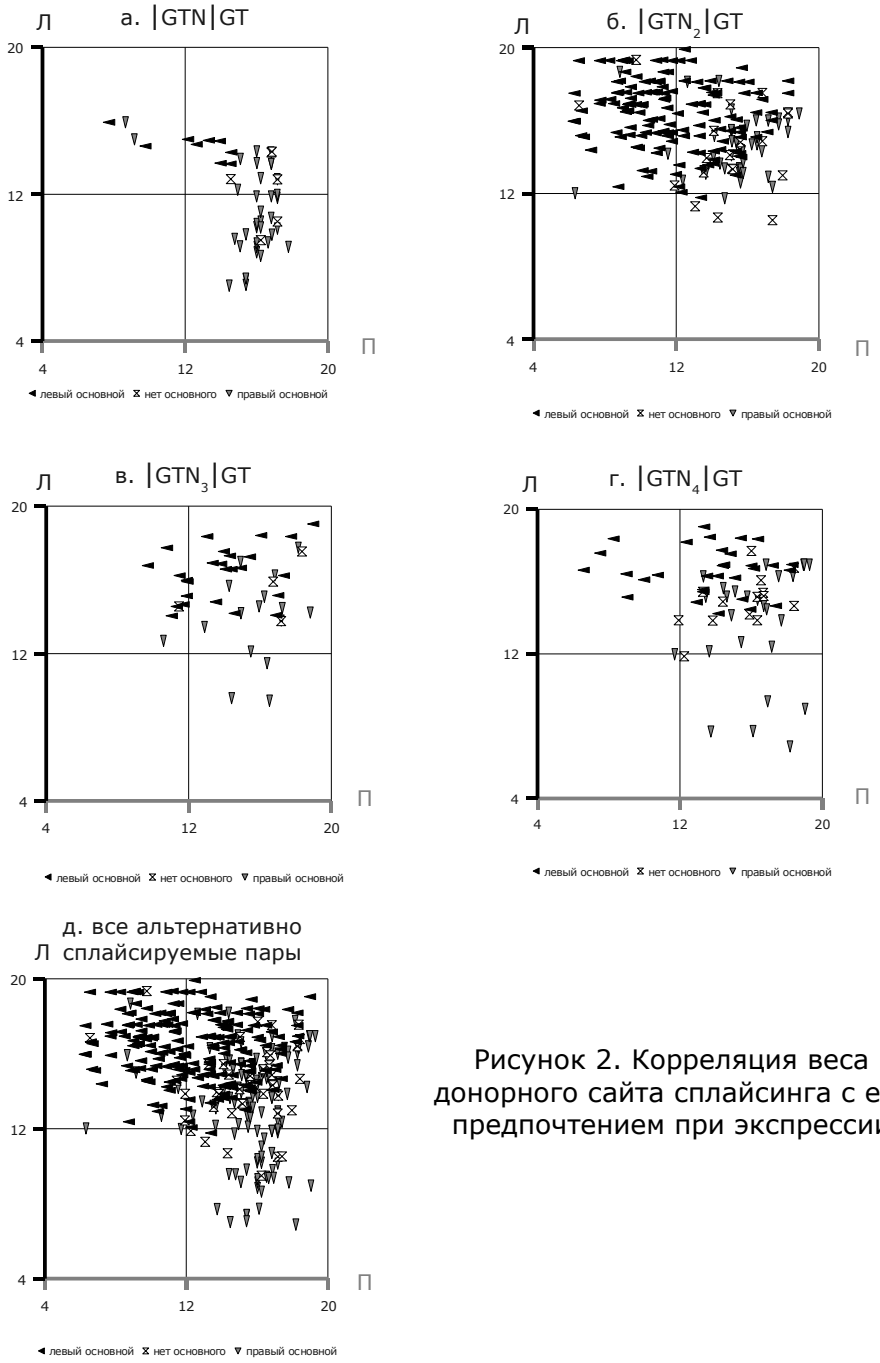



Рисунок 2. Корреляция веса донорного сайта сплайсинга с его предпочтением при экспрессии

Таблица 3. Использование альтернативных донорных сайтов сплайсинга в белок-кодирующих изоформах (транслируемость)

| левый транслируемый | правый транслируемый | сдвиг сайта (нукл.) | | | | |
|---|---|---|---|---|---|---|
| | | 3 | 4 | 5 | 6 | всего |
| + | + | 14 | 31 | 20 | 52 | 117 |
| + | − | 7 | 121 | 15 | 10 | 153 |
| − | + | 28 | 23 | 5 | 3 | 59 |
| − | − | 3 | 39 | 6 | 8 | 56 |
| | всего | 52 | 214 | 46 | 73 | 385 |

Пары с двумя нетранслируемыми сайтами располагались в нетранслируемых областях. В парах вида |GTN$_3$|GT с одним транслируемым сайтом, как правило, правый сайт был транслируемым, тогда как в парах вида |GTN$_4$|GT — левый сайт. Как и ожидалось (из-за большого расстояния между сайтами и сдвига на число нуклеотидов, кратное трём), наибольший процент пар, в которых обе изоформы порождают белок, оказался среди пар вида |GTN$_4$|GT (71%).

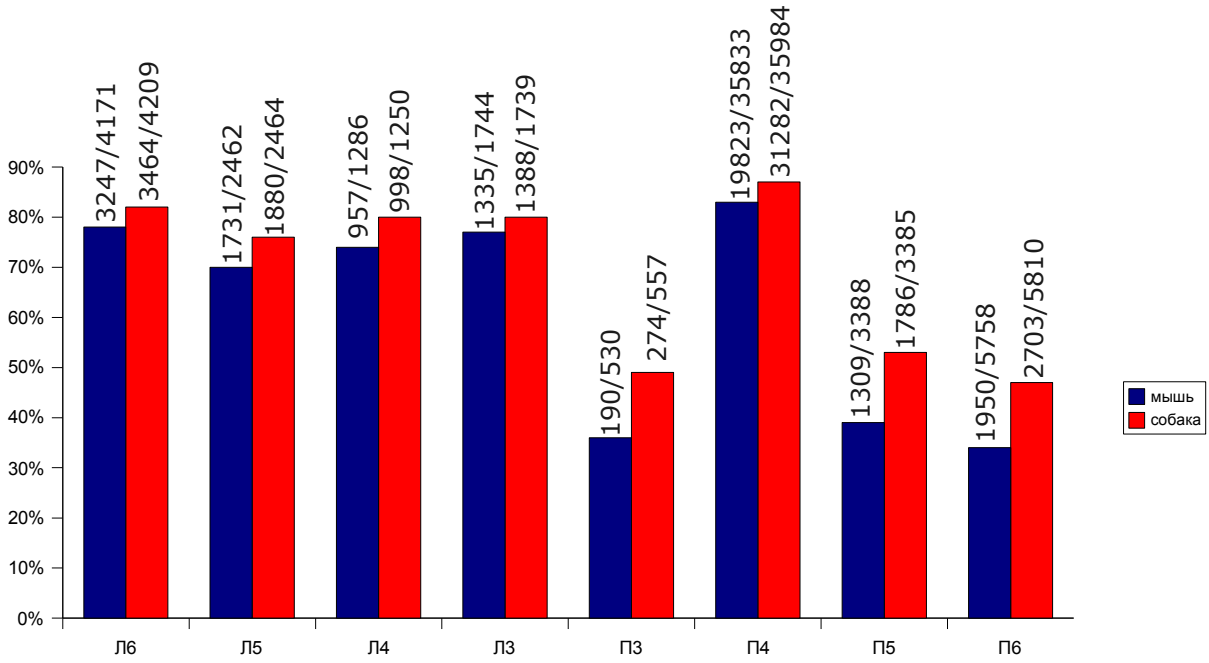



Рисунок 3. Сохранение потенциальных донорных сайтов вблизи сохранённого активного донорного сайта в геноме мыши и в геноме собаки. Над каждым столбцом гистограммы — отношение количества потенциальных донорных сайтов, находящихся рядом с сохранённым в геноме мыши (собаки) активным донорным сайтом к общему числу таких активных сайтов, выделены группы по положению потенциального сайта относительно активного

Транслируемая изоформа, как правило, оказывается основной по данным EST-маркеров. Однако, нетранслируемые изоформы могут быть мишенями нонсенс-мотивированной деградации или других управляемых клеточных механизмов деградации, поэтому они могут быть недопредставлены в базах EST-маркеров. Когда обе изоформы могут породить белок, сдвига интенсивности экспрессии в пользу левого или правого сайта не наблюдается.

Сайт сплайсинга человека считался консервативным в ортологичном гене мыши или собаки, если его позиция в соответствующем геноме могла быть определена при помощи программ BLAT (Kent 2002) и Pro-Gen (Novichkov et al 2001) и определённый таким образом ортологичный сайт содержал GT в позициях (+1, +2).

Из 126326 донорных сайтов сплайсинга человека, картированных на тройки ортологичных генов человека, мыши и собаки, 88696 (70%) были консервативны в геноме мыши и 89280 (71%) — в геноме собаки. Количество и консервативность потенциальных сайтов рядом с консервативными донорными сайтами сплайсинга показаны на рисунке 3. Как и ожидалось, интронные потенциальные сайты менее консервативны, чем экзонные, исключение составляют потенциальные сайты в позиции п4, согласующиеся с консенсусом активного сайта. Наименее консервативны GT в позиции п3, т. к. они вступают в конфликт с консенсусом.

Консервативность левых и правых сайтов в зависимости от частоты использования и величины сдвига отражена в таблице 4. Как и ожидалось, основные сайты чаще оказываются

консервативными, чем минорные, и пары, сохраняющие рамку считывания, более консервативны, чем пары, сдвигающие её.

Таблица 4. Сохранение перекрывающихся альтернативно сплайсируемых донорных сайтов сплайсинга

(а) в геноме мыши

| сдвиг сайта | 3 | | 4 | | 5 | | 6 | | всего | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Л | П | Л | П | Л | П | Л | П | Л | П |
| левый основной | 8/9 (90%) | 5/9 (60%) | 120/148 (80%) | 97/148 (70%) | 22/26 (80%) | 12/26 (50%) | 25/31 (80%) | 13/31 (40%) | 175/214 (80%) | 127/214 (60%) |
| нет основного | 4/6 (70%) | 3/6 (50%) | 10/21 (50%) | 8/21 (40%) | 1/4 (30%) | 2/4 (50%) | 10/15 (70%) | 11/15 (70%) | 25/46 (50%) | 24/46 (50%) |
| правый основной | 22/37 (60%) | 28/37 (80%) | 24/45 (50%) | 29/45 (60%) | 11/16 (70%) | 12/16 (80%) | 16/27 (60%) | 24/27 (90%) | 73/125 (60%) | 93/125 (70%) |
| всего | 34/52 (70%) | 36/52 (70%) | 152/214 (70%) | 128/214 (60%) | 35/46 (80%) | 22/46 (50%) | 51/73 (70%) | 45/73 (60%) | 272/385 (70%) | 231/385 (60%) |

(б) в геноме собаки

| сдвиг сайта | 3 | | 4 | | 5 | | 6 | | всего | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Л | П | Л | П | Л | П | Л | П | Л | П |
| левый основной | 8/9 (90%) | 5/9 (60%) | 118/148 (80%) | 91/148 (60%) | 23/26 (90%) | 8/26 (30%) | 25/31 (80%) | 10/31 (30%) | 174/214 (80%) | 114/214 (50%) |
| нет основного | 6/6 (100%) | 5/6 (80%) | 8/21 (40%) | 6/21 (30%) | 0/4 (0%) | 1/4 (30%) | 11/15 (70%) | 12/15 (80%) | 25/46 (50%) | 24/46 (50%) |
| правый основной | 23/37 (60%) | 29/37 (80%) | 23/45 (50%) | 28/45 (60%) | 11/16 (70%) | 12/16 (80%) | 21/27 (80%) | 24/27 (90%) | 78/125 (60%) | 93/125 (70%) |
| всего | 37/52 (70%) | 39/52 (80%) | 151/214 (70%) | 131/214 (60%) | 33/46 (70%) | 252/46 (50%) | 57/73 (80%) | 49/73 (70%) | 278/385 (70%) | 244/385 (60%) |

**Обсуждение.** Структура консенсуса может определять функциональные особенности перекрывающихся донорных сайтов. Консенсус донорного сайта сплайсинга содержит готовое ядро для донорного сайта на 4 нуклеотида правее, таким образом, левый сайт сильнее и в альтернативно сплайсируемых парах он, как правило, является основным. Ранее было показано, что альтернативный донорный сайт, как правило, сдвигает рамку считывания (эта статистика определяется, в основном, сдвигами на 4 нуклеотида, обусловленными консенсусом) (Tadokoro et al 2005, Akerman and Mandel-Gutfreund 2006), причём этот сдвиг не компенсируется сдвигом акцепторного сайта на втором конце интрона. Напротив, альтернативный акцепторный сайт, как правило, сохраняет рамку считывания (Akerman and Mandel-Gutfreund 2006).

Нами показано, что в 40% пар перекрывающихся донорных сайтов только левый (5′) сайт, а в 15% пар — только правый (3′) сайт порождает транслируемую изоформу, таким образом, вторая изоформа может индуцировать регулируемую деградацию. Хиллер и соавторы (Hiller et al 2006) экспериментально подтвердили использование обоих перекрывающихся донорных сайтов вида |GYN|GYN для семи генов человека, но не обнаружили тканеспецифичных различий в экспрессии парных изоформ. Более того, были исследованы различные аллели гена STAT3, и разницы в экспрессии изоформ для разных генотипов не было обнаружено. Если каждый из перекрывающихся донорных сайтов порождает транслируемую изоформу, полученные белки могут различаться адгезивными свойствами (Vogan et al 1996) или внутриклеточной локализацией (Tadokoro et al 2005), но на данный момент мне не известно ни одного сообщения о тканеспецифичной экпрессии изоформ, порождённых перекрывающимися альтернативными донорными сайтами, тогда как изоформы, порождаемые перекрывающимися акцепторными сайтами часто экспрессируются тканеспецифично (Tadokoro et al 2005). Эти наблюдения показывают, что функциональная роль тандемных донорных сайтов состоит скорее в равномерном поддержании концентраций белков, чем в их тканеспецифичной регуляции.

Остаётся неразрешённым вопрос о функциональной роли изоформ, которые либо порождают сильно изменённый укороченный белок, либо становятся мишенью для нонсенс-мотивированной деградации (НМД). Данные подтверждают, что эти сайты используются достаточно часто, что является серьёзным аргументом в пользу их функциональной важности. Было бы очень заманчиво приписать им регуляторную роль — например, намеренного уничтожения транскриптов в определённых условиях. Впрочем, подобную функцию принято приписывать непродуктивному альтернативному сплайсингу в целом (ср. Lareau et al 2004, 2007), а он широко распространён: показано (Lewis et al 2003), что 45% альтернативно сплайсируемых генов человека могут порождать изоформы с преждевременным стоп-кодоном — потенциальные мишени НМД.

## Глава 4. Нуклеотидные замены в альтернативных и постоянных белок-кодирующих участках генов

В последние несколько лет секвенирование полных геномов эукариот, а также масштабные проекты секвенирования мРНК сделали возможным проведение полногеномных исследований альтернативного сплайсинга. В данной работе проведено полногеномное исследование нуклеотидных замен в альтернативно сплайсируемых генах млекопитающих (на примере человека и мыши) и насекомых (на примере двух видов дрозофилы).

Рассмотрено 3029 генов человека и 790 генов плодовой мушки *Drosophila melanogaster*, альтернативно сплайсируемых в кодирующей области, и их ортологи в геномах мыши (*Mus musculus*) и другого вида плодовой мушки, *Drosophila pseudoobscura*, соответственно.

Анализировался только консервативный альтернативный сплайсинг, причём рассматривались только сайты, подтверждённые выравниванием полноразмерных транслированных мРНК с геномной последовательностью.

**Определения.** В альтернативно сплайсируемом гене назовём постоянными участки ДНК, образ которых всегда присутствует в кодирующей части сплайсируемой мРНК, а альтернативными — участки, образ которых может как присуствовать в кодирующей части, так и вырезаться как часть интрона (или целый интрон) при сплайсинге. Таким образом, нуклеотидная последовательность гена разбивается на интронные, экзонные некодирующие, постоянные и альтернативные участки. Экзон может быть постоянным, альтернативным, некодирующим или состоять из нескольких постоянных, альтернативных, а также некодирующих участков.

На нуклеотидных выравниваниях ортологичных генов были размечены постоянные (П) и альтернативные (А) участки. Среди альтернативных участков были выделены N-концевые

($A^N$), внутренние ($A^I$, от internal — „внутренний“) и С-концевые ($A^C$): альтернативный участок будем называть N-концевым, если в разметке гена нет постоянных участков, находящихся ближе к 5' концу, и С-концевым, если в разметке гена нет постоянных участков, находящихся ближе к 3' концу, иначе будем называть его внутренним альтернативным участком.

После разметки выравнивания каждого гена составлялись метавыравнивания (конкатенированные выравнивания) участков пяти классов (П, А, $A^N$, $A^I$, $A^C$). Рассматривались метавыравнивания двух типов: локальные, объединяющие кодирующие участки одного класса в пределах одного гена (например, все внутренние альтернативные участки гена BRCA1), и глобальные, объединяющие участки одного класса всех генов какой-либо выборки (например, внутренние альтернативные участки генов человека и мыши). Неполные кодоны, а также кодоны с делециями в выравнивании в метавыравнивания не включались. Для оценки $d_N$ и $d_S$ для парного выравнивания необходимо наличие достаточно длинного выравнивания (Ina 1995), мы использовали порог 80 п.н. Использование метавыравниваний позволило учесть не только длинные альтернативные фрагменты генов, такие, как кассетные экзоны, но и совсем короткие, такие, как удлинения — участки между двумя альтернативными донорными или двумя акцепторными сайтами. Для генов с длинными альтернативами сравнивалось эволюционное поведение альтернативных участков разных классов и постоянных участков, используя локальные метавыравнивания.

Оценивались три эволюционных параметра. Плотность несинонимичных замен ($d_N$) служит мерой расхождения аминокислотных последовательностей, соответствующих гомологичным участкам двух генов, и характеризует „насыщение“ кодирующего участка несинонимичными заменами. Плотность синонимичных замен ($d_S$) позволяет судить как об интенсивности мутаций в том или ином кодирующем участке (в сравнении с $d_N$), так и об эволюции „небелковых“ элементов гена, например, регуляторных последовательностей, таких, как экзонные энхансеры сплайсинга. Нормировка $d_N$ и $d_S$ согласована, и, в то время как $d_N$ и $d_S$ оценивают количество нуклеотидных замен с момента расхождения двух видов и зависят от времени, их отношение $\omega=d_N/d_S$ есть уже не функция времени, но характеристика давления отбора на этот участок. Если давление отбора отсутствует (т. е. любые замены нейтральны для организма), $\omega\approx1$. При $\omega>1$ заключают, что рассматриваемый участок белка и соответствующий ему кодирующий участок гена находятся под действием положительного отбора.

**Результаты.** Мутации пиримидиновых оснований в пиримидиновые или пуриновых в пуриновые в геномной ДНК происходят гораздо чаще, чем мутации пиримидиновых оснований в пуриновые и наоборот. Отношение скоростей транзиций и трансверсий R, необходимое для вычисления синонимичного и несинонимичного потенциала нуклеотидной позиции, оценивалось методом Ины (Ina 1995, глава 2) по метавыравниваниям всех кодирующих участков. Получилось, что для человека и мыши R=5,28, для двух дрозофил R=2,24.

В ортологичных белках у *H. sapiens* и *M. musculus* идентичны 81% аминокислот, у *D. melanogaster* и *D. pseudoobscura* — 64% аминокислот.

Эволюционные параметры оценивались как для генома в целом (по глобальным метавыравниваниям), так и для классов медленно, средне и быстро эволюционирующих генов (по глобальным метавыравниваниям) и, где это было возможно, для отдельных генов (по локальным метавыравниваниям).

Как в генах млекопитающих, так и в генах плодовых мушек плотность нуклеотидных замен в альтернативных областях выше, чем в постоянных: одновременно $d_N(A)>d_N(П)$ и $d_S(A)>d_S(П)$ (рисунок 4). Более того, частота аминокислотных замен на альтернативных участках выше, чем на постоянных, т. к. $d_N(A)>d_N(П)$, а давление отбора на аминокислотную последовательность на альтернативных участках меньше, чем на постоянных: $\omega(A)>\omega(П)$.

Таким образом, на альтернативных участках положительный отбор усиливается и / или отрицательный отбор ослабевает по сравнению с постоянными участками.

Распределение синонимичных замен в генах млекопитающих и мух отличается. В генах мух в альтернативных областях больше синонимичных замен, чем в постоянных областях, а в генах млекопитающих существенной разницы в плотности синонимичных замен в альтернативных и постоянных областях не наблюдалось. Анализ отдельных генов подтверждает эти закономерности.

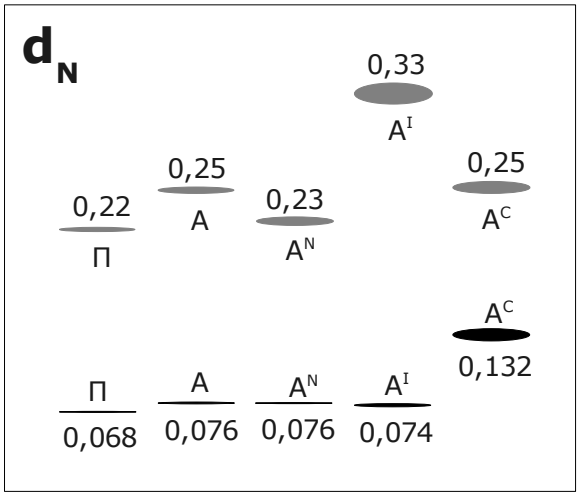
Среди всех классов альтернативных участков белков мух, аминокислотная последовательность N-концевых альтернативных участков наиболее консервативна, а внутренних альтернативных участков наименее консервативна, $d_N(A^N)<d_N(A^C)<d_N(A^I)$. При этом даже в наиболее консервативных альтернативных участках — N-концевых, плотность несинонимичных замен выше, чем в постоянных, $d_N(A^N)>d_N(П)$. В то время как различия между постоянными и альтернативными участками обусловлены большей плотностью нуклеотидных замен в целом, различия между классами альтернативных участков обусловлены разным соотношением синонимичных и несинонимичных замен: $d_N(A^N)<d_N(A^C)<d_N(A^I)$ и одновременно $d_S(A^N)>d_S(A^C)>d_S(A^I)$.

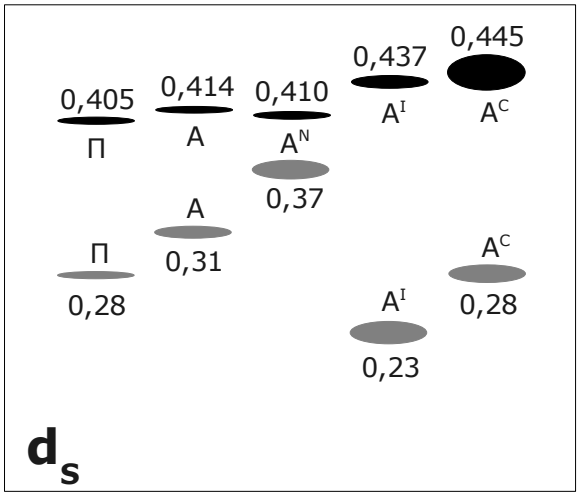
У млекопитающих плотность нуклеотидных замен на альтернативных участках распределена иначе, чем у мух. Плотность нуклеотидных замен в альтернативных участках возрастает в направлении от 5' конца к 3' концу гена, в то время как замены в постоянных участках распределены равномерно по всей длине гена. $d_S$ и ω неожиданно резко возрастают на С-концевых альтернативных участках. На качественном уровне соотношение скоростей замен в постоянных и альтернативных участках сохраняется для генов, эволюционирующих с разной скоростью.

**Обсуждение.** Эволюционное поведение различных функциональных участков генома существенно отличается. Гены с медленно эволюционирующей последовательностью дуплицируются чаще (Davis and Petrov 2004), хотя вскоре после дупликации скорость эволюции может увеличиться, т. к. отрицательный отбор ослабевает (Kondrashov et al 2002, Conant and Wagner 2003) и действие отбора на две копии может быть различным (Zhang et al 2003). Дюре и Муширу (Duret and Mouchiroud 2000) показали, что в генах, которые экспрессируются в большом числе тканей, уровень несинонимичных нуклеотидных замен ниже, чем в генах, которые экспрессируются в ограниченном числе тканей, а уровень синонимичных нуклеотидных замен в этих группах генов примерно одинаков. Пал, Папп и Хёрст (Pal et al 2001) показали, что интенсивно экспрессирующиеся гены более консервативны, чем гены, экспрессирующиеся менее интенсивно. Результаты, полученные в нашей работе, согласуются с этими наблюдениями, если принять, что постоянные участки экспрессируются в большем числе тканей и/или более интенсивно, чем альтернативные участки: первое объяснение работает для альтернативных участков тканеспецифичных изоформ, второе — для альтернативных участков изоформ, экспрессирующихся в различных тканях равномерно.

Молодые участки генов склонны эволюционировать быстро. Несколько исследований (Kondrashov et al 2002, Conant and Wagner 2003, Zhang et al 2003, Jordan et al 2004) показывают ослабление отрицательного отбора в паралогах вскоре после дупликации. В нашей работе показано, что в альтернативных участках генов положительный отбор сильнее и/или отрицательный отбор слабее, чем в постоянных, таким образом, эта закономерность может быть обобщена с целых генов на фрагменты генов.

Так как были рассмотрены только альтернативы, подтверждённые полноразмерными мРНК, находившимися на момент анализа в базах данных, некоторые возможности альтернативного сплайсинга могли быть упущены, и часть альтернативных участком могли



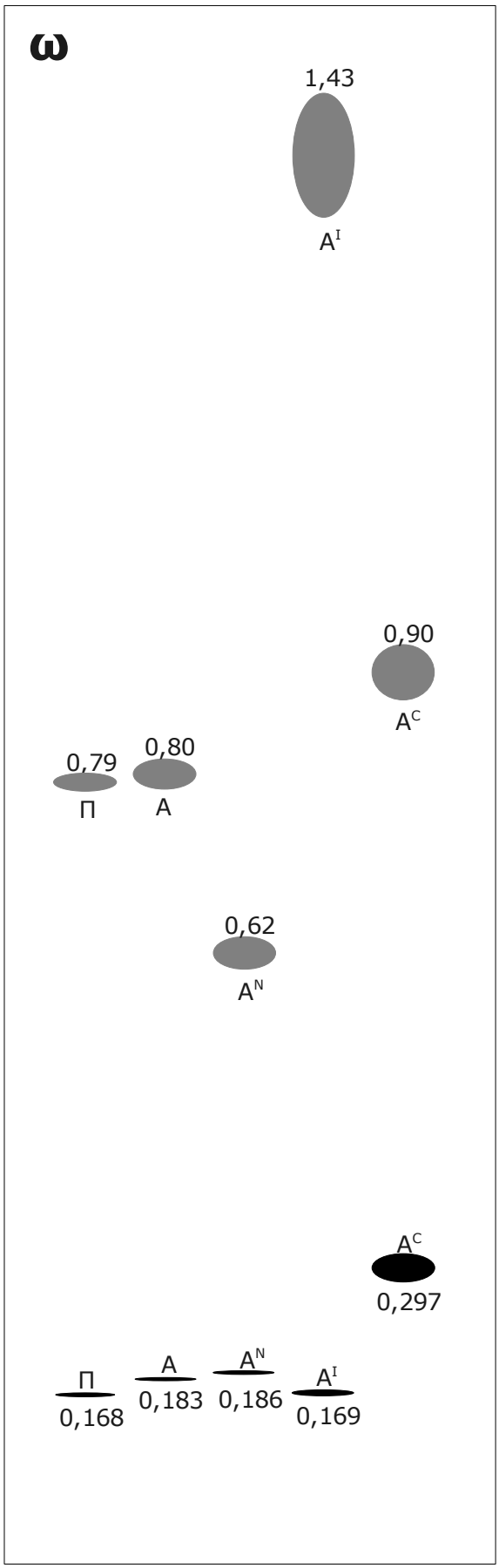




Рисунок 4. Оценки эволюционных параметров $d_N$, $d_S$ и ω, полученные с помощью глобальных метавыравниваний. Чёрным показаны результаты для человека и мыши, серым — для двух дрозофил. Указана медиана результатов оценивания параметров для 2000 выравниваний, полученных в результате бутстреппинга, высота эллипса равна $3(q_3 - q_1)$, где $q_1$ и $q_3$ — первый и третий квартили. Другие обозначения см. в тексте

быть аннотированы как постоянные. Но это могло лишь сделать наблюдаемые эффекты менее отчётливыми, но не изменить их содержание.

Недавно Цин и Ли (Xing and Lee 2005) описали такое же поведение несинонимичных замен в альтернативных и постоянных участках, но иное поведение синонимичных замен: они получили пониженный уровень синонимичных замен в альтернативных участках, особенно тканеспецифичных кассетных экзонах (Xing and Lee 2005). Одно из предложенных авторами объяснений этого эффекта — повышенное содержание энхансеров сплайсинга в альтернативных экзонах (Yeo et al 2005). Естественно предположить, что такие энхансеры будут консервативны, как в гене BRCA1 (Hurst and Pal 2001, Orban and Olah 2001), и это может привести к дополнительным требованиям консервативности синонимичных позиций. Однако, это объяснение, по-видимому, некорректно: хотя $d_S$ действительно ниже в энхансерах сплайсинга, доли постоянных и альтернативных участков, покрытых энхансерами, одинаковы (Parmley et al 2006), и если давление отбора на уровне мРНК в альтернативных и постоянных участках одинаково, на величину ω это не повлияет.

В нашем исследовании понижение $d_S$ на альтернативных участках не было обнаружено и уровни нуклеотидных замен отличаются от полученных в (Xing and Lee 2005). Так как полученные нами результаты согласованы между собой, статистически значимы для всех типов выборок генов и участков генов и не являются следствием недостаточной чистоты данных, должны быть другие объяснения такому несовпадению. Во-первых, в нашем исследовании принимались во внимание все типы альтернатив, а не только кассетные экзоны, как в (Xing and Lee 2005). При этом были рассмотрены короткие альтернативные участки. Во-вторых, при оценке эволюционных параметров были использованы разные методы: мы реализовали метод Ины (Ina 1995, см. Материалы и методы), в то время как Цин и Ли использовали метод наибольшего правдоподобия, реализованный в пакете PAML (Yang 1997, http://abacus.gene.ucl.ac.uk/software/paml.html). С другой стороны, мы рассматривали только альтернативы, подтверждённые полноразмерными мРНК, и не различали основные и минорные изоформы.

Возможно также, что общая длина регуляторных сайтов, подверженных отрицательному отбору мала по сравнению с общей длиной альтернативных участков. При этом ответственные за альтернативный сплайсинг *cis*-регуляторные элементы могли быть расположены вне самих участков, в том числе в некодирующих областях.

## Выводы

1. Показано, что альтернативно сплайсируемые пары донорных сайтов в геноме человека участвуют в контроле экспрессии генов на пост-транскрипционном уровне: наиболее распространены пары со сдвигом на 4 нуклеотида, смещающим рамку считывания, и в 61% случаев один из сайтов пары порождает транслируемую изоформу, в то время как вторая изоформа может стать мишенью нонсенс-мотивированной деградации мРНК.

2. Установлено, что в большинстве альтернативно сплайсируемых пар перекрывающихся донорных сайтов уровни экспрессии изоформ резко отличаются, и области значений весов сайтов и консенсусов последовательностей для однозначных пар и для двузначных пар с одним основным сайтом пересекаются. Таким образом, для выбора сайта в паре необходима дополнительная регуляция.

3. Разработана техника метавыравниваний, которая позволяет учитывать при анализе скоростей нуклеотидных замен даже небольшие альтернативные участки.

4. Показано, что уровень несинонимичных нуклеотидных замен в альтернативных областях генов выше, чем в постоянных.

5. Показано, что в альтернативных участках генов усилено действие положительного отбора, и/или ослаблено действие отрицательного отбора. Это может быть связано с относительной молодостью альтернативных участков.

6. Показаны таксоноспецифичные особенности эволюции альтернативных участков генов. У млекопитающих плотность как синонимичных, так и несинонимичных замен на альтернативных участках увеличивается в направлении от 5' к 3' концу и наблюдается резкий скачок на С-концевых альтернативных участках. У дрозофил суммарная плотность замен на альтернативных участках генов примерно постоянна, но доля синонимичных и несинонимичных среди них различна. Плотность синонимичных замен в синонимичных позициях выше всего на N-концевых альтернативных участках, а плотность несинонимичных замен в несинонимичных позициях — на внутренних альтернативных участках.

7. Показано, что внутренние альтернативные участки генов дрозофил находятся под положительным отбором: плотность несинонимичных замен в несинонимичных позициях превышает плотность синонимичных замен в синонимичных позициях почти в полтора раза.

8. Сделан общий вывод о том, что альтернативно сплайсируемые участки генов служат „экспериментальной площадкой“ молекулярной эволюции.

Я выражаю глубокую благодарность Михаилу Сергеевичу Гельфанду за чуткое научное руководство, постоянное внимание к моей работе и поддержку, а также искреннюю признательность Валентине Боевой, Ольге Калининой, Антону Митягину, Рамилю Нуртдинову, Дмитрию Малько и Дмитрию Виноградову.

## Список публикаций по теме диссертации

### Статьи

1. *Ermakova E.O., Nurtdinov R.N., Gelfand M.S.* Overlapping alternative donor splice sites in the human genome // J Bioinform Comput Biol. 2007. V. 5. №5. P.991-1004.

2. *Ermakova E.O., Nurtdinov R.N., Gelfand M.S.* Fast rate of evolution in alternatively spliced coding regions of mammalian genes // BMC Genomics. 2006. V. 7 №1. 84.

3. *Ермакова Е.О., Малько Д.Б., Гельфанд М.С.* Эволюционные отличия альтернативных и постоянных белок-кодирующих участков альтернативно сплайсируемых генов Drosophila // Биофизика. 2006. Т. 51. №4. С.581-588.

4. *Нуртдинов Р.Н., Неверов А.Д., Малько Д.Б., Космодемьянский И.А., Ермакова Е.О., Раменский В.Е., Миронов А.А., Гельфанд М.С.* EDAS — база данных альтернативно сплайсированных генов человека // Биофизика. 2006. Т. 51. №4. С.589-592.

### Тезисы конференций

1. *Ermakova E.O., Nurtdinov R.N., Gelfand M.S.* Overlapping alternative donor splice sites // Информационные технологии и системы. ИТиС’07, Звенигород, 18-21 сентября 2007. С.241-244.

2. *Malko D.B., Ermakova E.O.* Evolution of splicing in insects // Proceedings of the 3-rd Moscow Conference on Computational Molecular Biology. MCCMB’07, Москва, 27-31 июля 2007. P.193.

3. *Ermakova E.O., Nurtdinov R.N., Gelfand M.S.* Ovelapping alternative donor splicing sites in the human genome // ISMB/ECCB 2007 Proceedings. ISMB/ECCB’07, Вена, Австрия, 21-25 июля 2007.

4. *Ermakova E.O., Malko D.B., Gelfand M.S.* Patterns of selection and evolution of the exon-intron structure in alternatively spliced genes of nine Droshophila species and the malarial mosquito // ISMB/ECCB 2007 SIG Meetings Program Materials. 4th Special Interest Group Meeting on Alternative Splicing AS-SIG 2007, Вена, Австрия, 19-20 июля 2007. P.145-146.

5. *Ermakova E.O.* Evolutionary patterns in alternatively spliced coding regions of mammalian and Drosophila genes // Proceedings of the 11th Human Genome Meeting. HGM2006, Хельсинки, Финляндия, 31 мая - 3 июня 2006. P.54.

6. *Ермакова Е.О.* Точечные нуклеотидные замены и эволюция различных функциональных участков генома млекопитающих // Материалы Международной школы „Биоинформатика, геномика, протеомика“. Школа „Биоинформатика, геномика, протеомика“, Алма-Ата, Казахстан, апрель 2006. С.13-17.

7. *Ermakova E.O.* Alternatively spliced regions evolve faster // Proceedings of the International Moscow Conference on Computational Molecular Biology. MCCMB’05, Москва, июль 2005. P.95-96.

8. *Ермакова Е.О., Гельфанд М.С.* Положительный отбор в альтернативных областях генов человека // Материалы XII международной конференции студентов, аспирантов и молодых учёных „Ломоносов“. XII Международная конференция студентов, аспирантов и молодых учёных „Ломоносов“, Москва, апрель 2005. С.15-16.

Ермакова Екатерина Олеговна

## ОСОБЕННОСТИ ЭВОЛЮЦИИ РАЗЛИЧНЫХ ФУНКЦИОНАЛЬНЫХ ОБЛАСТЕЙ АЛЬТЕРНАТИВНО СПЛАЙСИРУЕМЫХ ГЕНОВ ЭУКАРИОТ

Рассматривались перекрывающиеся альтернативные донорные сайты сплайсинга, расположенные на расстоянии от 3 до 6 нуклеотидов друг от друга, и потенциальные сайты сплайсинга, находящиеся на таком же расстоянии от активного сайта сплайсинга. Показано, что альтернативно сплайсируемые пары донорных сайтов в геноме человека участвуют в контроле экспрессии генов на пост-транскрипционном уровне. В большинстве альтернативно сплайсируемых пар перекрывающихся донорных сайтов уровни экспрессии изоформ резко отличаются. Разработана техника метавыравниваний, которая позволяет учитывать при анализе скоростей нуклеотидных замен даже небольшие альтернативные участки. Изучено поведение точечных замен в альтернативно сплайсируемых кодирующих областях генов млекопитающих, на материале полных геномов человека и мыши, и насекомых, на примере полных геномов двух видов плодовой мушки. В альтернативных кодирующих участках генома нуклеотидные замены фиксируются чаще, чем в постоянных, и давление отбора ослаблено как на уровне мРНК, так и на уровне белка. Отдельно исследовано поведение нуклеотидных замен в концевых и внутренних участках гена. Рассмотрены таксоноспецифичные особенности эволюции альтернативных участков генов. Внутренние альтернативные участки генов дрозофилы находятся под положительным отбором, а в альтернативных участках генов человека, соответствующих С-концу белка, отрицательный отбор слабее и/или положительный отбор сильнее, чем в других альтернативных участках. Сделан общий вывод о том, что альтернативно сплайсируемые участки генов служат „экспериментальной площадкой“ молекулярной эволюции.

Ermakova Ekaterina Olegovna

## CHARACTERISTICS OF EVOLUTION OF DIFFERENT FUNCTIONAL REGIONS OF ALTERNATIVELY SPLICED EUKARYOTIC GENES

Overlapping alternative donor splice sites with the site shift from 3 through 6 nucleotides and similar potential splice sites were considered and their role in post-transcriptional expression control was described, showing dramatical differences of expression levels for most pairs of overlapping donor splice sites. The meta-alignments technique was developed. allowing us to analyze evolutionary patterns in relatively short alternatively spliced regions. It was applied to nucleotide substitutions in alternatively spliced genes of mammals (human and mouse genomes) and insects (two fluitfly genomes). Nucleotide substitutions are more abundant in alternative regions than in constitutive regions, and negative selection is reduced at the mRNA level and at the protein level. The pattern of nucleotide substitutions in internal and terminal regions is different. The evolution of alternative regions has taxon-specific features. The internal alternative regions of fruitfly genes evolve under positive selection. The negative selection is weaker and/or positive selection is stronger in the C-terminal alternative regions of mammals compared to other alternative regions. Overall, this study demonstrates that alternative splicing serves as a testing ground for molecular evolution.