МОЛЕКУЛЯРНАЯ БИОЛОГИЯ, 2004, том 38, № 1, с. 82–91

КОМПЬЮТЕРНАЯ ГЕНОМИКА

УДК 577.2.08:681.3

ПРЕДСКАЗАНИЕ И КОМПЬЮТЕРНЫЙ АНАЛИЗ ЭКЗОН-ИНТРОННОЙ СТРУКТУРЫ ГЕНОВ ЧЕЛОВЕКА

© 2004 г. А. А. Миронов*, М. С. Гельфанд

Государственный научный центр "ГосНИИ генетика", Москва, 113545

Поступила в редакцию 03.09.2003 г.

Представлен обзор работ авторов по компьютерному анализу генома человека. Работа состоит из двух частей – одна посвящена разработке методов предсказания экзон-интронной структуры генов, а вторая – исследованию альтернативного сплайсинга. В первой части работы описаны идеи методов предсказания структуры генов с использованием информации о гомологии продукта гена с известным белком, или геномной последовательности с последовательностью гомологичного гена из другого организма. Тестирование предложенных методов показало их высокую эффективность. С использованием разработанных методов и баз данных EST был проведен анализ сплайсинга генов человека. Нами было впервые показано, что количество альтернативно сплайсируемых генов составляет не менее 35% от общего числа генов. Далее, проведено сравнение альтернативного сплайсинга в геномах человека и мыши. Было показано, что 50% альтернативно сплайсируемых генов (25% от общего числа генов) имеют специфические изоформы, представленные в одном организме и не представленные в другом.

Ключевые слова: экзон-интронная структура генов, альтернативный сплайсинг.

ПРЕДСКАЗАНИЕ КОДИРУЮЩИХ ОБЛАСТЕЙ В ГЕНОМАХ

Одной из важнейших целей секвенирования геномов является определение набора генов генома. Распознавание белоккодирующих областей в прокариотических геномах – задача давно известная и, в общем, решенная, хотя и там остаются проблемы (например, точное картирование стартов генов). Поиск же кодирующих областей в эукариотических генах – задача куда более сложная, поскольку в этих организмах кодирующие области разорваны "бессмысленными" последовательностями – интронами, длина которых часто в десятки раз превосходит "осмысленные" – экзоны. Поэтому предсказание кодирующих областей – это, прежде всего, задача предсказания экзон-интронной структуры. Основные методы предсказания экзон-интронной структуры представлены в работах [1–3]. Следуя традиции, в дальнейшем для краткости изложения будем называть экзонами только кодирующую часть геномной последовательности.

Обычно система предсказания экзон-интронной структуры базируется на следующих соображениях. Во-первых, в подавляющем большинстве случаев интроны начинаются и заканчиваются консервативными динуклеотидами: GT – в начале интрона и AG – в его конце. Во-вторых, в областях, прилегающих к экзон-интронным границам, есть определенное предпочтение нуклеотидов в других позициях, что позволяет строить разного рода профили (весовые матрицы) для распознавания границ. Однако построенные таким образом распознающие правила все еще достаточно слабы. В-третьих, если нас интересует только кодирующая часть гена, то для более четкого выбора сайтов сплайсинга можно использовать соображение, что на том, что получится после вырезания интронов, не должно быть стоп-кодонов. Но и это не позволяет надежно определять структуру генов. Можно применить статистику кодонов и тем самым значительно улучшить качество предсказания (до 70%). Использование статистик дикодонов, статистических свойств интронов и слабых корреляций нуклеотидов в сайтах сплайсинга позволяет еще увеличить качество предсказания экзон-интронных структур генов (до 85%). Впервые комплексный статистический подход, учитывающий локальные (сайты сплайсинга) и глобальные (статистика белоккодирующих областей) свойства последовательностей, предложен в работе [4]. В настоящее время наиболее популярная техника решения задачи поиска белоккодирующих областей основана на использовании скрытых Марковских цепей [5], учитывающих все перечисленные соображения.

Тем не менее, применение статистических методов предсказания структуры генов при анализе больших геномных фрагментов имеет несколько трудно преодолимых недостатков. Во-первых, нет возможности локализовать границы генов и

* Эл. почта: aa_mironov@pochtamt.ru

предсказания часто объединяют несколько генов, закодированных в последовательности, в один очень большой ген, или, напротив, расщепляют гены на два и более. Во-вторых, применение этих методов затруднительно при наличии ошибок в геномной последовательности. Особенно они чувствительны к сдвигу рамки. Наконец, в-третьих, эти методы позволяют предсказывать только одну изоформу, игнорируя альтернативный сплайсинг.

С 1994 г. используют идеи о применении гомологий для определения структуры генов [6–8]. Такой подход не всегда применим, поскольку не всегда можно найти достаточно надежного гомолога, но в тех случаях, когда это можно сделать (~70% случаев), надежность предсказания можно существенно увеличить, часто до 100%. Пусть нам дана геномная последовательность, для простоты, содержащая один ген. Применив программу BLAST [9], мы можем найти в банке данных (например, в GenBank) белки, которые гомологичны белку, закодированному в данной последовательности. Используя наивный подход, мы можем просто разложить полученный список выравниваний на геномную последовательность и объявить их экзонами. Однако практика показывает, что такой подход дает качество предсказания, сравнимое со статистическим предсказанием структуры гена *ab initio* – примерно 85% . Проблема в том, что выравнивания, предъявленные программой BLAST, зачастую не согласованы с сайтами сплайсинга. Кроме того, этот подход заведомо теряет сравнительно короткие экзоны. Поэтому нами был предложен новый подход – сплайсированное выравнивание. Можно поставить несколько разных, с математической точки зрения, задач, связанных с предсказанием структуры генов на основе сходства.

Сайтовая задача. Пусть нам известны потенциальные сайты сплайсинга и последовательность гомологичного белка или кДНК. Надо найти такую экзон-интронную структуру, которая обеспечит наилучшее выравнивание сплайсированной последовательности с целевым белком или кДНК. Такой подход требует достаточно высокой гомологии гена и целевого белка и не позволяет использовать такое мощное средство, как статистика кодонов. С другой стороны, этот подход может быть применен в случае, когда последовательность гена прочитана не достаточно чисто и содержит в себе ошибки. Эта задача решается с помощью динамического программирования наподобие метода Смита-Ватермана [10], но с добавленными делециями специального вида (интронами) [11].

Блочная задача. Эта задача возникает тогда, когда нам известны не только сайты сплайсинга, но также и потенциальные экзоны, которых может быть очень много. Потенциальные экзоны могут быть предсказаны с помощью, например, статистики кодонов и сайтов сплайсинга. Этот метод позволяет использовать гораздо более удаленные гомологи.

Геномная задача. Существует также третья задача, когда нам известны две гомологичные (в смысле кодируемого белка) геномные последовательности из разных организмов. В этом случае можно определить структуру гена, опираясь на то, что экзоны кодируют белок и поэтому находятся под более строгим давлением отбора, в то время как интроны гораздо менее консервативны.

АЛГОРИТМЫ ПРЕДСКАЗАНИЯ ЭКЗОН-ИНТРОННОЙ СТРУКТУРЫ ГЕНОВ

Исторически сначала мы разработали алгоритм для решения более сложной блочной задачи. В настоящее время этот подход известен под названием алгоритм Прокруст (Procrustes, [1]). Пусть нам известны геномная последовательность, последовательность гомологичного белка и набор потенциальных экзонов. При этом потенциальных экзонов может быть достаточно много, и они могут перекрываться друг с другом. Задача заключается в том, чтобы построить цепочку экзонов, которая наилучшим способом выравнивается с данным белком. Наивный алгоритм поиска такой цепочки предполагает независимое выравнивание каждого из потенциальных экзонов с белком и затем построение оптимальной цепочки. Нетрудно оценить ожидаемое количество операций для этой задачи. Оно будет примерно равно $O(\Sigma l_{exn} * l_{prot})$, где Σl_{exn} – суммарная длина всех потенциальных экзонов, l_{prot} – длина аминокислотной последовательности. Однако потенциальные экзоны могут перекрываться. В этом случае представляется лишним много раз производить выравнивание одного и того же участка геномной последовательности (представленной в разных экзонах) с аминокислотной последовательностью. Поэтому предложена усовершенствованная схема, при которой каждый участок геномной последовательности или просматривается один раз, если он покрыт хоть одним потенциальным экзоном, или ни разу, если он не покрыт ни одним потенциальным экзоном. В этом случае время работы алгоритма оценивается как $O(\kappa l_{gene} * l_{prot})$, где κ – покрытие геномной последовательности потенциальными экзонами, l_{gene} – длина геномной последовательности. Надо отметить, что алгоритм Прокруст был не только первым алгоритмом предсказания экзон-интронной структуры генов на основе сходства, но также и первым подходом, позволившим комбинировать гомологию и статистическое предсказание структуры генов. Исследование свойств алгоритма Прокруст [12] показало, что качество предсказания зависит от

качества полученного выравнивания и при сходстве, превышающем 35%, предсказания можно считать достаточно надежными. В частности, в ряде случаев для предсказания структуры генов человека можно использовать даже последовательности прокариотических белков.

При определении нуклеотидных последовательностей неизбежно возникают ошибки, в том числе ошибки типа сбоя рамки и ошибки в сайтах сплайсинга. Уровень ошибок особенно высок на первых этапах секвенирования. Например, если уровень ошибок равен 1% (что характерно для однократного прочтения), то 2% сайтов сплайсинга будут содержать критические ошибки и 4% экзонов будет определено неправильно. Кроме того, в экзонах с заметной частотой будут появляться паразитные стоп-кодоны, не говоря уже о возможных сдвигах рамки. В таких условиях никакой алгоритм, основанный на статистике кодонов и сайтов сплайсинга, не сможет дать разумное предсказание. Нами был разработан метод Proframe предсказания структуры генов, который даже в таких тяжелых случаях может давать хорошие предсказания [13]. Это метод основан на той же идее использования сходства с гомологичным белком. В этом случае используется сайтовая задача, а алгоритм основан на динамическом программировании типа Смита-Ватермана, в который добавлены переходы, связанные с интронами и сдвигами рамки.

Описанные выше методы показали высокую эффективность, однако они не дают возможности находить гены, для продуктов которых нет известных гомологов. То, что белоккодирующие последовательности эволюционируют с гораздо меньшей скоростью, чем некодирующие последовательности (межгенные участки и интроны), позволяет использовать сравнение геномных последовательностей для предсказания структуры генов. При таком сравнении нуклеотидных последовательностей двух родственных генов белоккодирующие экзоны выглядят как участки с высокой локальной гомологией на фоне сильно дивергировавших или вообще не гомологичных некодирующих областей. Использованию геномных сравнений для предсказания структуры генов способствует массовое секвенирование геномов эукариот и, в особенности, синтенных районов. В последнее время этот подход был с успехом применен для анализа кластера генов на хромосоме 12p13 человека и в синтенной области на хромосоме 6 мыши [14] и для предсказания генов в локусе *bli-4* нематод *Caenorhabditis elegans* и *C. briggsae* [15]. В этих работах использовали сходство нуклеотидных последовательностей, анализируемое с помощью построения точечных матриц сходства [15] или поиска локальных гомологий [14]. В то же время известно, что анализ белковых последовательностей гораздо более чувствителен к наличию отдаленного сходства, чем анализ нуклеотидных последовательностей. Кроме того, такой анализ позволяет отличить гомологию белоккодирующих экзонов от сходства некодирующих, например, регуляторных областей [16, 17].

Нами разработан вариант сплайсированного выравнивания (Pro-Gen [18]), позволяющий производить попарное сравнение геномных последовательностей на уровне белков и предсказывать экзон-интронную структуру генов высших эукариот. При этом в отличие от предыдущих работ консервативность экзон-интронной структуры не предполагалась. Этот вариант алгоритма сплайсированного выравнивания использует сайтовый подход, т.е. строится не предварительный список потенциальных экзонов, а полное выравнивание с учетом возможных интронов. Алгоритм строит стандартную матрицу выравнивания. Здесь возникает трудность, связанная с тем, что размер матрицы выравнивания равен произведению длин исходных геномных последовательностей, что при среднем размере гена 30 т.п.н. приводит к необходимости использовать около 100 Мбайт оперативной памяти, а для более длинных генов и того больше. Поэтому в программной реализации запоминались только переходы на сайтах сплайсинга.

Следует отметить, что методы сплайсированного выравнивания работают плохо: если не использовать ограничений снизу на размеры экзонов и интронов – алгоритм предсказывает большое количество очень коротких экзонов. Это неудивительно, особенно при исследовании геномов высших эукариот, имеющих гигантские интроны. Действительно, по чисто случайным причинам можно найти короткий фрагмент (1–5 аминокислотных остатков), гомологичный целевому белку. Для борьбы с этим “мозаичным” эффектом приходится вводить ограничения на размеры интронов и экзонов, а также вводить специальные штрафы на инициацию интрона.

Все методы сплайсированного выравнивания просматривают достаточно большую матрицу, и время работы алгоритма пропорционально произведению длин сравниваемых последовательностей. С другой стороны, для поиска кандидатов для сравнения используются быстрые алгоритмы типа BLAST. Представляется целесообразным использовать выравнивания, полученные с помощью программы BLAST при построении сплайсированного выравнивания. Действительно, каждое локальное выравнивание разбивает всю матрицу выравнивания на квадранты и делает ненужным просмотр около половины полной матрицы. Последние версии программ сплайсированного выравнивания используют это соображение и рабо-

тают достаточно быстро – типичное сравнение требует времени порядка 1с.

Алгоритм геномного выравнивания оказался полезным также и для анализа прокариотических геномов. Задача поиска генов в бактериях не представляет в настоящее время больших трудностей, за исключением стартов генов, которые предсказываются не достаточно надежно. Кроме того, возможные сдвиги рамок также снижают точность предсказания генов. Нами предложен метод исправления аннотаций генов, основанный на сравнительном анализе предсказаний в родственных геномах. Используется алгоритм выравнивания транслированных геномных последовательностей с учетом сбоя рамки. При этом выравнивание может начинаться только на возможных стартовых кодонах. Поскольку давление отбора на некодирующие области меньше и имеет другой характер, чем в кодирующих областях, такой подход позволяет с большой точностью находить старты генов и сбои рамок. В частности, нам удалось уточнить положения около 200 генов даже в таком хорошо охарактеризованном геноме, как *E. coli*. Сопоставление наших исправлений с экспериментально картированными стартами генов показало, что в 95% случаев наши предсказания точно совпадают с экспериментальными стартами [19].

ИССЛЕДОВАНИЕ АЛЬТЕРНАТИВНОГО СПЛАЙСИНГА

Секвенирование геномов часто сопровождается секвенированием большого количества фрагментов транскрибированных и сплайсированных РНК (EST), а также полноразмерных РНК. Для картирования генома человека было накоплено более 4 млн. последовательностей EST (expression sequence tags). EST – это короткие последовательности, прочитанные с тотальной мРНК клетки с одного прохода, как правило, со случайных или с поли-Т праймеров. Поскольку это прочтение с одного прохода, то качество этих последовательностей низкое. С помощью сравнения последовательности EST можно кластеризовать и собирать в контиги, которые в результате могут дать более или менее полные последовательности матричных РНК, экспрессируемых в данной клетке [20, 21].

Большое количество секвенированных последовательностей EST и EST-контигов позволяет использовать эти данные для предсказания экзон-интронной структуры. Поскольку база данных EST содержит последовательности, полученные из разных тканей и на разной стадии развития, то эти данные являются подходящим инструментом для исследования альтернативного сплайсинга. При исследовании альтернативного сплайсинга используют две стратегии. Одна – кластеризация EST, их множественное выравнивание и анализ нарушений правильной структуры выравниваний [22]. Другой подход основан на картировании EST на геномные последовательности и определение экзон-интронной структуры генов и всех их вариантов. Для картирования EST на геномные последовательности можно также использовать алгоритмы сплайсированного выравнивания. Нами разработана эффективная программа картирования последовательностей EST на геномные последовательности (Pro-EST), которая с успехом использовалась в наших исследованиях.

Задача использования EST для геномной аннотации и предсказания экзон-интронной структуры не столь тривиальна, как может показаться на первый взгляд. Это исследовалось несколькими группами, в частности GRAIL [23]. Одной из основных трудностей было то, что значительное число EST картируется в межгенные и интронные области, или является продуктом неправильного или неполного сплайсинга. По-видимому, до 20% EST, представленных в базах данных, – артефакты [24]. Поэтому наиболее информативны EST или EST-контиги, которые покрывают более одного экзона. Использование программ поиска локальных гомологий типа BLAST не достаточно, так как они не позволяют точно локализовать экзон-интронные границы.

В работе использована разработанная нами программа предсказания экзон-интронной структуры по данным EST (Pro-EST), основанная на алгоритме сплайсированного выравнивания [1], и база данных EST-контигов TIGR [25]. Составлена выборка из 400 аннотированных генов человека. Повторы отфильтрованы программой RepeatMasker [26]. EST-контиги, соответствующие отобранным генам, выбраны из TIGR Human Gene Index с использованием программы BLASTN.

Процедура поиска альтернативного сплайсинга

Экзон-интронную структуру определяли с помощью программы Pro-EST. Эта программа искала сайты сплайсинга с очень низким порогом, затем искала цепочку экзонов, обеспечивающую наилучшее выравнивание с последовательностью сравнения (EST-контигом). Для этого использовали алгоритм сплайсированного выравнивания. Критерием качества выравнивания считали нормированный вес выравнивания – отношение веса выравнивания к весу тривиального выравнивания последовательности сравнения с собой [12]. Таким образом, для каждого гена был получен ряд предсказаний, индуцированных разными последовательностями сравнения. Далее применяли постобработку результатов предсказания – предсказанные структуры объединяли в *карту*, или *суперструктуру*, если предсказания пересекались без противоречий. Это объединение прово-

Таблица 1. Число EST-контигов, соответствующих одной суперструктуре

Число контигов	0	1	2	3	4	5	6	7	8
% суперструктур	10.6	48.3	19.3	10.8	6.6	1.8	1.4	0.8	0.4

Примечание. Столбец 0 представляет число химерных структур.

Таблица 2. Число суперструктур, в которых представлен один EST-контиг

Число суперструктур	0	1	2	3	4	5	6	>6
% контигов	27.1	55.5	12.2	2.0	1.6	0.2	0.7	0.7

Примечание. Столбец 0 показывает количество одиночных не подтвержденных контигов.

Таблица 3. Распределение числа альтернативно сплайсирующихся генов

Число суперструктур	1	2	3	4	5	6	7	8	>9
% генов	65.6	18.6	4.6	5.4	1.3	1.5	0.5	1.0	1.5

Примечание. Столбец 1 соответствует генам, альтернативный сплайсинг для которых не обнаружен.

дили по следующему алгоритму. Рассматривали все элементы структуры вида интрон-экзон-интрон (*е-триплет*). Два таких элемента объединяли, если правый интрон первого *е*-триплета совпадал с левым интроном второго *е*-триплета. Таким образом, при объединении допускалось даже слабое перекрытие экзонов, если оно подтверждалось положением интрона. С другой стороны, даже длинный участок совпадения экзонов не служил основанием для объединения, если структуры имели различные предсказанные интроны. Такое, в частности, могло быть получено от включения в выборку не полностью сплайсированной РНК.

Процедуру продолжали до тех пор, пока к структуре нельзя было добавить ни одного нового *е*-триплета. Так были сконструированы все возможные структуры. Заметим, что эта процедура предполагает отсутствие корреляции между сплайсингом разных интронов, поэтому возможно возникновение химерных структур, противоречащих всем использованным EST-контигам. Однако использование сравнения с короткими EST не является подходящим методом для анализа такого рода корреляций. Для этого необходим анализ последовательностей полноразмерных кДНК.

Контиги или суперструктуры, не имеющие общих экзонов или интронов с другими суперструктурами или аннотированными CDS, исключали из дальнейшего рассмотрения. Есть два типа суперструктур такого рода. Во-первых, они могут оказаться целиком за пределами кодирующей области или быть порождены соседними не аннотированными генами. Во-вторых, такие суперструктуры содержат только один экзон и поэтому ничего не говорят о структуре гена, в частности, они могли появиться от несплайсированной РНК.

EST-контиги и суперструктуры

Использование геномной последовательности в качестве якоря является дополнительным инструментом для кластеризации EST. Таблица 1 показывает распределение числа EST-контигов, которые можно кластеризовать с помощью геномной последовательности. Такая дополнительная кластеризация возможна почти в 50% случаев. Химерные суперструктуры, не поддержанные отдельными контигами, составляют 10% всех суперструктур. Оставшиеся 40% карт сформированы более чем одним контигом.

Представительность индивидуальных EST-контигов в суперструктурах описана в табл. 2. Более половины контигов (55%) представлены в одной суперструктуре и немного более четверти (27%) – одиночные контиги. Примерно 3% контигов участвуют в действительно сложных событиях альтернативного сплайсинга.

В табл. 3 представлено число предсказанных альтернативных экзон-интронных структур на один ген. Более трети генов имеют, как минимум, два варианта экзон-интронной структуры. Альтернативные структуры были классифицированы в соответствии с позициями на мРНК. Так, можно различать альтернативы на 5'-конце (5'-вилка), альтернативы на 3'-конце (3'-вилка) и внутренние альтернативы (петли, включая выпячивания). 5'-вилки обнаружены в 73 генах (54%

альтернативно сплайсирующихся генов), петли – в 41 гене (30%), 3'-вилки – в 64 генах (47%).

Проанализировано распределение частных вариантов альтернативного сплайсинга. Установлено, что 23% петель образованы альтернативным акцепторным сайтом, 16% имеют альтернативный донорный сайт, в 27% генов экзон является кассетным, т.е. присутствует в одном из вариантов и отсутствует в другом. Было всего несколько случаев, когда в структуре исчезал интрон, существующий в альтернативе. 25% случаев не поддаются простой классификации и содержат комбинацию из описанных выше структур. Далее, в 22% 5'-вилок присутствует альтернативный 5'-концевой экзон, 18% имеют разные старты транскрипции и дополнительный интрон. Наконец, 11% 3'-вилок имеют альтернативный терминальный экзон, 35% имеют разные точки окончания (сайты полиаденилирования) и дополнительный интрон в одном варианте. Остальные случаи классифицированы как сложные.

Классифицируя альтернативный сплайсинг с функциональной точки зрения, мы обнаружили, что 80% альтернативно сплайсирующихся генов имеют альтернативы в 5'-нетранслируемой области, 20% имеют альтернативы в кодирующей области и 19% – в 3'-нетранслируемой области (сумма превышает 100%, поскольку многие гены имеют по несколько типов альтернатив).

Истинные альтернативы или ошибки сплайсинга?

Чтобы отличить альтернативный сплайсинг от аберрантного, проанализированы случаи, когда нарушалась рамка считывания. В 161 случае альтернативные районы целиком содержались в аннотированной кодирующей области. В 95 случаях (59%) альтернативы отличались на целое количество кодонов, 23 случая включали несколько (обычно два) экзонов, которые восстанавливали рамку считывания. Сдвигающие рамки альтернативы, связанные с потерей/вставкой экзона, наблюдали в 40 случаях. Альтернативы, связанные с выбором сайта сплайсинга, обнаружены в 74 случаях, и только в 4 случаях сохранялся интрон.

Таким образом, возможность сохранения интрона можно исключить в подавляющем большинстве случаев. Для более надежного отличия аберрантного сплайсинга от реального альтернативного сплайсинга требуется детальный анализ каждого случая и большой объем экспериментальной работы. Предварительное наблюдение заключается в том, что некомпенсированный сбой рамки преимущественно наблюдается в окрестности 3'-конца кодирующей области и поэтому затрагивает С-конец белка. Кроме того, более трети сбоев рамки может быть объяснено, если допустить сплайсинг по неканоническим сайтам.

На момент выполнения этой работы существовала оценка частоты альтернативного сплайсинга в геноме человека около 6%. В настоящее время эта оценка превышает нашу и составляет около 50% [27, 28]. Анализ последовательностей EST позволил создать ряд баз данных по альтернативному сплайсингу [29, 30].

Влияние альтернативного сплайсинга на структуру белков

В начале проекта "Геном человека" количество генов человека оценивали как 80–120 тыс. генов. Однако определение и анализ структуры генома человека показало, что он содержит около 35 тыс. генов [27, 28]. Разнообразие же белков в значительной степени определяется альтернативным сплайсингом. Возникает естественный вопрос: как отличаются доменные структуры различных изоформ?

С этой целью сделана выборка белков из полностью секвенированных высших эукариот из Swiss-Prot, для которых охарактеризован альтернативный сплайсинг (1780 белков, 4804 изоформ) [31]. Затем, с использованием InterProScan [32, 33] для этих белков определена доменная организация. Далее локализованы альтернативные участки аминокислотной последовательности относительно доменной структуры белков. Анализ полученных позиций показал, что границы альтернативных участков, как правило, находятся вне доменов и других структурных или функциональных элементов. Оценка статистической значимости такого предпочтения составляет величину от 10^{-5} до 10^{-48}. Это означает, что альтернативный сплайсинг существенно чаще включает или выключает целые домены или другие структурно-функциональные элементы, либо меняет петлевые участки, чем можно было ожидать при случайном распределении альтернативно сплайсированных участков и белковых доменов. Однако в 28% случаев участок альтернативного сплайсинга находится внутри доменов. Для 48 белков из нашей выборки известны пространственные структуры. В 43 из 71 альтернативных вариантов (60.5%) происходит удаление (добавление) существенной части гидрофобного ядра белка, в результате чего белок теряет (приобретает) функциональность одного из доменов. Остальные 28 вариантов не влияют на структуру доменов и локализованы вне гидрофобного ядра и в 10 случаях заменяют короткий фрагмент белка, а в 18 – удаляют (добавляют) короткие фрагменты.

Для более детального анализа выделены случаи, когда альтернативные варианты перекрывают не более 50 остатков домена. Для этих случа-

Таблица 4. Выборка пар альтернативно сплайсируемых ортологичных генов мыши и человека

Выборка	Человек	Мышь
Альтернативно сплайсируемые гены	126	124
Кассетные экзоны	177	123
Альтернативный акцепторный сайт	51	46
Альтернативный донорный сайт	52	53
Сохраненный интрон	12	29
Всего альтернатив	285	252

ев, с использованием Swiss-Prot [36] и ProSite [35], показано, что альтернативный сплайсинг чаще (статистическая значимость 10^{-14}) меняет функционально-значимые остатки, чем при случайном расположении. Таким образом, можно предположить, что альтернативный сплайсинг позволяет варьировать функцию белков, не изменяя скелет его пространственной структуры.

Исследование консервативности альтернативного сплайсинга

В настоящее время уже очевидно, что альтернативный сплайсинг является существенным источником разнообразия белков в эукариотической клетке. Секвенирование генома мыши [36] показало, что только один процент генов мыши не имеет гомологов в геноме человека и наоборот один процент генов человека не имеет гомологов в геноме мыши. Средняя степень сходства белков мыши и человека составляет около 80%. Около 40% некодирующих областей могут быть выровнены на нуклеотидном уровне. Основная структура ортологичных генов также совпадает. Возникает естественный вопрос: насколько различается альтернативный сплайсинг в этих геномах? Наивный подход к решению этого вопроса заключается в следующем – найдем все альтернативы ортологичных генов мыши и человека и сравним их. Как отмечалось выше, основным источником информации об альтернативном сплайсинге служат последовательности EST. Поэтому наблюдаемые различия в изоформах могут быть объяснены просто отсутствием соответствующих последовательностей EST в базах данных, а не природным различием в альтернативном сплайсинге.

Наше исследование основано на другом подходе [37, 38]. Рассматриваем все доступные альтернативы для гена одного из организмов. Затем для каждой из изоформ находим аминокислотную последовательность и по этой последовательности с помощью программы ProFrame [13] находим соответствующую изоформу в другом организме. Если некоторые сегменты аминокислотной последовательности не выравниваются с геномной последовательностью другого организма, то можно с уверенностью заявлять, что соответствующий экзон, или его часть, не представлен в другом организме, и, следовательно, мы наблюдаем неконсервативный альтернативный сплайсинг.

Исследование консервативности сплайсинга потребовало подготовки выборки пар ортологичных альтернативно сплайсируемых генов человека и мыши. Использована база данных по альтернативному сплайсингу млекопитающих AsMam DB [29]. Для каждого гена мыши, представленного в этой базе данных, находился соответствующий ген человека с помощью программы BLAST. Аналогично, в базе данных по альтернативному сплайсингу человека HAS DB [30] отобраны альтернативно сплайсируемые гены человека, и для них найдены соответствующие гены мыши. В результате получена выборка из 161 пары генов. Структура выборки представлена в табл. 4. Отличие количества альтернативно сплайсируемых генов в таблице от размера выборки объясняется тем, что в некоторых парах в одном организме известен альтернативный сплайсинг, а в другом –

Таблица 5. Консервативность альтернативного сплайсинга

Выборка	Человек		Мышь	
	консервативны	не консервативны	консервативны	не консервативны
Кассетные экзоны	74	26	39	9
Альтернативный акцепторный сайт	16	10	17	6
Альтернативный донорный сайт	19	15	16	9
Сохраненный интрон	5	0	10	4
Всего альтернатив	114	51	82	28
	69%	31%	75%	25%
Всего генов	41	44	30	26
	48%	52%	54%	46%

Таблица 6. Консервативность альтернативного сплайсинга: результаты анализа баз данных по альтернативному сплайсингу и полноразмерных кДНК

Выборка	Человек				Мышь			
	консервативны		не консервативны		консервативны		не консервативны	
	всего	кДНК	всего	кДНК	всего	кДНК	всего	кДНК
Кассетные экзоны	74	56	26	25	70	39	9	5
Альтернативный акцепторный сайт	16	18	10	7	24	17	6	6
Альтернативный донорный сайт	19	13	15	5	15	16	9	6
Сохраненный интрон	5	4	0	3	8	10	4	7
Всего альтернатив	114	96	51	30	117	82	28	24
	69%	76%	31%	24%	83%	75%	25%	17%
Всего генов	41	45	44	28	68	30	26	22
	48%	62%	52%	38%	76%	54%	46%	24%

нет. В результате применения описанной процедуры обнаружены неконсервативные изоформы, представленные в табл. 5.

Данные таблицы 5 показывают, что почти половина из рассмотренных генов имеет изоформы, специфические для человека и для мыши. Полученный результат можно объяснить загрязненностью базы данных EST. Поэтому проведен контрольный эксперимент, в котором использованы только те альтернативы, которые подтверждены полноразмерным секвенированием кДНК. Результаты приведены в табл. 6. Видно, что количество неконсервативных изоформ и генов, имеющих неконсервативные изоформы, несколько сократилось, но качественно результат остался прежним. При анализе этой таблицы надо иметь в виду, что небольшой размер выборки не может дать хорошую количественную оценку, а только качественную. Тем не менее, представляется удивительным, что при всей биологической схожести человека и мыши такое большое количество генов имеет специфические изоформы, и, следовательно, специфические белки. Для целого ряда генов проведен специальный анализ литературы и найдены подтверждения нашего наблюдения, а именно: в этих работах описаны и специально исследованы изоформы, которые, по нашим данным, являются специфическими для одного или другого организма. Кроме того, надо отметить, что наши оценки по построению достаточно жестки: вполне возможно, что ряд изоформ, которые, по нашим наблюдениям, следует отнести к консервативным, на самом деле таковыми не являются.

Сделанные нами наблюдения подтверждают, что исследование альтернативного сплайсинга – одна из важнейших задач анализа генома человека. Проведенные исследования и полученные результаты ставят целый ряд новых проблем. Нам представляется важным исследовать тканевую специфичность изоформ, в том числе и неконсервативных. Представленность транскриптов в базах данных EST связывают с уровнем транскрипции [39]. Было бы интересно оценить таким образом уровень транскрипции разных изоформ. Еще одна задача связана с анализом однонуклеотидого полиморфизма. Структуру генов можно представить как мозаику константных экзонов (фрагментов, которые присутствуют во всех изоформах), альтернативных экзонов (фрагментов, которые в некоторых изоформах отсутствуют), неконсервативных экзонов (фрагментов, которые отсутствуют в зрелых мРНК мыши). Задача заключается в оценке частоты и качества полиморфизмов в этих районах. Кроме того, можно оценить степень межвидовой консервативности в этих районах. Очень важны экспериментальные исследования альтернативного сплайсинга. Не секрет, что массовые данные по EST содержат значительный уровень шума. Детальные экспериментальные исследования разных изоформ мРНК и белков для конкретных генов могли бы пролить дополнительный свет на роль альтернативного сплайсинга в функционировании генома.

В заключение хотелось бы заметить, что в начале работ над программой “Геном человека” компьютерные методы анализа геномов играли в основном вспомогательную роль. В настоящее же время значительная часть биологически важных результатов получается благодаря именно компьютерному анализу геномов. Эта эволюция хорошо видна и в описанных работах, которые начались с разработки методов предсказания экзон-интронной структуры генов, а продолжились анализом альтернативного сплайсинга.

Эти исследования поддержаны грантами Российского фонда фундаментальных исследований, Российской национальной программы “Геном че-

ловека", Международного фонда Сороса, ИНТАС, Медицинского Института Ховарда Хьюза и Людвиговского института раковых исследований.

Надо отметить, что Александр Александрович Баев с самого начала Национального проекта "Геном человека" уделял большое внимание развитию методов компьютерного анализа геномов.

СПИСОК ЛИТЕРАТУРЫ

1. Gelfand M.S., Mironov A.A., Pevzner P.A. 1996. Gene recognition via spliced sequence alignment. *Proc. Natl. Acad. Sci. USA*. **93**, 9061–9066.
2. Fickett J.W. 1996. The gene identification problem: An overview for developers. *Comput. Chem*. **20**, 103–118.
3. Fickett J.W. 1996.Finding genes by computer: the state of the art. *Trends Genet*. **12**, 316–320.
4. Gelfand M.S. 1990. Computer prediction of the exon-intron structure of mammalian pre-mRNAs. *Nucleic Acids Res*. **18**, 5865–5869.
5. Burge C., Karlin S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol*. **268**, 78–94.
6. Snyder E.E., Stormo G.D. 1995. Identification of protein coding regions in genomic DNA. *J. Mol. Biol*. **248**, 1–18.
7. Gelfand M.S., Mironov A.A., Pevzner P.A. 1994. *Gene recognition via spliced sequence alignment. Technical report*. PennState Univ.
8. Rogozin I.B., Milanesi L., Kolchanov N.A. 1996. Gene structure prediction using information on homologous protein sequence. *Comput. Appl. Biosci*. **12**, 161–170.
9. Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. 1990. Basic local alignment search tool. *J. Mol. Biol*. **215**, 403–410.
10. Waterman M.S. 1984. Efficient sequence alignment algorithms. *J. Theor. Biol*. **108**, 333–337.
11. Birney E., Thompson J.D., Gibson T.J. 1996. PairWise and SearchWise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. *Nucleic Acids Res*. **24**, 2730–2739.
12. Mironov A.A., Roytberg M.A., Pevzner P.A., Gelfand M.S. 1998. Performance-guarantee gene predictions via spliced alignment. *Genomics*. **51**, 332–339.
13. Mironov A.A., Novichkov P.S., Gelfand M.S. 2001. ProFrame: similarity-based gene recognition in eukaryotic DNA sequences with errors. *Bioinformatics*. **17**, 13–15.
14. Ansari-Lari M.A., Oeltjen J.C., Schwartz S., Zhang Z., Muzny D.M., Lu J., Gorrell J.H., Chinault A.C., Belmont J.W., Miller W., Gibbs R.A. 1998. Comparative sequence analysis of a gene-rich cluster at human chromosome 12p13 and its syntenic region in mouse chromosome 6. *Genome Res*. **8**, 29–40.
15. Thacker C., Marra M.A., Jones A., Baillie D.L., Rose AM. 1999. Functional genomics in Caenorhabditis elegans: An approach involving comparisons of sequences from related nematodes. *Genome Res*. **9**, 348–359.
16. Lipman D.J. 1997. Making (anti)sense of non-coding sequence conservation. *Nucleic Acids Res*. **25**, 3580–3583.
17. Duret L., Bucher P. 1997. Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol*. **7**, 399–406.
18. Novichkov P.S., Gelfand M.S., Mironov A.A. 2001. Gene recognition in eukaryotic DNA by comparison of genomic sequences. *Bioinformatics*. **17**, 1011–1018.
19. Baytaluk M.V., Gelfand M.S., Mironov A.A. 2002. Exact mapping of prokaryotic gene starts. *Brief Bioinform*. **3**, 181–194.
20. Mironov A.A., Gelfand M.S. 1998. Gene recognition using EST data: Unexpectedly frequent alternative splicing of human genes. *Proc. of the first international conference on bioinformatics*. **2**, 249–250.
21. Mironov A.A., Fickett J.W., Gelfand M.S. 1999. Frequent alternative splicing of human genes. *Genome Res*. **9**, 1288–1293.
22. Hanke J., Brett D., Zastrow I., Aydin A., Delbruck S., Lehmann G., Luft F., Reich J., Bork P. 1999. Alternative splicing of human genes: more the rule than the exception? *Trends Genet*. **15**(10), 389–390.
23. Xu Y., Uberbacher E.C. 1997. Automated gene identification in large-scale genomic sequences. *J. Comput. Biol*. **4**, 325–338.
24. Wolfsberg T.G., Landsman D. 1997. A comparison of expressed sequence tags (ESTs) to human genomic sequences. *Nucleic Acids Res*. **25**, 1626–1632.
25. Adams M.D., Kerlavage A.R., Fleischmann R.D., Fuldner R.A., Bult C.J., Lee N.H., Kirkness E.F., Weinstock K.G., Gocayne J.D., White O., et al. 1995. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature*. **377**(6547 Suppl), 3–174.
26. http://www.genome.washington.edu/UWGC/analysistools/repeatmask.htm.
27. International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature*. **409**(6822). 860–921.
28. Venter J.C., et al. 2001. The sequence of the human genome. *Science*. **291**(5507), 1304–1351.
29. Ji H., Zhou Q., Wen F., Xia H., Lu X., Li Y. 2001. AsMamDB: an alternative splice database of mammals. *Nucleic Acids Res*. **29**, 260–263.
30. Modrek B., Resch A., Grasso C., Lee C. 2001. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res*. **29**, 2850–2859.
31. Kriventseva E.V., Koch I., Apweiler R., Vingron M., Bork P., Gelfand M.S., Sunyaev S. 2003. Increase of functional diversity by alternative splicing. *Trends Genet*. **19**(3), 124–128. Review.
32. Zdobnov E.M., Apweiler R. 2001. InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*. **17**, 847–848.
33. Pruess M., Fleischmann W., Kanapin A., Karavidopoulou Y., Kersey P., Kriventseva E., Mittard V., Mulder N., Phan I., Servant F., Apweiler R. 2003. The Proteome Analysis database: a tool for the in silico analysis of whole proteomes. *Nucleic Acids Res*. **31**, 414–417.
34. Boeckmann B., Bairoch A., Apweiler R., Blatter M.C., Estreicher A., Gasteiger E., Martin M.J., Michoud K., O'Donovan C., Phan I., Pilbout S., Schneider M. 2003.

The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*. **31**, 365–370.

35. Falquet L., Pagni M., Bucher P., Hulo N., Sigrist C.J., Hofmann K., Bairoch A. 2002. The PROSITE database, its status in 2002. *Nucleic Acids Res*. **30**, 235–238.
36. International Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature*. **420**(6915), 520–562.
37. Нуртдинов Р.Н., Миронов А.А., Гельфанд М.С. 2002. Консервативен ли альтернативный сплайсинг генов млекопитающих? *Биофизика*. **47**(4), 197–203.
38. Nurtdinov R.N., Artamonova I.I., Mironov A.A., Gelfand M.S. 2003. Low conservation of alternative splicing patterns in the human and mouse genomes. *Hum. Mol. Genet*. **12**, 1313–1320.
39. Baranova A.V., Lobashev A.V., Ivanov D.V., Krukovskaya L.L., Yankovsky N.K., Kozlov A.P. 2001. In silico screening for tumour-specific expressed sequences in human genome. *FEBS Lett*. **200**, 143–148.

Prediction and Computer Analysis of the Exon–Intron Structure of Human Genes

A. A. Mironov and M. S. Gelfand

State Research Center GosNIIgenetika, Moscow, 113545 Russia
e-mail: aa_mironov@pochtamt.ru

Abstract—This review of the original works on computer analysis of the human genome considers the development of methods to predict the exon–intron structure of genes and analysis of alternative splicing. Prediction of the gene structure is based on homology between the gene product and a known protein or between the genomic sequences of the gene and its homolog from another organism. The methods were tested and proved highly efficient. Human gene splicing was analyzed with original methods and EST databases. Genes with alternative splicing were for the first time shown to account for no less than 35% total genes. Alternative splicing was compared for the human and mouse genomes. Species-specific isoforms were demonstrated for 50% alternatively spliced genes (25% total genes).