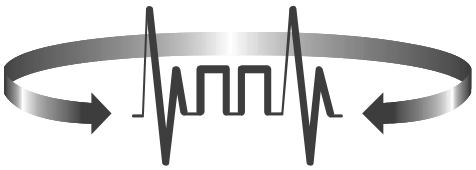

30-я КОНФЕРЕНЦИЯ МОЛОДЫХ УЧЕНЫХ
И СПЕЦИАЛИСТОВ ИППИ РАН

Информационные технологии и системы

ИТиС'07

СБОРНИК ТРУДОВ КОНФЕРЕНЦИИ

г. Звенигород, 18-21 сентября 2007 г.

ББК 32.81
И 741

Информационные Технологии и Системы (ИТиС’07)
Москва: ИППИ РАН, 2007. – 370с.

Издание содержит труды 30-й конференции молодых ученых и специалистов Инсиститута проблем передачи информации им. А.А.Харкевича Российской академии наук (ИППИ РАН) «Информационные технологии и системы» (ИТиС’07). Конференция явилась продолжением серии традиционных конференций, организованных Советом молодых ученых и специалистов ИППИ РАН в предшествующие годы.

Основная цель Конференции ИТиС’07 – дать возможность молодым ученым и специалистам различных подразделений ИППИ РАН, а также студентам, аспирантам и молодым ученым других институтов РАН, отраслевых институтов, университетов и вузов, познакомиться с коллегами и обменяться научными достижениями по основным для ИППИ РАН направлениям научной деятельности: теория передачи и защиты информации; математическая теория информации и управления, многокомпонентные случайные системы; информационно-коммуникационные технологии и их применение в сложных системах и сетях; информационные процессы в живых системах и биоинформатика; компьютерная лингвистика и моделирование естественного языка.

Все включенные в данный сборник работы опубликованы в том виде, в котором они были представлены авторами, среди которых молодые ученые и специалисты ИППИ РАН, ИМБ РАН, ИМГ РАН, ИОГен РАН, ИПУ РАН, МГТУ им. Н.Э. Баумана, МГУ им. М.В. Ломоносова, МИФИ (ГУ), МТУСИ, МФТИ (ГУ), НИИ биомедицинской химии РАМН, УФСИН России по Псковской области, ФГУП ГосНИИГенетика и Центрального НИИ эпидемиологии Роспотребнадзора.

Труды Конференции могут представлять интерес для ученых, студентов и аспирантов, специализирующихся в областях науки, связанных с перечисленными выше научными направлениями.

Конференция проведена при финансовой поддержке Целевой программы Президиума РАН «Поддержка молодых ученых 2007 г.».

ISBN 978-5-7834-0193-0

Кластеризация и поиск промоторов у хлоропластов

А.В. Селиверстов, В.А. Любецкий
Институт проблем передачи информации им. А.А. Харкевича
Российской академии наук
slvstv@iitp.ru, lyubetsk@iitp.ru

Аннотация

Предложен алгоритм для быстрого поиска кластеров в наборе векторов; координаты вектора - частотные характеристики данной нуклеотидной последовательности. С помощью этого алгоритма приведено исследование 5′-лидерных областей генов у хлоропластов. В частности, предсказаны промоторы генов psbA, psbB, psbE, psaA у семенных растений и гена ndhF у крестоцветных и близких семейств. На этой основе предполагается, что промотор располагается на 5′-границе внутри слабо консервативных участков в 5′-лидерных областях генов, так что левее расположен неконсервативный участок.

1. Введение

Поиск консервативных участков внутри 5'-лидерных областей ортологичных генов у представителей некоторой филогенетической группы имеет большое значение при поиске сайтов белок-ДНКовой и других регуляций экспрессии генов. Известно несколько подходов к отбору из большого исходного набора нуклеотидных последовательностей части, которая состоит из попарно похожих последовательностей, по всей их длине или лишь на некотором сравнительно коротком их участке. В типичной ситуации заранее не известно, какие из исходных последовательностей содержат такой участок, неизвестна даже доля «мусорных» последовательностей, т.е. таких, которые не содержат такого участка; отсутствует и какое-либо структурное описание таких участков.

Обычно для этого используется множественное выравнивание последовательностей в целом или тот или иной поиск локального консервативного участка (например, как клики в графе). Соответствующие алгоритмы имеют время работы экспоненциально возрастающее с увеличением числа последовательностей. Известно несколько эвристических алгоритмов, которые дают хорошее выравнивание для достаточно близких последовательностей. Однако при исследовании лидерных областей часто приходится работать с достаточно далекими последовательностями и искать слабо консервативные участки. При этом среди последовательностей много таких, которые значительно отличаются от остальных, - это мешает применять эвристические алгоритмы. Поэтому представляет интерес представление исходного набора последовательностей точками в пространстве небольшой размерности так, чтобы расстояние между точками отражало сходство биологически значимых характеристик последовательностей, которые часто заранее неизвестны; в этой работе нами предлагается такое представление. Затем выполняется кластеризация этих точек; эта задача хорошо известна в вычислительной геометрии, [1]. Наконец, происходит множественное выравнивание последовательностей, попавших в один достаточно большой кластер; для этого нами применялась программа MultAlign, предоставленная А.А. Мироновым, который её разработал.

2. Методы

Соответствие между нуклеотидными последовательностями, т.е. словами в алфавите {A, C, G, T}, и точками в произведении некоторого числа рациональных прямых, т.е. упомянутое выше **представление**, определяется следующим образом. Фиксируем натуральное число k. Любому данному слову сопоставляется точка в $16k$-мерном пространстве, координаты которой равны частотам встречаемости в этом слове пар букв, у которых расстояние между буквами принадлежит одному из множеств {0}, {1}, {2, 3}, {4, 5, 6, 7},…, $\{2^{k-2},\ldots, 2^{k-1}-1\}$. Таким образом, первые 16 координат соответствуют всевозможным парам букв, идущих подряд. Следующие 16 координат соответствуют парам букв, разделённых одной произвольной буквой. Следующие 16 координат соответствуют парам

букв, разделённых двумя произвольными буквами и т.д.

Практический интерес представляют слова, состоящие из нескольких сотен букв. Поэтому параметр k не превосходит десяти. Важно отметить, что усреднение по разным расстояниям между рассматриваемыми позициями букв, которое происходит при таком представлении, позволяет изучать сходство последовательностей, имеющих делеции и вставки. При работе с близкими последовательностями можно выбирать значение k=1, хотя учёт частот только подряд идущих пар букв представляется слишком грубым. Замена пар букв на тройки букв и более привело бы к тому, что каждый набор букв имел бы низкую частоту встречаемости.

Итак, набору исходных нуклеотидных последовательностей фиксированной длины соответствует набор точек в 16k-мерном пространстве. Этот набор **определяется** как кластер, если расстояние между точками из набора и центром минимального объемлющего набор «стандартного» прямоугольного параллелепипеда достаточно мало по сравнению со средним расстоянием между этим центром и случайными наборами той же мощности из последовательностей той же длины, которые являются подсловами у слов из некоторого множества A. «Стандартного» означает, что стороны параллелепипеда параллельны координатным осям. Вместо параллелепипеда можно рассматривать сферу, но это приводит к значительному усложнению вычислительной стороны дела.

В качестве множества A мы брали множество 5'-лидерных областей всех генов из рассматриваемых геномов.

В качестве расстояния бралась сумма модулей разностей координат точек.

Для исходного набора последовательносей легко алгоритмически проверяется, образуют ли соответствующие точки в пространстве кластер. Если нет, то наш алгоритм по очереди удаляет точки, начиная с наиболее удалённых от центра объемлющего параллелепипеда, пока не останется кластер. Обычно имеются биологические сведения, позволяющие контролировать удаление точек до получения кластера; это необходимо потому, что понятие кластера включает порог, выбор которого нужно делать по ходу вычислений. Например, такими биологическими сведениями является эволюционная близость рассматриваемых геномов.

3. Тестирование

Известно, [2], что ген *rbcL*, кодирующий большую субъединицу рибулёзо-1,5-бисфосфаткарбоксилазы – главного фермента цикла Кальвина-Бенсона, у *Astasia longa* транскрибируется в пластидах и содержит восемь экзонов, непосредственно перед первым из которых расположена протяжённая 5'-нетранслируемая область мРНК длиной 147н.

Виды *Astasia longa* и *Euglena gracilis* принадлежат таксону Euglenozoa, т.е. относятся к простейшим животным. Вид *Mesostigma viride* - водоросль из таксона Streptophyta, объединяющего высшие растения и некоторые водоросли. Третий вид эволюционно далек от первых двух. Был рассмотрен набор из 5'-лидерных областей длиной 147н перед всеми генами у *Mesostigma* и также перед геном *rbcL* у *Astasia* и *Euglena*. В соответствующем наборе точек при k=1 нашим алгоритмом найден кластер, содержащий пять точек, из которых три соответствуют гену *rbcL* из этих трёх видов.

4. Поиск промоторов

4.1. Предварительные сведения

В работе [3] в 5'-лидерных областях генов *clpP*, *psaA*, *psbA*, *psbB* найдены консервативные участки, общие для высших растений и некоторых водорослей.

Эти участки примыкают к инициирующим кодонам генов и, вероятно, участвуют в регуляции на уровне трансляции. Для гена *psbA* такая регуляция экспериментально найдена у хламидомонады. Нами продолжено исследование 5'-лидерных областей генов с целью изучения регуляции на уровне транскрипции. Напомним, что промотором называется область в 5'-лидерной области гена, которая служит для связывания сигма-субъединицы РНК-полимеразы и важна для инициации транскрипции. Известно, что многие промоторы хлоропластов и бактерий близки между собой и включают два относительно консервативных участка каждый из шести нуклеотидов, расположенных вне транскрибируемой области вблизи -35 и -10 нуклеотидов, считая от старта траскрипции, [4].

4.2. Результаты

Области около нескольких промоторов, указанных в аннотации для хлоропласта чёрной сосны *Pinus thunbergii*, хорошо выравниваются с областями перед ортологичными генами цветковых, [5].

На множественном выравнивании 5'-лидерных областей генов *psbA*, *psbE*, *psbB*, кодирующих некоторые белки второй фотосистемы, видны консервативные промоторы вида TTG–21н–(-10 бокс), где -10 бокс промотора для *psbA* и *psbE* абсолютно консервативен и равен TATACT, а для *psbB* равен TAGAAT у цветковых и TAAAAT у *Pinus thunbergii* и *Psilotum nudum* (в последнем случае ортолог для гена *psbB* назван *psbT*).

Кроме того, у цветковых растений транскрибируемые области перед генами *psbA* и *psbE* содержат длинные консервативные области. При этом области, расположенные перед -35 боксом размеченного промотора, не выравниваются.

Для гена *psbB* выравниваются лишь две короткие области: первая – около промотора, вторая – вблизи старта трансляции. Причём вторая область консервативна также у некоторых водорослей, включая *Chaetosphaeridium globosum*, *Chara vulgaris*, *Nephroselmis olivacea*.

Промотор перед геном *psaA*, кодирующим белок первой фотосистемы, также консервативен у цветковых и сосны (*P. thunbergii*). На множественном выравнивании видны две консервативные области: первая – от -35 бокса простирается довольно далеко в транскрибируемую область; вторая примыкает непосредственно к инициирующему кодону и консервативна не только у растений, но и у многих водорослей, включая диатомовые, криптофитовые, красные, харовые и зелёную *Nephroselmis olivacea*. Возможно, она связана с регуляцией трансляции, [3].

Одновременно обнаружено несколько промоторов, специфичных для небольших таксономических групп. Интересен промотор перед геном *ndhF*, имеющий -10 бокс с консенсусом TATAAA. Он найден у всех крестоцветных, а также у *Citrus sinensis*, *Gossypium* spp., *Eucalyptus globulus* и *Oenothera elata*.

Во всех случаях этот участок примыкает к слабо консервативной области, простирающейся вплоть до открытой рамки считывания *ndhF*. Такое выравнивание не продолжается ни на один вид других семейств, включая относительно близкие на дереве видов бобовые (*Glycine max*, *Lotus corniculatus*, *Phaseolus vulgaris*), огурец (*Cucumis sativus*), тополь (*Populus alba*, *P. trichocarpa*), герань *Pelargonium x hortorum*, хотя они входят вместе с крестоцветными и миртовыми в таксономическую группу rosids.

В целом рассмотренные 5'-лидерные области гена *ndhF* у *Eucalyptus globulus* ближе всего к таковым у *Citrus sinensis*.

4.3. Обсуждение

Консервативные участки, найденные в начале транскрибируемого участка (первая область), с одной стороны отделены от инициирующего кодона протяженным слабо консервативным промежутком. Поэтому они не связаны с регуляцией трансляции.

С другой стороны, такое положение консервативного участка не характерно для связывания транскрипционного фактора (репрессора), который обычно либо перекрывает промотор, либо расположен перед ним, взаимодействуя с сигма- и альфа-субъединицами РНК-полимеразы. Таким образом, протяжённые консервативные области в начале траскрибируемых областей генов *psbA*, *psbE*, *psbB* и *psaA* содержат альтернативный промотор или связаны с устойчивостью мРНК.

Это наблюдение представляет метод для предсказания промоторов как 5'-границ слабо консервативных областей в 5'-лидерных областях генов.

5. Благодарности

Авторы благодарны Л.И. Рубанову за обсуждение методов исследования и выражают особую признательность Е.А. Лысенко, привлекшему внимание к проблеме поиска промоторов.

Работа поддержана грантом Международного научно-технического центра (ISTC 2766).

6. Литература

[1] D.T. Lee, F.P. Preparata, “Computational geometry – A survey”, *IEEE Transactions on Computers*, C33(12), 1984, pp. 1072-1101.

[2] G. Siemeister, W. Hachtel, “Structure and expression of a gene encoding the large subunit of ribulose-1,5-bisphosphate carboxylase (*rbcL*) in the colourless euglenoid flagellate *Astasia longa*”, *Plant Mol. Biol*., 14(5), 1990, pp. 825-833.

[3] A. Seliverstov, V. Lyubetsky, “Translation regulation of intron-containing genes in chloroplasts”, *J. Bioinform. Comput. Biol*., 4(4), 2006, pp. 783-792.

[4] E.A. Lysenko, “Plant sigma factors and their role in plastid transcription”, *Plant Cell Rep.,* 2007 (in press).

[5] T. Wakasugi, J. Tsudzuki, S. Ito, K. Nakashima, T. Tsudzuki, and M. Sugiura, "Loss of all *ndh* genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii*", *Proc. Natl. Acad. Sci. U.S.A*., 91 (21), 1994, pp. 9794-9798.