

На правах рукописи



Курмангалиев Ербол Жанузакович

СРАВНИТЕЛЬНО-ГЕНОМНЫЙ АНАЛИЗ
ПОСТТРАНСКРИПЦИОННЫХ И ПОСТТРАНСЛЯЦИОННЫХ МЕХАНИЗМОВ
РЕГУЛЯЦИИ СТРУКТУРЫ И ФУНКЦИИ БЕЛКОВ

03.01.09 — математическая биология, биоинформатика

АВТОРЕФЕРАТ

диссертации на соискание учёной степени
кандидата биологических наук

Москва — 2011

Работа выполнена 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 физико-математических наук,

Макеев Всеволод Юрьевич,

Институт общей генетики им. Н.И.Вавилова РАН, Москва

кандидат биологических наук, профессор

Фришман Дмитрий Иосифович,

Технический университет Мюнхена, г. Мюнхен, Германия

Ведущая организация:

Учреждение Российской академии наук Институт молекулярной биологии

им. В.А. Энгельгардта РАН, Москва

Защита диссертации состоится 25 ноября 2011 года в 14-00 часов на заседании диссертационного совета Д 002.077.04 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 __ октября 2011 года

Ученый секретарь диссертационного совета

доктор биологических наук, профессор

Рожкова Г.И.



Общая характеристика работы

Актуальность темы

До полного секвенирования генома человека высказывались самые различные предположения об общем количестве генов. При этом достаточно общепринятой была точка зрения, что количество генов возрастает со сложностью биологических организмов. Поэтому неожиданными оказались результаты первичного анализа полного генома человека, который выявил менее 30000 генов. Для сравнения, в геноме круглого червя *Caenorhabditis elegans* содержится примерно 20000 генов.

Однако разнообразие белков не ограничивается лишь общим числом генов. У высших эукариот одним из основных механизмов повышения разнообразия генных продуктов на посттранскрипционном уровне является альтернативный сплайсинг. Существуют различные оценки общего числа альтернативно сплайсируемых генов человека. Массовое секвенирование EST-последовательностей и их анализ показали, что не менее трети генов человека альтернативно сплайсируются. Дальнейшее увеличение объема данных о транскриптоме человека только увеличивало эту оценку, которая сейчас достигла 90-95%.

С другой стороны, неясно, какая доля из предсказанных по EST-последовательностям вариантов транскриптов являются функциональными, а какая является результатом ошибок механизма сплайсинга или экспериментальными артефактами. Традиционно исследователей интересует в первую очередь функциональный альтернативный сплайсинг, и они пытаются очистить анализируемые наборы данных от нефункциональных транскриптов. В то же время, анализ последствий подобных ошибок может дать информацию о том, как функционирует сам механизм сплайсинга. Одной из ошибок сплайсинга является удержание интрона. Другим интересным объектом являются мутации в сайтах сплайсинга, нарушающие их распознавание, что может приводить к различным последствиям. По некоторым данным, мутации, нарушающие сплайсинг пре-мРНК, потенциально являются одной из наиболее частых причин наследственных заболеваний. Поэтому исследования последствий ошибок сплайсинга, и, в частности, мутаций, затрагивающих сплайсинг, может также иметь важное практическое значение.

В то же время, структурное и функциональное разнообразие генных продуктов не ограничивается лишь набором транскриптов, получаемых при альтернативном сплайсинге. Уже после синтеза белки могут подвергаться дальнейшим посттрансляционным модификациям путем ковалентного присоединения различных функциональных групп или протеолитического расщепления. Посттрансляционные модификации играют важнейшую роль в самых разнообразных клеточных процессах путем влияния на активность белков, их клеточную локализацию и взаимодействия с другими белками. Одним из наиболее важных и распространенных типов посттрансляционных модификаций эукариот является обратимое фосфорилирование белков по остаткам серина, треонина и тирозина. По некоторым оценкам, треть белков, закодированных в геноме человека, содержат ковалентно связанный фосфат. В последние годы, с развитием новых методов в протеомике и масс-спектропии, количество фосфосайтов, идентифицированных в белках различных модельных организмов, достигло десятков тысяч. Накопление таких больших объемов протеомных данных позволяет

проводить системные исследования посттрансляционных модификаций методами биоинформатики. Одной из таких задач является исследование возможной связи посттранскрипционных и посттрансляционных механизмов регуляции генов.

Другой интересный вопрос заключается в том, как эволюционируют сайты посттрансляционных модификаций. Как и все функционально важные участки белков, сайты модификаций более консервативны по сравнению с окружающими их аминокислотными остатками. Однако, кроме функциональной значимости, модифицированные аминокислоты отличаются от своих немодифицированных аналогов и по химическим свойствам. Поэтому мутации в сайтах модификаций и их немодифицированных аналогах, приводящие к их замене на другие типы аминокислот, скорее всего, будут иметь различное влияние на структуру и функцию белков. Это должно приводить к различиям в частотах замен модифицированных и немодифицированных аминокислот на другие типы аминокислотных остатков.

Цель и задачи исследования

Целью данной работы было исследование различных аспектов посттранскрипционной и посттрансляционной регуляции структуры и функции белков с помощью компьютерного анализа новейших доступных данных. В частности, были исследованы ошибки альтернативного сплайсинга и эволюционные паттерны посттрансляционных модификаций.

При этом были решены следующие задачи:

- собрана выборка интронов, для которых наблюдались случаи удержания в транскриптах;
- проведен сравнительный анализ удержанных и конститутивных интронов;
- собрана выборка экзонов, мутации в сайтах сплайсинга которых приводили к пропуску этих экзонов;
- проведен сравнительный анализ экзонов, в которых мутации в сайтах сплайсинга приводят к пропуску экзона и к активации скрытых сайтов сплайсинга;
- исследована возможная связь между альтернативным сплайсингом и посттрансляционными модификациями путем анализа распределения сайтов фосфорилирования между различными участками белков;
- проведена реконструкция эволюции сайтов посттрансляционных модификаций, в том числе, сайтов фосфорилирования и сайтов ацетилирования;
- проведен сравнительный анализ паттернов замен сайтов посттрансляционных модификаций и их немодифицированных аналогов на другие типы аминокислот.

Научная новизна и практическое значение

В данной работе был проведен систематический анализ ошибок сплайсинга. Выявлены достоверные различия между конститутивными и удержанными интронами по ряду параметров (в т.ч. по длине, по качеству сайтов сплайсинга, по плотности потенциальных цис-регуляторных элементов, по расположению в генах). Впервые было показано, что доля удержанных интронов монотонно возрастает от 5` - к 3` -концу транскриптов.

Проведен анализ экзонов с мутациями в сайтах сплайсинга. Собрана выборка экзонов, мутации сайтов сплайсинга в которых вызывают пропуск экзона. Выявлены значимые различия между экзонами, в которых мутации в сайтах сплайсинга приводят к пропуску экзона либо к активации скрытого сайта сплайсинга (в т.ч. по длине, по весу сайтов сплайсинга, по плотности потенциальных цис-регуляторных элементов, по наличию эквивалентных скрытых сайтов сплайсинга в непосредственной близости от сайтов с мутациями).

Получены данные о возможной связи альтернативного сплайсинга и посттрансляционных модификаций. Выдвинута гипотеза о различиях в частотах замен модифицированных и немодифицированных аминокислотных остатков на другие типы аминокислот. Впервые проведен эволюционный анализ паттернов замен сайтов посттрансляционных модификаций. На примере сайтов фосфорилирования выявлены значимые различия между векторами замен модифицированных и немодифицированных аминокислотных остатков. В частности показано, что фосфорсерины в среднем чаще заменяются на глутамат и аспартат, по сравнению с нефосфорилированными остатками серина. Реализованная методика анализа паттерна замен сайтов фосфорилирования может использоваться для исследования других типов посттрансляционных модификаций.

В целом, полученные в этом исследовании результаты могут использоваться при предсказании возможных последствий мутаций в сайтах сплайсинга и посттрансляционных модификаций, что может иметь практическое значение в исследованиях наследственных заболеваний и найти применение в персонализированной медицине.

Апробация работы

Материалы исследования по теме диссертации были представлены на следующих конференциях: XIV Международной конференции студентов, аспирантов и молодых ученых “Ломоносов” (Москва, апрель 2007); 3rd International Moscow Conference on Computational Molecular Biology (МССМВ’07, Москва, июль 2007); 30-й конференции “Информационные технологии и системы” (ИТиС’07, Звенигород, сентябрь 2007); 31-й конференции “Информационные технологии и системы” (ИТиС’08, Геленджик, сентябрь, 2008); 4th International Moscow Conference on Computational Molecular Biology (МССМВ’09, Москва, июль 2009); 33-й конференции “Информационные технологии и системы” (ИТиС’10, Геленджик, сентябрь 2010).

Объем и структура диссертации

Диссертационная работа изложена на 110 страницах и состоит из введения, трех глав, выводов, списка цитированной литературы и приложений. **Глава 1** содержит обзор литературы по теме диссертации. **Глава 2** содержит описание данных, методов и программного обеспечения использовавшихся при решении задач, поставленных в диссертации. **Глава 3** содержит описание полученных результатов и их обсуждение. Список литературы содержит 206 наименований. В приложениях приведены дополнительные материалы, не вошедшие в основные разделы диссертации. Работа содержит 21 рисунок и 10 таблиц.

Результаты и обсуждение

Ошибки сплайсинга

Сплайсинг, как и любой другой биологический процесс, подвержен ошибкам. Такой сплайсинг по сути также является альтернативным сплайсингом, и приводит к возникновению продуктов, отличных от канонических изоформ гена. В зависимости от того, идет ли распознавание сайтов сплайсинга через экзон или интрон, ошибки сплайсинга будут приводить к различным последствиям. Ошибки при распознавании через экзон должны приводить к полному удалению экзона или, при наличии сильного скрытого сайта, к использованию последнего. В то же время, ошибки при распознавании через интрон должны приводить к удержанию интронов.

В этой главе описаны результаты анализа ошибок сплайсинга в транскриптах человека. В частности, мы провели сравнительный анализ удержанных и конститутивно сплайсируемых интронов, а также мутаций сайтов сплайсинга, приводящих к пропуску экзона или активации скрытых сайтов сплайсинга.

Удержанные интроны

Удержание интрона является наименее изученным типом альтернативного и ошибочного сплайсинга. В отличие от других типов сплайсинга, заключающихся в выборе между различными сайтами сплайсинга, удержание интрона представляет собой полное отсутствие сплайсинга. Большинство случаев удержания интрона, наблюдаемых в базах данных транскриптомов человека, по всей видимости, являются результатом ошибок сплайсинга или присутствия в образце недосплайсированных транскриптов. Обычно такие транскрипты считаются экспериментальными артефактами и не рассматриваются при анализе альтернативного сплайсинга. Мы попытались рассмотреть удержанные интроны, встречающиеся в базах данных кДНК человека, с другой стороны: они являются промежуточными продуктами процесса сплайсинга и их анализ может дать дополнительную информацию о том, как происходит сплайсинг.

Для определения факторов, влияющих на удержание интронов, был проведен сравнительный анализ удержанных и конститутивно сплайсируемых интронов человека. Мы сравнили длины интронов и их фланкирующих экзонов, силу сайтов сплайсинга интрона, силу дистальных сайтов сплайсинга фланкирующих экзонов (акцепторный сайт 5'-экзона и донорный сайт 3'-экзона), плотности экзонных цис-регуляторных элементов сплайсинга, а также относительную локализацию интронов по длине генов.

Распределения длин интронов между удержанными и конститутивными интронами значительно различались (двухвыборочный тест Колмогорова-Смирнова, $P < 10^{-15}$). Удержанные интроны были значительно короче конститутивных: 84% удержанных интронов были короче 1000 н.п., по сравнению с 40% конститутивных. Медианы длин удержанных и конститутивных интронов были равны 337 и 1481 н.п. соответственно. В то же время, никакой значимой разницы в длинах фланкирующих экзонов замечено не было.

Сила сайтов сплайсинга интронов и их фланкирующих экзонов для удержанных и конститутивных интронов вычислялась с использованием позиционной матрицы весов.

Сайты сплайсинга удержанных интронов были более слабыми: распределения весов сайтов сплайсинга удержанных и конститутивных интронов значимо различались как для донорных, так и для акцепторных сайтов (двухвыборочный тест Колмогорова-Смирнова, $P < 10^{-15}$). Медианы весов донорных сайтов были равны 18.2 и 18.8 для удержанных и конститутивных интронов соответственно. Аналогично, для акцепторных сайтов медианы весов равнялись 18.0 для удержанных и 19.1 для конститутивных интронов.

Весы донорных сайтов 3'-фланкирующих экзонов удержанных и конститутивных интронов были схожими. В случае 5'-фланкирующего экзона веса акцепторных сайтов оказались значительно ниже у удержанных интронов, по сравнению с конститутивными: медианы 18.6 и 19.1 соответственно (двухвыборочный тест Колмогорова-Смирнова, $P < 10^{-15}$).

Потенциальные цис-регуляторные элементы сплайсинга предсказывались в обоих классах интронов с помощью трех опубликованных программ ESEfinder, RESCUE-ESE, и PESX. Плотность большинства предсказанных экзонных энхансеров сплайсинга (exonic splicing enhancers) была значительно выше в удержанных интронах, тогда как плотность экзонных сайленсеров сплайсинга (exonic splicing silencers) была выше в конститутивных интронах (Рисунки 1 и 2).

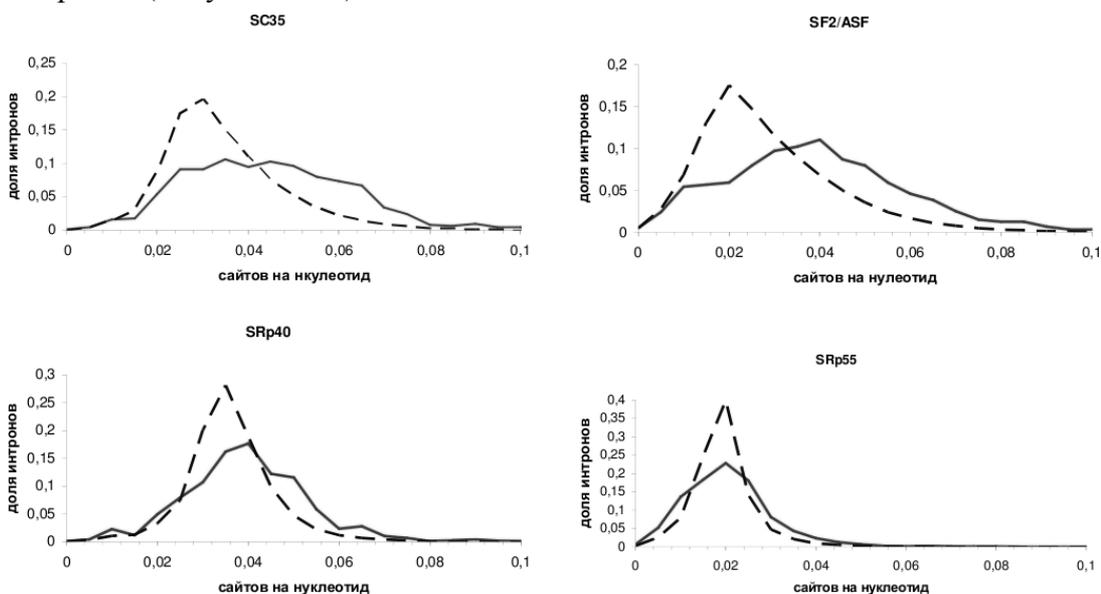


Рисунок 1. Распределение плотностей ESE-мотивов предсказанных программой ESEfinder. Сплошная: удержанные интроны; пунктирная: конститутивные интроны.

При этом средняя плотность всех четырех мотивов ESEfinder была выше в удержанных интронах (Рисунок 1). Наибольшая разница между медианами плотностей наблюдалась для сайтов связывания SF2/ASF (медианы плотностей 0.040 и 0.028 для удержанных и конститутивных интронов соответственно), в то время как наименьшая разница была в случае сайтов связывания SRp55 (медианы плотностей 0.0217 и 0.0215, незначимая разница). Плотности PESE-октамеров (энхансеры) также были выше в удержанных интронах, тогда как плотность PESS-октамеров (сайленсеры) была выше в конститутивных интронах (Рисунок 2). В отличие от них, плотность ESE-гексамеров, предсказанных RESCUE-ESE, была значительно выше в конститутивно сплайсируемых интронах, чем в удержанных (Рисунок 2). Все эти различия были статистически значимыми (двухвыборочный тест Колмогорова-Смирнова, $P < 10^{-15}$).

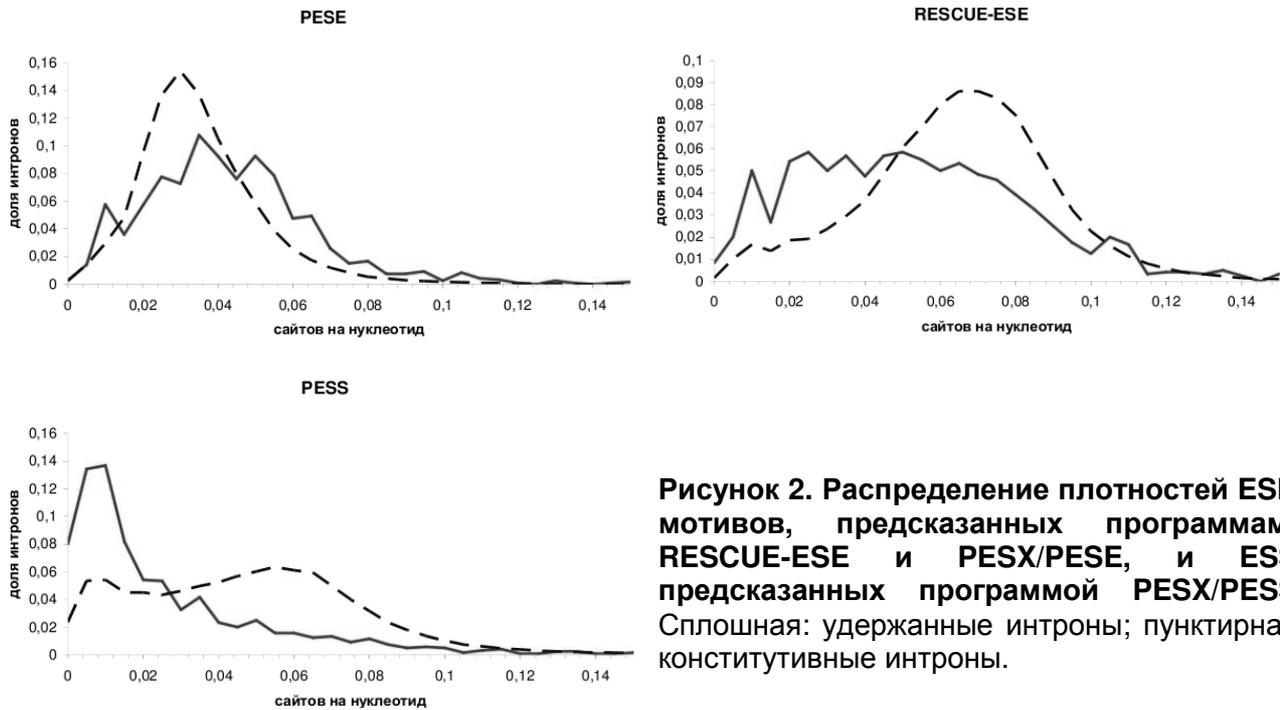


Рисунок 2. Распределение плотностей ESE-мотивов, предсказанных программами RESCUE-ESE и PESX/PESE, и ESS, предсказанных программой PESX/PESS. Сплошная: удержанные интроны; пунктирная: конститутивные интроны.

Относительная позиция интрона в гене была определена как отношение $ОП=P/D$, где P было расстоянием от 5'-конца гена до 5'-конца интрона (донорный сайт), и D было длиной гена (расстоянием между 5'- и 3'-концами, в соответствии с аннотацией RefSeq). Так как терминальные интроны и экзоны могут иметь нестандартное распределение длины, вычисления проводились несколькими способами. Во-первых, мы использовали несплайсированные гены, по аннотациям в RefSeq, и в таких случаях расстояние вычислялись по геномным последовательностям. Во-вторых, мы рассматривали сплайсированные гены: все интроны были удалены и изучаемый интрон был сжат до одной точки ("тень интрона"); расстояния вычислялись по последовательностям мРНК. В-третьих, мы рассматривали сплайсированные гены с удаленным последним экзоном. И, наконец, мы определяли позицию интрона как отношение порядкового номера изучаемого интрона (от начала гена) к их общему числу в гене.

Конститутивные интроны (показаны светлым на Рисунке 3) смещены к 3'-концам несплайсированных генов (Рисунок 3B), и к 5'-концам при вычислениях на сплайсированных транскриптах (Рисунок 3C). Это согласуется с понижением плотности интронов и повышением длины экзонов по направлению от 5'-концов к 3'-концам генов. При удалении последнего 3'-терминального экзона, распределение становится почти равномерным (Рисунок 3D).

В случае с удержанными интронами ситуации абсолютно другая (двухвыборочный тест Колмогорова-Смирнова, $P < 10^{-15}$ для относительной позиции интронов в случае со сплайсированными генами и сплайсированными генами без последнего экзона, и $P < 10^{-9}$ для несплайсированных генов; тест Хи-квадрат, $P < 10^{-15}$ для порядкового номера интрона). Распределение удержанных интронов (показаны темным на Рисунке 3) значительно смещено в направлении 3'-конца генов во всех случаях, по сравнению с конститутивными интронами. Таким образом, доля удержанных интронов монотонно увеличивается в направлении от 5'-конца к 3'-концу генов (кривая треугольников, Рисунок 3).

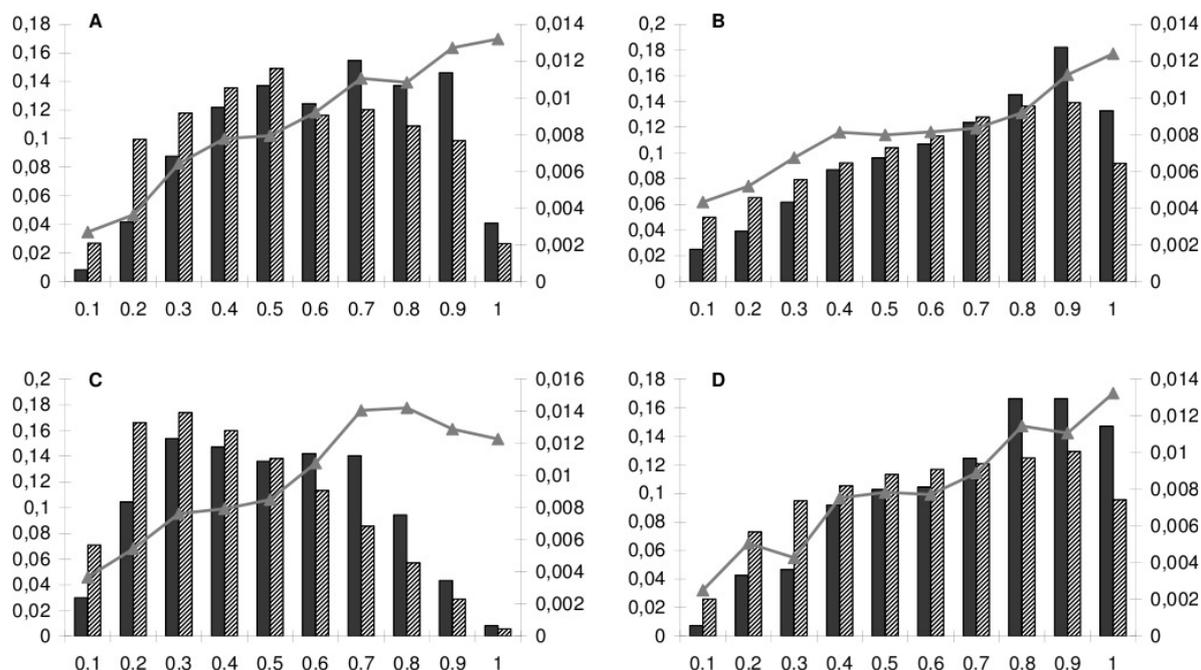


Рисунок 3. Распределение относительной позиции интронов. А: относительный порядковый номер интрона; В: несплайсированные гены; С: сплайсированные гены; D: сплайсированные гены без последнего экзона. Левая вертикальная ось: доля интронов в каждом сегменте, отдельно для удержанных (темное) и конститутивных (светлое) интронов. Точки 0 и 1 на горизонтальной оси соответствуют 5`- и 3`-концам генов, соответственно. Правая вертикальная ось и кривая треугольников: доля удержанных интронов среди всех интронов в данном сегменте.

Мутации в сайтах сплайсинга

Еще одним способом изучения механизмов сплайсинга является анализ последствий мутаций в сайтах сплайсинга. Такие мутации имеют два основных возможных последствия: пропуск экзона или активацию скрытого сайта сплайсинга, в то время как удержание интрона происходит достаточно редко. Одним из наиболее важных условий активации скрытого донорного сайта является наличие достаточно сильного кандидата в непосредственной близости от мутированного сайта.

Выборки экзонов с мутациями, нарушающими сайты сплайсинга, были получены из базы данных Менделевских наследственных заболеваний OMIM и литературы. Рассматривались лишь мутации, непосредственно затрагивающие донорные и акцепторные сайты сплайсинга. Экзоны, в зависимости от того, вызывают ли они пропуск экзона либо активацию скрытого сайта, разделялись на S-экзоны и C-экзоны, соответственно. Мутации донорных и акцепторных сайтов рассматривались как отдельно, так и вместе, для увеличения статистической значимости наблюдений. Результаты сравнительного анализа S-экзонов и C-экзонов просуммированы в Таблице 1.

S-экзоны оказались значимо короче C-экзонов (медианы длин 114 и 136). Никакой значимой разницы в длинах фланкирующих интронов замечено не было.

Были сравнены веса самих сайтов с мутациями, а также соседних сайтов (для экзонов с мутациями в донорных сайтах – ближайшие акцепторные сайты; аналогично, для экзонов с мутациями в акцепторных сайтах – ближайшие донорные сайты с 5`- и 3`-стороны). Акцепторные сайты, подверженные мутациям, оказались слабее в C-экзонах по сравнению с S-экзонами, с медианами весов 18.72 и 19.59 соответственно (тест Манна-Уитни, $P=0.05$).

Никакой значимой разницы между S-экзонами и C-экзонами в распределениях весов донорных сайтов, подверженных мутациям, а также всех остальных рассмотренных сайтов, обнаружено не было.

Далее, мы анализировали экзоны на наличие потенциальных скрытых сайтов сплайсинга в непосредственной близости от сайтов с мутациями. Для оценки относительного обогащения потенциальными скрытыми сайтами областей, окружающих сайты с мутациями, мы провели поиск эквивалентных сайтов сплайсинга в пределах близлежащих от изучаемого сайта интрона и экзона. Под эквивалентным сайтом сплайсинга мы понимали потенциальный сайт сплайсинга того же типа, что и изучаемый сайт, с весом равным весу последнего или выше. Сайты с мутациями, для которых не удавалось найти эквивалентный сайт в близлежащих интроне и экзоне, исключались из рассмотрения. Относительное обогащение областей вблизи от мутированных сайтов потенциальными скрытыми сайтами сплайсинга оценивали по расстоянию до ближайшего эквивалентного сайта. И в случае мутаций в донорных, и в акцепторных сайтах, разница между S-экзонами и C-экзонами была очень велика: эквивалентные сайты находились намного ближе к сайтам с мутациями C-экзонов, чем S-экзонов. Медиана расстояния от донорного сайта с мутацией C-экзонов до ближайшего эквивалентного сайта была равна 75 нукл., в то время как в случае S-экзонов медиана равнялась 220 нукл.. Аналогично, в случае акцепторных сайтов медианы расстояния до ближайших эквивалентных сайтов были равны 66 нукл. и 185 нукл. для C-экзонов и S-экзонов, соответственно.

Плотность потенциальных цис-регуляторных элементов в S-экзонах и C-экзонах вычислялась, так же, как и при сравнении удержанных и конститутивных интронов. В большинстве случаев медиана плотности предсказанных экзонных энхансеров сплайсинга (ESE) была выше в C-экзонах по сравнению с S-экзонами. В тоже время, медиана плотности экзонных сайленсеров сплайсинга (ESS), наоборот, была выше в S-экзонах (Таблица 1). Однако в связи с малыми размерами выборок, эта разница была значима лишь в нескольких случаях: медианы плотностей были статистически значимо выше в C-экзонах, чем в S-экзонах с мутациями в донорных сайтах, для сайтов связывания SF2/ASF (тест Манна-Уитни, $P=0,048$) и SRp40 (тест Манна-Уитни, $P=0,006$), а также октамеров PESE (тест Манна-Уитни, $P=0,007$).

Обсуждение результатов анализа удержанных интронов и мутаций в сайтах сплайсинга

В целом, результаты данного исследования хорошо согласуются с существующими биологическими моделями. Удержанные интроны относительно короткие – поэтому попарное распознавание сайтов, вероятнее всего, происходит через интрон, что в случае ошибок сплайсинга должно приводить к удержанию интронов.

Относительно низкий вес сайтов сплайсинга в удержанных интронах, как и отсутствие потенциальных скрытых сайтов с достаточно большим весом в непосредственной близости от аутентичных сайтов S-экзонов, свидетельствует о том, что веса сайтов является адекватной оценкой силы сайтов сплайсинга и их функциональности. При этом, относительная редкость потенциальных скрытых сайтов в непосредственной близости от C-

экзонов не ограничивается исключительно экзонами с мутациями в донорных сайтах. С другой стороны, мы не смогли подтвердить наблюдение о том, что для С-экзонов с мутациями в донорных сайтах характерны сильные акцепторные сайты.

В отличие от предыдущих исследований, которые интересовались, в первую очередь, функциональным (консервативным) удержанием интронов, мы не акцентировали внимание на возможной функциональности. Одним из последствий этого является то, что большинство удержанных интронов, рассмотренных в данном исследовании, по всей видимости, не функциональны, так как всего 3,3% из них сохраняют рамку считывания. Но это не исключает возможную роль таких интронов в регуляции на уровне белков или мРНК. Однако и сама процедура, и полученные результаты указывают на то, что большая часть удержанных интронов в нашем исследовании принадлежат к недосплайсированным транскриптам. Таким образом, низкие веса сайтов удержанных интронов могут иметь два объяснения. Во-первых, удержанные интроны могут происходить от недосплайсированных транскриптов (слабые сайты приводят к низкой эффективности сплайсинга). Во-вторых, удержанные интроны могут быть примером регулируемого альтернативного сплайсинга: известно, что функциональные альтернативные сайты сплайсинга имеют более низкий вес, чем конститутивные сайты сплайсинга.

Было показано, что удержанные интроны человека и растений чаще находятся в 5'- и особенно в 3'-нетранслируемых областях транскриптов, по сравнению с кодирующими областями мРНК. Это объяснялось уничтожением плохо сплайсированных копий мРНК с помощью механизма нонсенс-мотивированной деградации. Однако это не объясняет наблюдаемое преобладание удержанных интронов в 5'-нетранслируемых областях. Наши наблюдения показывают монотонное увеличение доли удержанных интронов по направлению от 5'-конца к 3'-концу генов. Это в некоторой степени согласуется с моделью котранскрипционного сплайсинга (в отличие от предположений о том, что сплайсинг одновременно начинается на всех интронах), которая была подтверждена в эксперименте.

Наблюдаемая разница в плотностях экзонных энхансеров сплайсинга между удержанными и конститутивными интронами, а также между S-экзонами и C-экзонами, тоже имеет естественное биологическое объяснение. Так, высокая плотность ESE-подобных сайтов на относительно коротких интронах может приводить к ошибочному распознаванию данного интрона как части экзона вместе с фланкирующими экзонами. Аналогично, высокая плотность экзонных энхансеров сплайсинга в экзонах с мутациями в сайтах сплайсинга может заставить аппарат сплайсинга сохранить данный экзон и использовать скрытый сайт, в то же время экзонные сайленсеры сплайсинга могут вызывать пропуск экзона. В то же время, неясно почему потенциальных энхансеров предсказанных RESCUE-ESE было больше в конститутивных интронах, чем в удержанных.

Наблюдаемая статистически значимая разница в весах аутентичных акцепторных сайтов С-экзонов и S-экзонов может иметь более сложное, но все же правдоподобное объяснение: экзоны со слабыми сайтами сплайсинга уже содержат больше энхансеров сплайсинга, по сравнению с экзонами с сильными сайтами, и, таким образом, имеют больше шансов стать С-экзонами.

	S-экзоны	C-экзоны	МУ	все внутренние экзоны
Длина экзона (нуклеотиды)				
мутации в донорных сайтах	114	147	0.024	
мутации в акцепторных сайтах	112.5	130	н/з	
все	114	136	0.020	123
Плотность цис-регуляторных элементов (предсказанных сайтов на нуклеотид)				
<i>ESEfinder: SC35</i>				
мутации в донорных сайтах	0.043	0.042	н/з	
мутации в акцепторных сайтах	0.038	0.045	н/з	
все	0.042	0.043	н/з	0.038
<i>ESEfinder: SF2/ASF</i>				
мутации в донорных сайтах	0.025	0.037	0.048	
мутации в акцепторных сайтах	0.036	0.041	н/з	
все	0.028	0.040	0.005	0.036
<i>ESEfinder: SRp40</i>				
мутации в донорных сайтах	0.034	0.043	0.006	
мутации в акцепторных сайтах	0.040	0.043	н/з	
все	0.035	0.043	0.004	0.040
<i>ESEfinder: SRp55</i>				
мутации в донорных сайтах	0.028	0.024	н/з	
мутации в акцепторных сайтах	0.022	0.023	н/з	
все	0.025	0.023	н/з	0.023
<i>RESCUE-ESE</i>				
мутации в донорных сайтах	0.090	0.108	н/з	
мутации в акцепторных сайтах	0.100	0.080	н/з	
все	0.091	0.094	н/з	0.099
<i>PESE</i>				
мутации в донорных сайтах	0.048	0.082	0.007	
мутации в акцепторных сайтах	0.057	0.055	н/з	
все	0.055	0.064	0.023	0.064
<i>PESS</i>				
мутации в донорных сайтах	0.012	0.008	н/з	
мутации в акцепторных сайтах	0.009	0.007	н/з	
все	0.011	0.007	н/з	0.007
Вес сайта сплайсинга				
<i>мутации в донорных сайтах</i>				
исходный донорный сайт	18.52	18.49	н/з	18.82
акцепторный сайт 5`-экзона	18.70	19.67	н/з	19.08
акцепторный сайт 3`-экзона	19.37	18.98	н/з	19.09
<i>мутации в акцепторных сайтах</i>				
исходный акцепторный сайт	19.59	18.72	0.050	19.08
донорный сайт 5`-экзона	18.44	18.56	н/з	18.82
донорный сайт 3`-экзона	18.48	18.51	н/з	18.79
Расстояния до ближайшего эквивалентного сайта (нуклеотиды)				
мутации в донорных сайтах	220	75	0.067	289
мутации в акцепторных сайтах	185	66	0.024	81

Таблица 1. Свойства S-экзонов, C-экзонов и всех внутренних экзонов из нашей выборки генов RefSeq. Для всех параметров экзонов указаны медианы. МУ: статистическая значимость различий между S-экзонами и C-экзонами, вычисленная тестом Манна-Уитни; н/з: не значимое.

Альтернативный сплайсинг и посттрансляционные модификации

Фосфорилирование является одним из наиболее важных и распространенных типов посттрансляционных модификаций белков. С другой стороны, у многоклеточных эукариот альтернативный сплайсинг является основным механизмом повышения структурного и функционального разнообразия белков. Для изучения возможной связи между посттрансляционными модификациями и альтернативным сплайсингом мы проанализировали распределение сайтов фосфорилирования в конститутивных и альтернативно сплайсируемых областях белков человека.

В этом анализе использовались сайты фосфорилирования, идентифицированные в высокопроизводительном эксперименте в клетках человека линии HeLa. Сайты были картированы на белки из базы данных EDAS, причем для каждого гена была выбрана наиболее длинная белковая изоформа. Далее, для каждого аминокислотного остатка модифицированных белков были получены данные по альтернативному сплайсингу, в том числе “включаемость” кодонов (доля последовательностей EST, содержащих кодон, среди всех последовательностей EST покрывающих данную область гена). В качестве контрольных выборок использовались кодоны тех же генов, кодирующие нефосфорилированные аналоги модифицированных аминокислот.

В зависимости от уровня “включаемости” сегментов, включающих анализируемые аминокислотные остатки, кодоны были классифицированы как конститутивные (принадлежащие к конститутивным сегментам, включаемость 100%) и альтернативно сплайсируемые (принадлежащие к альтернативно сплайсируемым сегментам белков, включаемость менее 99%). Кроме того, около 30% всех остатков, включая сайты фосфорилирования, имели долю включения 99%. Мы их отнесли к группе неопределенных, так как это могло быть как результатом альтернативного сплайсинга, так и следствием ошибок сплайсинга или экспериментальных артефактов.

Анализ данных показал, что сайты фосфорилирования имеют тенденцию находиться в альтернативно сплайсируемых областях белков. Во всех случаях эта тенденция слаба, но статистически достоверна. В 24% случаев сайты фосфорилирования были альтернативно сплайсируемыми, а в 46% случаев – конститутивными. В то же время, для нефосфорилированных остатков эти величины равны 21% и 51%, соответственно (тест Хи-квадрат, $P < 10^{-15}$). Аналогично, при отдельном рассмотрении фосфорилированных серинов, 24% находились в альтернативно сплайсируемых областях и 45% в конститутивных (тест Хи-квадрат, $P = 2 \times 10^{-8}$). Сходные данные получены в случае фосфорилированных треонинов – 25% и 44%, соответственно (тест Хи-квадрат, $P = 1 \times 10^{-4}$). В то же время, для относительно небольшой выборки фосфотиозинов значимой разницы не обнаружено.

Результаты, полученные в этом исследовании, согласуются с ранними наблюдениями о корреляции альтернативного сплайсинга и функционально важных участков белков. Другим возможным объяснением полученного результата может являться то, что, как для сайтов фосфорилирования, так и для альтернативно сплайсируемых областей показана тенденция к расположению на поверхности белков. Кроме того, на предсказанных структурах белков было показано, что фосфосерины и фосфотреонины предпочитают

неструктурированные области, в то время как фосфотирозины преимущественно располагаются в областях с упорядоченной вторичной структурой. В свою очередь, сами неструктурированные области белков часто кодируются в альтернативно сплайсируемых сегментах генов. Эти гипотезы можно будет проверить после того, как станут доступны пространственные структуры достаточного количества фосфорилируемых белков.

Паттерны эволюции сайтов посттрансляционных модификаций

Важнейшая роль посттрансляционных модификаций в различных внутриклеточных процессах предполагает консервативность таких сайтов в процессе эволюции. В ряде работ проводился эволюционный анализ сайтов фосфорилирования и было показано, что фосфосайты более консервативны, чем их нефосфорилированные аналоги, расположенные в тех же областях белков. В настоящей работе мы исследовали другой аспект эволюции сайтов посттрансляционных модификаций. Так как модифицированные аминокислоты химически являются отдельными типами аминокислот, в процессе эволюции они могут вести себя отлично от своих немодифицированных аналогов (с точки зрения частот замен на другие типы аминокислот).

Для изучения различий в эволюции между модифицированными аминокислотами и их немодифицированными аналогами мы реконструировали эволюцию фосфорилированных остатков серина и треонина. В частности, мы изучали эволюцию фосфосеринов человека, плодовой мушки и дрожжей. Кроме того, была проведен эволюционный анализ другого распространенного типа посттрансляционных модификаций, сайтов ацетилирования лизинов в белках человека.

Паттерны замен сайтов фосфорилирования

Для каждого из трех организмов мы использовали наборы сайтов фосфорилирования, полученных в более чем одном массовом эксперименте. Фосфорилированию чаще всего подвергаются остатки серина, реже треонина, и ещё реже тирозина. В связи с очень маленькими размерами выборок фосфотирозинов, этот тип сайтов был исключен из анализа.

Фосфорилирование белков – процесс динамичный, поэтому пересечение между наборами фосфосайтов, идентифицированных в различных экспериментах, из различных клеточных линий и тканей, было относительно небольшим. Сайты, которые были фосфорилированы в более чем одном массовом эксперименте, по всей видимости, модифицируются конститутивно, или, как минимум, представляют собой выборку наиболее надежных фосфосайтов.

Мы исследовали эволюцию модифицированных и немодифицированных остатков серина и треонина отдельно в восьми видах позвоночных, одиннадцати видах дрозофил и пятнадцати видах грибов. Для каждого сайта фосфорилирования мы реконструировали эволюцию в соответствующей группе организмов с помощью модифицированного алгоритма максимального правдоподобия. Так как мы не могли определить точный момент в ходе эволюции, начиная с которого данный сайт стал модифицироваться, мы приняли, что этот момент совпадает с моментом появления самой первой аминокислоты того же типа в данном дереве. Далее мы считали количество замен предковых потенциальных сайтов модификации на другие типы аминокислот, и вычисляли вектора замен.

Эволюционные особенности и частота фосфосеринов и фосфотреонинов сильно зависят от структурного контекста, и они очень часто располагаются в неструктурированных областях белков. Известно, что неструктурированные и регулярные области белков сильно различаются по аминокислотному составу. Чтобы избежать влияния этого факта на результаты нашего анализа, мы разделили все фосфосайты и немодифицированные остатки серина и треонина из контрольных выборок на сайты, расположенные в регулярных и неструктурированных областях белков (см. ниже).

Лишь часть фосфосеринов из наших начальных выборок имели замены на другие аминокислоты. Очень небольшая доля из них были локализованы в регулярных участках белков. Поэтому дальнейший анализ проводили только на остатках серина и треонина из неструктурированных участков белков. Конечные выборки фосфорилированных и нефосфорилированных остатков серина и треонина включали сайты с хотя бы одной заменой на другие аминокислоты, расположенные в неструктурированных областях фосфопротеинов. Отдельно рассматривалась подвыборка сайтов фосфорилирования, которые были подтверждены в более чем одном эксперименте (выборка “надежных сайтов”).

Для оценки статистической значимости различий между векторами замен мы применили процедуру бутстрепа контрольных выборок немодифицированных остатков серина и треонина. Размер каждой из 10000 случайных контрольных выборок соответствовал размеру анализируемой выборки фосфосайтов. Так, к примеру, в случае фосфосеринов человека каждая из контрольных выборок содержала по 4277 нефосфорилированных остатков серина. В случае же сравнительного анализа подвыборки фосфосайтов человека, обнаруженных в более чем одном массовом эксперименте, бутстрепы содержали по 906 контрольных сайтов.

В то же время, особенности белковых областей, окружающих сайты фосфорилирования, скорее всего, не ограничиваются лишь их неструктурированностью. В частности, важную роль в фосфорилировании может иметь вторичная структура, экспозиция на поверхности белка, доступность растворителям и т.п. Поэтому для уменьшения влияния данных факторов на результаты нашего сравнительного анализа, мы ввели дополнительные контрольные выборки. Они включали нефосфорилированные остатки серина и треонина, находящиеся в тех же областях белков, что и сайты модификаций. Размеры дополнительных контрольных выборок в каждом случае равнялись размерам анализируемых выборок, содержащих все фосфосайты. Достаточное количество немодифицированных остатков серина было набрано на максимальном расстоянии 10, 11 и 9 а.о. от фосфосеринов для дрожжей, дрозофилы и человека соответственно. В случае фосфотреонинов эти расстояния равнялись 29, 26, 20 а.о., соответственно.

Результаты сравнительного анализа векторов замен фосфорилированных и нефосфорилированных остатков серина показаны на Рисунке 6 (левая панель). Частоты замен фосфосеринов и серинов из контрольных выборок сильно варьировали между различными группами организмов. Однако некоторые тенденции были общими для всех таксонов и статистически значимыми.

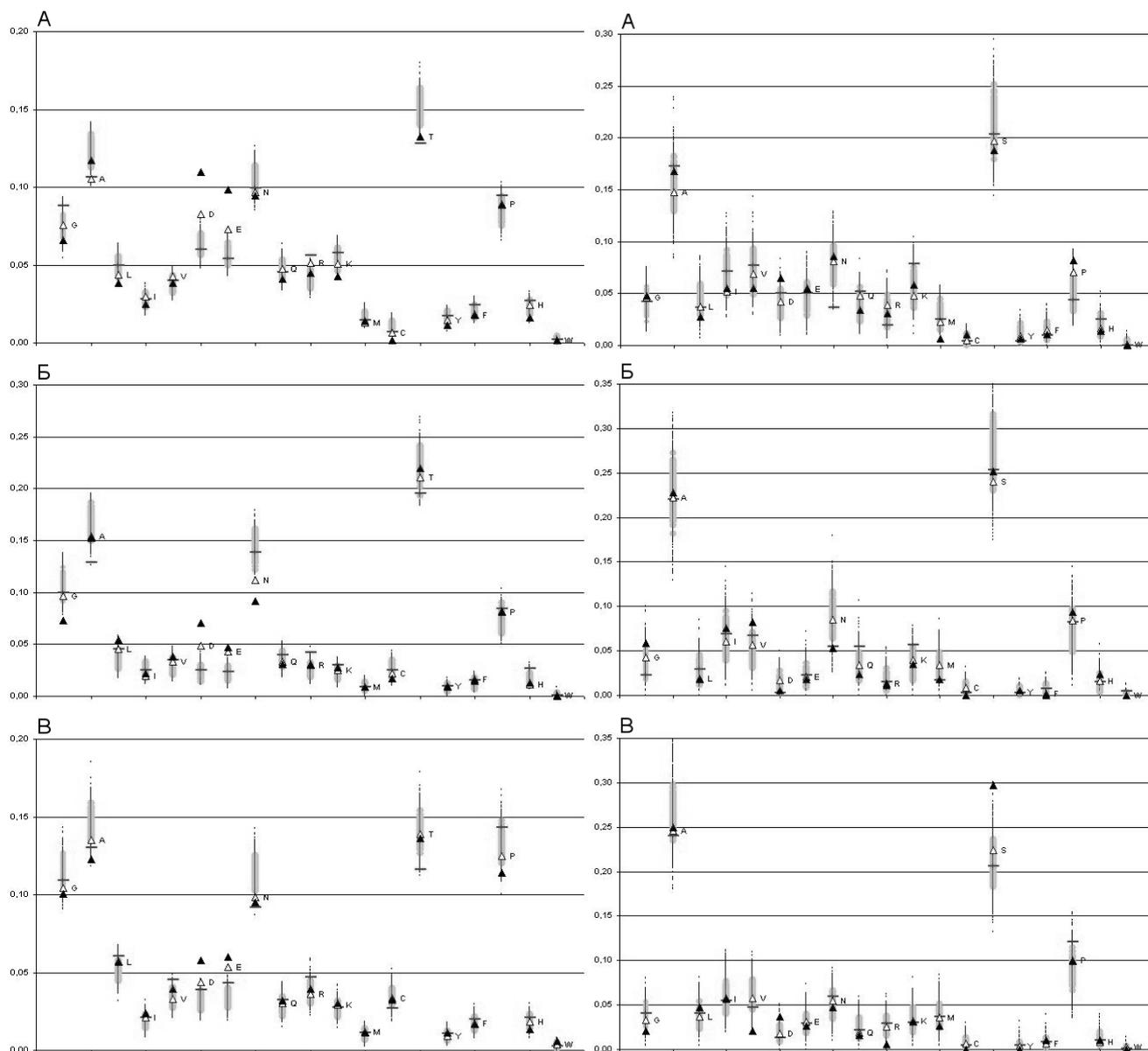


Рисунок 4. Вектора замен серинов (левая панель) и треонинов (правая панель) на другие типы аминокислот. Частоты замен серинов и треонинов из неструктурированных участков фосфопротеинов среди грибов (А), дрозофил (Б) и позвоночных (В): для всех фосфосайтов – белые треугольники; для фосфосайтов, обнаруженных в нескольких экспериментах – черные треугольники; 10000 контрольных выборок – облака больших светло-серых и маленьких темно-серых точек, соответственно; дополнительные контрольные выборки ближайших к фосфосайтам нефосфорилированных серинов и треонинов – горизонтальные полосы.

Вопреки ожиданиям, предпочтения замен фосфосеринов на треонины и тирозины, которые также являются фосфорилируемыми аминокислотами, обнаружено не было. В то же время, при фосфорилировании остаток серина приобретает отрицательный заряд, и, как видно из полученных результатов, во всех трех группах организмов в процессе эволюции фосфосерины намного чаще заменялись на отрицательно заряженные остатки глутамата и аспартата (по сравнению с нефосфорилированными остатками серина). В обоих случаях частота замен остатков фосфосерина была намного выше, чем во всех 10000 случайных контрольных выборках ($P \ll 10^{-4}$). В случае подвыборки более надежных фосфосайтов, подтвержденных в нескольких экспериментах, частота замен на аспартат и глутамат была

еще выше, и также лежала вне интервала частот, соответствующих 10000 бутстрепов (который в данном случае шире, так как размеры выборок значительно меньше).

Кроме этого, был и ряд других тенденций, общих для всех трех таксонов. В частности, фосфосерины относительно реже заменялись на остатки аланина, цистеина и аспарагина, по сравнению с контрольными выборками (Рисунок 4, левая панель). Однако в этих случаях немодифицированные серины из тех же областей белков (дополнительные контрольные выборки) также заменялись на данные аминокислоты реже остальных нефосфорилированных серинов. Вероятнее всего, эти сдвиги в частотах замен связаны не с посттрансляционными модификациями, а с особенностями участков белков, в которых расположены фосфосайты.

В случае же фосфотреонинов никаких значимых и общих для всех трех таксонов различий между частотами замен фосфосайтов и треонинами из контрольных выборок отмечено не было (Рисунок 4, правая панель). Это может быть связано с тем, что размеры выборок фосфотреонинов были значительно меньше размеров выборок фосфосеринов. В частности, как и в случае фосфосеринов, не было замечено никакого предпочтения к заменам фосфосайтов на другие фосфорилируемые типы аминокислот (серины и тирозины).

Паттерны замен сайтов ацетилирования

В дополнение к анализу эволюции фосфорилирования белков, мы провели аналогичное исследование сайтов ацетилирования лизина, также широко распространенного и консервативного типа посттрансляционных модификаций. В анализ были включены массовые протеомные данные по ацетилированию белков человека, полученные в двух высокопроизводительных экспериментах. Эти наборы данных были получены из различных типов клеток. Более того, различались и субклеточные фракции, из которых выделялись ацетилированные пептиды. Пересечение между наборами ацетилированных белков, идентифицированных в этих двух исследованиях, было крайне малым (8% белков были общими для двух наборов ацетилпротеинов). На уровне конкретных сайтов ацетилирования, это пересечение было ещё меньше – мы обнаружили всего 2% ацетиллизинов, общих для этих двух выборок. В связи с этим, было принято решение рассматривать эти наборы данных отдельно.

Мы изучали эволюцию сайтов ацетилирования человека среди восьми видов позвоночных. Анализ проводился аналогично исследованию эволюции фосфосайтов. Для каждого сайта фосфорилирования мы реконструировали эволюцию и вычисляли вектора замен (Рисунок 5). Для оценки статистической значимости различий между векторами замен была применена процедура бутстрепа контрольных выборок.

Аналогично фосфосайтам, лишь часть ацетилированных лизинов из начальных выборок имели замены на другие аминокислоты. В отличие от сайтов фосфорилирования, мы не разделяли сайты ацетилирования по расположению в неструктурированных и регулярных областях белков.

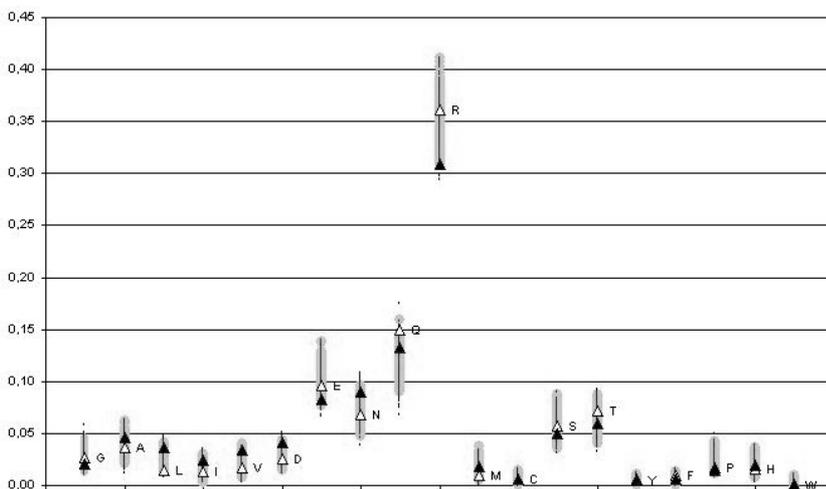


Рисунок 5. Вектора замен лизинов на другие типа аминокислот. Частоты замен лизинов среди позвоночных: для ацетилированных лизинов из клеточных линий MV4-11, A549 и Jurkat – белые треугольники; для ацетилированных лизинов из клеток печени – черные треугольники; 10000 контрольных выборок – облака больших светло-серых и маленьких темно-серых точек, соответственно.

Как видно из результатов анализа (Рисунок 5), вектора замен ацетилированных и неацетилированных остатков лизина также имели некоторые различия, но результаты, полученные для двух различных наборов данных, были несогласованными. По всей видимости, как и в случае с фосфорилированными остатками треонина и тирозина, доступных на сегодняшний день данных по ацетилированию белков недостаточно для получения значимых результатов.

Обсуждение результатов анализа эволюции посттрансляционных модификаций

Мы провели сравнительный эволюционный анализ паттернов замен сайтов посттрансляционных модификаций белков и их немодифицированных аналогов. В частности, была реконструирована и изучена эволюция сайтов фосфорилирования серина и треонина, а также сайтов ацетилирования лизина.

Нами было показано, что в процессе эволюции фосфосерины намного чаще заменяются на отрицательно заряженные остатки глутамата и аспартата (по сравнению с нефосфорилированными остатками серина). Эта тенденция наблюдалась во всех трех группах организмов. Во всех таксонах эта разница была статистически значимой ($P < 10^{-4}$), а в случае подвыборок надежных сайтов, подтвержденных в нескольких экспериментах, этот эффект был ещё более выраженным. Это наблюдение интересно тем, что искусственная замена серина на аспартат или глутамат, так называемая фосфомиметическая мутация, часто используется для функционального подтверждения фосфорилирования серина.

С другой стороны, методы, используемые в массовых фосфопротеомных экспериментах, часто основаны на предварительной аффинной селекции отрицательно заряженных пептидов. Это приводит к обогащению фосфопептидов кислотными остатками. Также было показано, что некоторые случаи фосфорилирования могут быть функционально консервативными, но при этом не сохранять конкретные позиции в быстро эволюционирующих неструктурированных областях белков. Так, в некоторых случаях функциональной целью фосфорилирования являются не конкретные сайты, а целые участки

белков. В таких случаях позиции сайтов фосфорилирования в ортологичных белках могут быть смещены относительно друг друга.

Все это, вместе со структурными особенностями белковых областей, в которых расположены фосфосайты, и проблемами, возникающими при выравниваниях таких областей, могло исказить результаты нашего анализа. Но все эти факторы должны оказывать такой же эффект и на выборку нефосфорилированных остатков серина, расположенных в тех же областях белков. Таким образом, эти факторы не могут объяснить наблюдаемую разницу между векторами замен фосфосеринов и серинов из дополнительных контрольных выборок, расположенных в непосредственной близости от фосфосайтов.

Необходимо учитывать, что анализируемые в данной работе вектора замен сайтов модификаций, вполне возможно, содержат ложные фосфосайты. Это может являться как экспериментальным артефактом массовых протеомных исследований, так и результатом нашей методики реконструкции эволюции фосфосайтов. Кроме того, существуют данные о том, что многие из идентифицированных сайтов фосфорилирования могут быть нефункциональными, и являться результатом неспецифичного фосфорилирования киназами. Паттерны замен таких нефункциональных фосфосайтов, скорее всего, будет схож с паттернами замен нефосфорилированных остатков серина. С другой стороны, контрольные выборки могут содержать еще не идентифицированные сайты фосфорилирования. Эти ложные, неспецифичные и неидентифицированные фосфосайты должны размывать наблюдаемую разницу между векторами замен модифицированных и немодифицированных остатков. Тем самым, реальная разница частот замен может быть ещё выше, чем было показано в настоящей работе.

В то же время, для других типов посттрансляционных модификаций значимых различий между паттернами замен модифицированных и немодифицированных аминокислотных остатков обнаружено не было. Это могло стать результатом малого количества доступных данных. Кроме того, дополнительное ограничение на объем данных о заменах сайтов модификаций на другие типы аминокислот, накладывало количество доступных ортологов из других организмов. В случае ацетилирования лизина результаты, полученные для двух различных наборов данных, оказались несогласованными. Это могло стать результатом низкой репрезентативности доступных на сегодняшний день данных по ацетилированию лизина. Поэтому результаты подобных исследований должны интерпретироваться крайне осторожно, и желательно основываться на данных, полученных из различных источников и для разнообразных групп организмов.

Основные результаты и выводы

1. Показано, что удержанные интроны значимо отличаются от конститутивно сплайсируемых по ряду параметров. В том числе, оказалось, что удержанные интроны в среднем короче, имеют более слабые сайты сплайсинга, содержат больше энхансеров и меньше сайленсеров сплайсинга, по сравнению с конститутивно сплайсируемыми интронами.
2. Показано, что удержанные интроны неравномерно распределены вдоль генов. В частности, доля удержанных интронов монотонно возрастает от 5`- к 3`-концу транскриптов.
3. Выявлен ряд значимых различий между экзонами, включение которых нарушается в результате мутаций в сайтах сплайсинга (S-экзоны), и экзонами, которые сохраняют включение, но сплайсируются по другому сайту (C-экзоны). Было показано, что S-экзоны в среднем короче, имеют более сильный акцепторный сайт, содержат меньше энхансеров и больше сайленсеров сплайсинга, по сравнению с C-экзонами. Кроме того, ближайшие потенциальные скрытые сайты сплайсинга расположены значительно ближе к аутентичным сайтам сплайсинга C-экзонов, по сравнению с S-экзонами.
4. Показано, что фосфосерины и фосфотреонины в среднем чаще располагаются в альтернативно сплайсируемых областях генов, по сравнению с нефосфорилированными остатками серина и треонина.
5. Была выдвинута и проверена гипотеза о том, что посттрансляционно модифицированные аминокислоты в процессе эволюции могут вести себя отлично от своих немодифицированных аналогов (с точки зрения частот замен на другие типы аминокислот). Выявлены значимые различия между векторами замен фосфорилированных и нефосфорилированных остатков серина. Оказалось, что фосфосерины чаще заменяются на глутамат и аспарат, по сравнению с нефосфорилированными остатками серина.

Список публикаций по теме диссертации

Статьи в научных журналах

1. Kurmangaliyev Y.Z., Gelfand M.S. Computational analysis of splicing errors and mutations in human transcripts // BMC Genomics. - 2008. - 9. - 13.
2. Курмангалиев Е.Ж., Гельфанд М.С. Сайты фосфорилирования тяготеют к альтернативно сплайсируемым областям белков // Молекулярная биология. - 2009. - 43. - 572-574.
3. Kurmangaliyev Y.Z., Goland A., Gelfand M.S. // Evolutionary patterns of phosphorylated serines // Biology Direct. - 2011. - 6. - 8.

Тезисы конференций

1. Курмангалиев Е.Ж. Использование удержанных интронов для изучения ошибок сплайсинга // Материалы XIV Международной конференции студентов, аспирантов и молодых ученых “Ломоносов”. - 2007. - 56.
2. Kurmangaliyev Y.Z. Exon skipping and activation of cryptic sites as consequences of splicing mutations // Proc. of 3rd Moscow Conference on Computational Molecular Biology (MCCMB'07). - 2007. - 168.
3. Курмангалиев Е.Ж. Последствия мутаций в сайтах сплайсинга: пропуск экзона и активация сайтов сплайсинга // Информационные технологии и системы (ИТиС'08): сборник трудов конференции. - 2007. - 255-256.
4. Kurmangaliyev Y.Z., Gelfand MS. Alternative splicing tends to involve phosphorylation sites // Информационные технологии и системы (ИТиС'08): сборник трудов конференции. - 2008. - 304-305.
5. Kurmangaliyev Y.Z. Patterns of evolution in protein phosphorylation sites // Proc. of 4th Moscow Conference on Computational Molecular Biology (MCCMB'09). - 2009. - 199-200.
6. Курмангалиев Е. Изучение эволюции сайтов ацетилирования лизина в белках позвоночных // Информационные технологии и системы (ИТиС'10): сборник трудов конференции. - 2010. - 415.

Автор выражает искреннюю благодарность своему научному руководителю Михаилу Сергеевичу Гельфанду, а также коллегам из УНЦ «Биоинформатика» ИППИ РАН. Автор также благодарит свою семью и друзей за терпение и поддержку при подготовке диссертации.