Описание набора сценариев «RepIns» для поиска вставок прямых повторов в некодирующих областях геномов близкородственных видов

Версия от 18.11.2010

Общие сведения

Набор сценариев «RepIns» состоит из трёх сценариев, соответствующих трём этапам алгоритма, подробно описанным ниже:

1. сценарий **ncds.py** формирует группы гомологичных некодирующих областей;
2. сценарий **malign.py** выполняет множественные выравнивания в пакетном режиме;
3. сценарий **repins.py** производит поиск вставок повторов и сбор статистики их длин.

Сценарии организованы таким образом, т.к. выполняют слабо связанные друг с другом задачи и могут быть использованы независимо друг от друга в рамках других задач.

Требования к среде исполнения

Набор сценариев написан на языке Python 2 и выполняется в среде интерпретатора Python версии 2.7, предустанавливаемого во многих современных операционных системах (информацию по установке и использованию интерпретатора Python можно получить на сайте http://python.org/). Для ввода данных в формате GenBank используется модуль Biopython версии 1.55, который должен быть соответствующим образом установлен (информацию по установке Biopython можно найти на сайте http://biopython.org/). Для выполнения множественного выравнивания нуклеотидных последовательностей используется программа MUSCLE, которая должна быть доступна в виде исполняемого файла (информацию по программе MUSCLE можно получить по адресу http://www.drive5.com/muscle/). Все перечисленные программные продукты свободно доступны и имеют лицензии, разрешающие их бесплатное использование (более подробную информацию об условиях использования смотрите на соответствующих сайтах). Набор сценариев не предъявляет специфических требований к операционной системе и, вероятно, может быть успешно использован (возможно, с незначительными модификациями) во множестве операционных систем семейств Windows, UNIX и Linux, но фактически работоспособность проверялась лишь в системе Windows XP.

Описание алгоритма

Алгоритм поиска вставок прямых повторов в некодирующих областях геномов близкородственных видов состоит из трёх основных этапов, выполняемых последовательно:

1. подготовка набора групп некодирующих последовательностей;
2. построение множественных выравниваний некодирующих последовательностей;
3. поиск вставок прямых повторов во множественных выравниваниях.

Подготовка набора групп некодирующих последовательностей

Функцией этого этапа является *вырезание* из геномных последовательностей *некодирующих участков* в соответствии с заданными аннотациями и двухуровневая *группировка* полученных последовательностей: по *гомологичным парам генов* и *семействам видов*.

Входными данными для этого этапа являются **набор аннотированных геномных последовательностей** и **список семейств**, объединяющих подмножества этого набора в близкородственные группы, каждая из которых исследуется в дальнейшем независимо.

Выходными данными этого этапа являются **наборы групп гомологичных некодирующих последовательностей**. Каждая *группа* набора содержит две или более некодирующие последовательности, расположенные между *соответствующими* парами генов из разных геномов семейства, где под *соответствующими* парами понимаются пары генов (A1, B1) и (A2, B2), где A1 — гомолог A2, B1 — гомолог B2. Каждый *набор* объединяет группы одного семейства.

Для каждого генома, в соответствии с его аннотацией, формируется список всех *межгенных промежутков*, т.е. непродолжаемых непрерывных областей геномной последовательности, не покрытых ни одним геном. Каждый промежуток снабжается списками ограничивающих его с каждой стороны генов (каждый ген представлен парой его координат, именем и номером экзона). Для каждого промежутка определяется *идентификатор группы* — упорядоченная пара специальным образом унифицированных имён генов (включающих номера экзонов), ограничивающих данный промежуток с разных сторон, записанных в лексикографическом порядке. Промежутки собираются внутри каждого семейства в группы по совпадению идентификаторов групп. Соответствующие им некодирующие последовательности, сгруппированные по семействам и по группам, являются результатом работы алгоритма. В случаях, когда ограничивающие некодирующую область гены расположены в порядке, обратном указанному в идентификаторе группы, она выводится в виде обратной комплиментарной последовательности.

Построение множественных выравниваний

Функцией этого этапа является построение *множественных выравниваний* соответствующих некодирующих последовательностей внутри каждого семейства видов.

Входными данными для этого этапа являются созданные на предыдущем этапе **наборы групп гомологичных некодирующих областей**.

Выходными данными этого этапа являются **выравнивания** гомологичных некодирующих областей.

На данном этапе для каждой группы некодирующих последовательностей строится множественное выравнивание. При этом сохраняются метаданные последовательностей и группировка последовательностей по гомологии соседних генов и семействам видов. Фактический алгоритм построения множественных выравниваний может быть различным в зависимости от используемой для этой цели программы. С алгоритмом, используемым программой MUSCLE можно ознакомиться на сайте программы: http://www.drive5.com/muscle/.

Поиск вставок прямых повторов

Функцией этого этапа является поиск в выравниваниях вставок специального типа и сбор статистики распределения частоты таких вставок в отдельных семействах по длинам образующих их последовательностей.

Входными данными для этого этапа являются **множественные выравнивания** гомологичных некодирующих областей.

Выходными данными этого этапа являются **список** вставок прямых повторов и **распределение** вставок по длинам образующих их повторов.

Для **каждой пары** P последовательностей **каждого выравнивания** независимо производится поиск вставок прямых повторов в соответствии с нижеописанным **алгоритмом**.

По паре P последовательностей выравнивания строится пара вспомогательных последовательностей: S и S'. Последовательность S содержит во всех абсолютно консервативных позициях P соответствующие буквы, во всех позициях, где есть хоть один символ делеции, — символы делеции, во всех остальных позициях — пробелы. Последовательность S' строится так же за исключением того, что делеции игнорируются (т.е. на месте делеций в последовательности S — последовательность S' содержит буквы в «условно абсолютно консервативных» позициях P (т.е. таких позициях, где множество различных букв состоит из одной буквы) и пробелы в остальных).

В S находятся координаты (пары начало-конец) всех вхождений регулярного выражения "[ACTG]-+[ACTG]", т.е. «буква нуклеотида, один или более символов делеции, буква нуклеотида» (точнее, координаты, соответствующие в этом выражении группе делеций). Из списка координат отсеиваются пары, между которыми в S' встречаются пробелы. Для каждой оставшейся пары координат — в S' читается соответствующее им слово, в нём находится минимальный период повтора и соответствующее образующее слово (т.е. минимальное слово, которое при циклическом повторении даёт в результате данную последовательность). Слева и справа (от координат вставки)

в S' ищется максимально длинная серия точных повторений образующего слова. Вставки, не имеющие повторов хотя бы с одной стороны, отсеиваются.

Каждая найденная вставка прямого повтора сохраняется в виде кортежа (L, R, G, F_i, F_o), элементы которого имеют следующие значения:

- L — координата начала вставки, т.е. первой позиции, содержащей символ делеции;
- R — координата конца вставки, т.е. первой буквой, следующей за делецией;
- G — образующее слово;
- F_i — кратность повторения G внутри вставки;
- F_o — кратность повторения G за границами вставки.

Из полученного в результате многократного выполнения вышеприведённого алгоритма множества кортежей исключаются повторяющиеся кортежи (из каждой серии одинаковых кортежей остаётся лишь один кортеж). Внутри каждого семейства видов производится подсчёт числа вставок, имеющих одинаковую длину образующего слова. Полученный **список** вставок и **распределение** их числа по длинам образующих являются **результатом** работы алгоритма.