ДОПОЛНИТЕЛЬНЫЕ МАТЕРИАЛЫ К СТАТЬЕ

БИОИНФОРМАТИЧЕСКИЕ ПОДХОДЫ ДЛЯ ДЕТЕКЦИИ ГИБРИДНЫХ ГЕНОВ И ПРОДУКТОВ *ТРАНС*-СПЛАЙСИНГА

© 2024 г. И. Ю. Мусатов[*, **, #], М. И. Сорокин[**], А. А. Буздин[*, *, ****]**

**ФГАОУ ВО "Московский физико-технический институт (национальный исследовательский университет)",*

Россия, 141701, Долгопрудный, Институтский переулок, 9

***Институт персонализированной онкологии и персонализированного здравоохранения ФГАОУ ВО Первого МГМУ им. И.М. Сеченова Минздрава России (Сеченовский университет),*

Россия, 119048, Москва, ул. Трубецкая, 8/2

****Институт биоорганической химии им. академиков М.М. Шемякина и Ю.А. Овчинникова РАН, Россия, 117997, Москва, ул. Миклухо-Маклая, 16/10*

*****ГНЦ РФ ФГБУ Национальный медицинский исследовательский центр эндокринологии Минздрава России,*

Россия, 117292, Москва, ул. Дм. Ульянова, 11

Термины и определения

В настоящей статье применяют следующие термины с соответствующими определениями.

Альтернативный сплайсинг (alternative splicing) – это клеточный процесс, при котором экзоны одного и того же гена соединяются в разных комбинациях, что приводит к образованию разных, но связанных транскриптов мРНК. Эти мРНК можно транслировать для производства различных белков с разными структурами и функциями – и все это из одного гена [1].

Апоптоз (apoptosis) – регулируемый процесс программируемой клеточной гибели, в результате которого клетка распадается на отдельные апоптотические тельца, ограниченные плазматической мембраной.

Гемопоэтическая стволовая клетка (ГПСК) (hematopoietic stem cell) – стволовая клетка, участвующая в гемопоэзе (т.е. кроветворении); клетка, дающая начало клеткам крови.

Геном человека (human genome) – все три миллиарда пар оснований дезоксирибонуклеиновой кислоты (ДНК), составляющих весь набор хромосом человеческого организма. Геном человека включает кодирующие области ДНК, которые кодируют все гены

[#]Автор для связи: (тел.: +7 (985) 147-97-18; эл. почта: musatov.mailbox@yandex.ru).

(20000–25000) человеческого организма, а также некодирующие области ДНК, которые не кодируют никаких генов [2].

Граф (graph) – это геометрическая фигура, которая состоит из точек и линий, которые их соединяют. Точки называют вершинами графа, а линии – ребрами.

Ориентированным графом (обозначается G) (directed graph) называется пара G = (V, E), где V – множество вершин, а E ⊂ V × V, множество ребер, соединяющих эти вершины.

Граф де Брюйна (иногда также граф де Брейна) **(de Brujin graph)** с параметром l для n-буквенного алфавита (обозначается B (n,l)) – это ориентированный граф G (V, E) [3], где V – множество всех слов длины l в заданном алфавите, (u,v) – вершины графа, соединенные линией, называемой ребром, а E ⊂ V × V – множество ребер.

Примечание 1 [3]

Граф де Брюйна для множества прочтений в задаче сборки генома:

Пусть имеются все различные суффиксы и префиксы прочтений длиной (l – 1).

Построим граф D, в котором префиксы и суффиксы являются вершинами. Будем соединять вершины V_1 и V_2 направленным ребром только тогда, когда существует прочтение с префиксом V_1 и суффиксом V_2. Каждое исходной прочтение будет соответствовать ребру графа. Такой граф будет являться подграфом графа де Брюйна B (4,l), т.к. получился таким же методом построения, просто по меньшему множеству строк. Каждый путь в графе в таком случае будет соответствовать некоторой строке. Начнем с (k – 1)-мера. Проходя по каждому ребру, будем приписывать последний нуклеотид соответствующего k-мера справа к имеющейся строке. В итоге получится строка длины (k – 1) + (количество рёбер в пути) [3].

Примечание 2

Наличие вышеприведенных условий для графа де Брюйна позволяет сделать следующий вывод: существует слово L длины l + 1 в заданном алфавите такое, что u = prefix (L) и v = suffix(L).

Граф де Брюйна основан на k-мерах. Для осмысления того, как именно происходит построение графа де Брюйна для генома, сформулируем упрощенную задачу и приведем ее решение [3].

Задача

Пусть имеется множество прочтений длины l нуклеотидов (букв). Требуется построить возможный геном, если известно, что все подстроки генома длины l входят в данное множество прочтений.

Объяснение

Геном, как большая строка, имеет свои подстроки – прочтения. Как геном, так и прочтения состоят из четырехсимвольного алфавита {A, T, G, C}, по-другому – нуклеотидов; для РНК-генома в алфавите тимин (T) заменяется на урацил (U). Часто длина прочтений лежит в диапазоне 100–1000 нуклеотидов, а сам геном может содержать порядка миллиона нуклеотидов у простейших организмов. При этом информацию о геноме мы получаем опосредовано – через прочтения, в ходе секвенирования.

Решение

Строится граф G, где вершины – суффиксы и префиксы длины l – 1 всех прочтений. Получается подграф графа де Брюйна B (4, l – 1). Подграф, поскольку в таком графе присутствуют не обязательно все 4^{l-1} вершины, где каждому ребру соответствует прочтение. Найдем в таком графе путь, который пройдет через каждую вершину только один раз (“эйлеров путь”). Такой путь существует, поскольку на исследуемый геном было наложено условие, что все подстроки длины l входят в множество прочтений. Такой путь является возможным ответом.

Примечание 3

Единственно верный ответ (реальный геном) можно получить не всегда, т.к. в графе не всегда есть такой эйлеров путь.

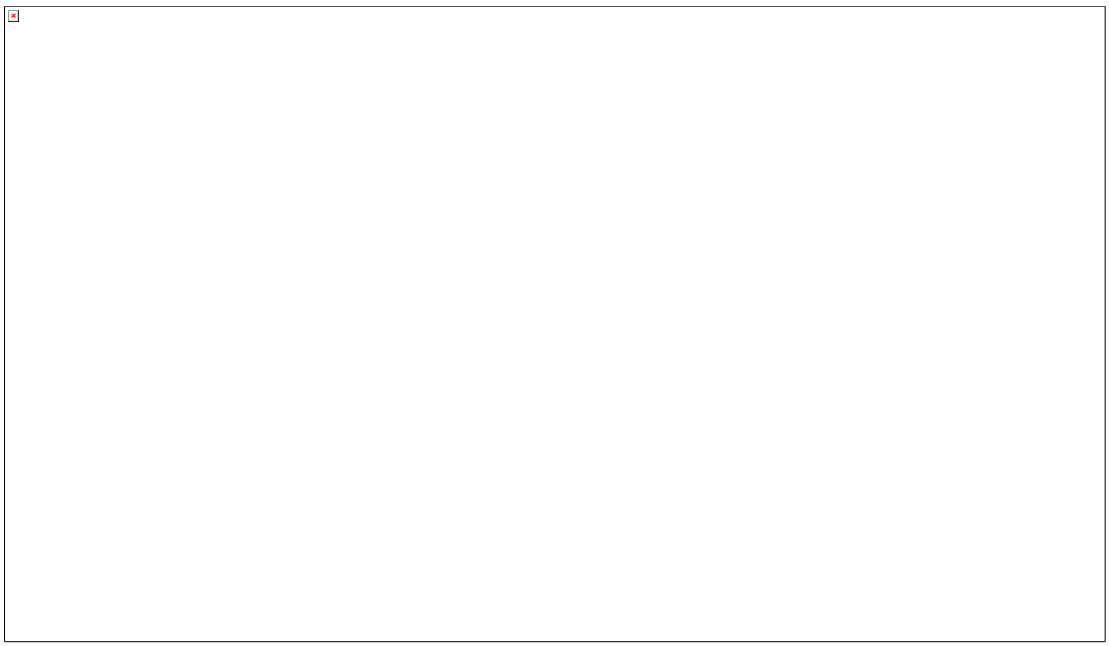

Рис. S1. Пример построения графа де Брюйна из ридов (указанных на рёбрах графа) – их префиксов и суффиксов для гипотетического генома ATGTATTAC.

Проблемы построения графа де Брюйна в реальных геномах:

1) существуют повторяющиеся прочтения;
2) существуют ошибки в самих прочтениях (делеции, инсерции, несовпадения, однонуклеотидные полиморфизмы);
3) прочтения имеют неодинаковую длину;
4) время работы алгоритма O($4^l \times l$).

Длина прочтений (read length) – число пар нуклеотидов (п.н. – bp) секвенированного ДНК-фрагмента.

Интроны (intron, intragenic region) – участки ДНК, соответствующие участкам, удаляемым из зрелого РНК транскрипта в процессе сплайсинга.

***k*-мер (*k*-mer)** – строка длиной *k* нуклеотидов (букв) в алфавите (A, T, G, C).

Примечание: для РНК алфавит будет состоять из (A, U, G, C).

Контиг (contig(uous)) [4] (применительно к геномным исследованиям, происходит от слова contiguous, т.е. “смежный”) – представляет собой набор сегментов ДНК- или РНК-последовательностей, которые перекрываются таким образом, что обеспечивают непрерывное представление некоторой геномной области.

MMP – алгоритм поиска максимального отображаемого префикса MMP (maximum mappable prefix) – алгоритм, осуществляющий поиск начальной подстроки максимально возможной длины, которая бы совпадала с эталонной строкой и начиналась с первого символа строки [5], в свою очередь, MMP – максимальный префикс, который можно сопоставить с геномной/транскриптомной последовательностью.

Мутагенез (mutagenesis) – процесс изменения в нуклеотидной последовательности ДНК, приводящий к мутациям. Различают естественный (спонтанный) и искусственный (индуцированный) мутагенез.

Мутация (mutation) – это изменение последовательности ДНК организма. Мутации могут возникать в результате ошибок репликации ДНК во время деления клеток, воздействия мутагенов или вирусной инфекции. Зародышевые мутации (происходящие в яйцеклетках и сперматозоидах) могут передаваться потомству, тогда как соматические мутации (происходящие в клетках организма) потомству не передаются [6].

Однонуклеотидный полиморфизм (SNP – single nucleotide polymorphism) – отличие в последовательности генома образца и эталонного генома, т.е. вариант генома, отличающийся в некотором положении на одно азотистое основание ДНК.

Открытая рамка считывания (open reading frame, ORF) – последовательность нуклеотидов в составе ДНК, которая при трансляции в аминокислоты не содержит стоп-кодонов в своей последовательности [7].

Примечание: генетический код считывается рибосомой по 3 п.н., это означает, что двухцепочечная молекула ДНК может быть прочитана в любой из шести возможных открытых рамок считывания – три в прямом направлении и три в обратном.

Префикс (prefix) – подстрока, начинающаяся с первого символа строки.

Пролиферация клеток (cell proliferation) – то же, что деление клеток.

Прочтения (reads) – ДНК/РНК-фрагменты, возникающие в результате секвенирования.

Путь в графе – некоторая определенная последовательность вершин и рёбер.

Пример: путём называется последовательность вида $v_0 e_1 v_1 \dots e_k v_k$, где $e_i \in E, e_i = (v_{i-1}, v_i)$, e – от англ. edges (рёбра), v – vertices (вершины), соответственно, E – множество всех рёбер, а V – множество всех вершин.

Реципрокные транслокации (reciprocal translocations) (reciprocal – обоюдный) – хромосомные перестройки, при которых происходит взаимный обмен участками между двумя хромосомами.

РНК-транскрипт (транскрипт) (RNA transcript) – фрагмент РНК, полученный в результате транскрипции ДНК в РНК.

Секвенирование длинными прочтениями (long-read sequencing) (также в тексте **“длинные прочтения”**) **–** метод определения ДНК/РНК-последовательности, в котором длина прочтения лежит в диапазоне 5000–100000 п.н.

Секвенирование короткими прочтениями (short-read sequencing) (также в тексте **“короткие прочтения”**) – метод определения ДНК/РНК-последовательности, в котором длина прочтения не превышает 300 п.н.

Соматические транслокации (somatic translocations) – транслокации, возникающие в соматических клетках в результате ошибок митоза; приводят к аномалиям, которые затрагивают определенную клеточную линию.

Стоп-кодон – это последовательность из трех нуклеотидов в ДНК или информационной РНК (мРНК), которая сигнализирует о необходимости прекратить синтез белка. Существует 64 различных кодона: 61 кодон определяет аминокислоты, три кодона являются стоп-кодонами (т.е. UAA, UAG и UGA) [8].

Структурный вариант (СВ) (genomic structural variant) – изменения генома, которые затрагивают более одной (обычно ≥5) пар оснований. Основные типы СВ включают делеции, инсерции, дупликации (тандемные или вкрапленные) и инверсии [9].

Примечание: синонимом термина “структурный вариант” в данной статье будем считать слово “изоформа” по отношению к гибридным генам.

Суффикс (suffix) – подстрока, заканчивающаяся на последний символ строки.

Тирозинкиназа (tyrosine kinase) – фермент, переносящий фосфатную группу от АТФ к остаткам тирозина определенных белков внутри клетки. Он функционирует как переключатель, “включая” и “выключая” многие клеточные процессы.

Точки разрыва (breakpoints) – участки ДНК, последовательность которых существенно отличается от эталонной последовательности, т.е. такие места на хромосоме, где ДНК может быть удалена, инвертирована или может произойти транслокация (определение взято из **COSMIC** – Catalogue of Somatic Mutations in Cancer) [10].

Примечание: термин “точка разрыва” используется также иногда в контексте поиска гибридных генов и транскриптов и может означать “точку слияния” (**fusion point**), т.е. фактически являться тем местом, где два гена слились в один. Поэтому стоит обращать внимание на контекст, в котором используется термин.

Транскриптом (transcriptome) – это полный набор молекул информационной РНК или мРНК, экспрессируемых организмом. Термин “транскриптом” также можно использовать для описания массива транскриптов мРНК, продуцируемых в определенном типе клеток или тканей. В отличие от генома, который отличается стабильностью, транскриптом активно изменяется [11].

Фиксированные формалином парафинизированные образцы ткани (FFPE sample) – образцы ткани, взятые у пациента для последующего анализа, фиксированные формалином и парафинизированные для консервации и сохранения структуры ткани. Нуклеиновые кислоты, в особенности РНК, в таких образцах, как правило, присутствуют в виде коротких фрагментов (деградированная РНК) [12].

“Филадельфийская хромосома” (Philadelphia chromosome) – дефектная, укороченная хромосома 22, возникающая в результате реципрокной транслокации t(9;22)(q34;q11) между хромосомами 9 (в ней находится ген *ABL1*) и 22 (в ней находится участок *BCR*). Содержит хромосомную перестройку *BCR-ABL*. В результате транслокации образуется укороченная 22-я хромосома и удлиненная 9-я хромосома и новый химерный белок BCR-ABL, содержащий участок белка ABL1, обладающий тирозинкиназной активностью. В норме тирозинкиназы являются медиаторами, активируемыми связыванием факторов роста с мембранными рецепторными белками и передающими сигнал к делению на ядро, но BCR-ABL представляет собой постоянно активную форму и вызывает неконтролируемое деление клетки.

Химерный (слитный/гибридный) ген (fusion gene) – ген, образовавшийся в результате слияния двух ранее независимых генов. Такой ген может образоваться в результате транслокации, хромосомной инверсии или делеции.

Хромосомная транслокация (chromosome translocation) – взаимный обмен участками между двумя хромосомами.

Хронический миелоидный лейкоз (ХМЛ, хронический миелобластный лейкоз, хроническая миелоидная лейкемия) (chronic myelogenous leukemia, CML) – болезнь, вызванная миелопролиферативным новообразованием. В случае хронического миелолейкоза происходит клональная пролиферация злокачественно измененной мультипотентной стволовой клетки, причем “хронический” указывает на медленную скорость развития заболевания. Причина заболевания – соматическая мутация в гемопоэтической стволовой клетке, в результате которой образуется укороченная “филадельфийская хромосома”. Филадельфийская хромосома возникает в случае неудачной транслокации между хромосомами 9 и 22, при которой образуется гибридный ген *BCR-ABL*, усиливающий тирозинкиназную активность, влияющую на скорость пролиферации клеток.

Хэширование – процедура, позволяющая преобразовать исходный массив данных произвольной длины в битовую строку фиксированной длины.

Примечание:

В контексте данного обзора, как правило, хэшироваться будут подстроки генома, прочтения, или транскрипты, состоящие из нуклеотидов, т.е. речь будет о преобразовании текстовой строки из алфавита {A, T (U – для РНК), G, C} произвольной длины в некоторое числовое значение, которое тем ближе по структуре последовательности своих символов к некоторому эталону, чем ближе вычисленный хэш (число) у исследуемой последовательности к значению хэша (числа) эталона.

Эйлеров путь в графе (Eulerian path) – это путь, который проходит по каждому ребру, причем ровно один раз.

Экзон (exon) (от англ. ex(pressi)on – экспрессия, выражение) – часть гена, кодирующая зрелый РНК-транскрипт, т.е. РНК, которая получается после того, как в процессе сплайсинга будут удалены все интроны.

Экспрессия (также экспрессия генов) **(gene expression)** – слово, обозначающее процесс, в ходе которого генная информация используется для получения функционирующего продукта гена.

СПИСОК ЛИТЕРАТУРЫ

1. Alternative Splicing Talking Glossary of Genomic and Genetic Terms. National Human Genome Research Institute. https://www.genome.gov/genetics-glossary/Alternative-Splicing
2. *Fridovich-Keil J.L.* // Human Genome. In: Britannica. https://www.britannica.com/science/human-genome
3. *Nurk S., Bankevich A., Antipov D., Gurevich A.A., Korobeynikov A., Lapidus A., Prjibelski A.D., Pyshkin A., Sirotkin A., Sirotkin Y., Stepanauskas R., Clingenpeel S.R., Woyke T., McLean J.S., Lasken R., Tesler G., Alekseyev M.A., Pevzner P.A.* // J. Comput. Biol. 2013. V. 20. P. 714–737. https://doi.org/10.1089/cmb.2013.0084
4. Contig // Talking Glossary of Genomic and Genetic Terms. National Human Genome Research Institute. https://www.genome.gov/genetics-glossary/Contig
5. *Dobin A., Davis C.A., Schlesinger F., Drenkow J., Zaleski C., Jha S., Batut P., Chaisson M., Gingeras T.R.* // Bioinformatics. 2013. V. 29. P. 15–21. https://doi.org/10.1093/bioinformatics/bts635
6. Mutation // Talking Glossary of Genomic and Genetic Terms. National Human Genome Research Institute. https://www.genome.gov/genetics-glossary/Mutation
7. Open-Reading-Frame // Talking Glossary of Genomic and Genetic Terms. National Human Genome Research Institute. https://www.genome.gov/genetics-glossary/Open-Reading-Frame
8. Stop Codon // Talking Glossary of Genomic and Genetic Terms. National Human Genome Research Institute. https://www.genome.gov/genetics-glossary/Stop-Codon
9. *Gawroński A.R., Lin Y.Y., McConeghy B., LeBihan S., Asghari H., Koçkan C., Orabi B., Adra N., Pili R., Collins C.C., Sahinalp S.C., Hach F.* // Nucleic Acids Res. 2019. V. 47. P. e38. https://doi.org/10.1093/nar/gkz067
10. Breakpoint // COSMIC – catalogue of somatic mutations in cancer https://cancer.sanger.ac.uk/cosmic/help/rearrangement/overview
11. Transcriptome Fact Sheet // Fact Sheets about Genomics. National Human Genome Research Institute. https://www.genome.gov/about-genomics/fact-sheets/Transcriptome-Fact-Sheet
12. *Yi Q.-Q., Yang R., Shi J.-F., Zeng N.-Y., Liang D.-Y., Sha S., Chang Q.* // J. Int. Med. Res. 2020. V. 48. https://doi.org/10.1177/0300060520931259