УДК 575.852

Любецкая Е.В. [1], *Горбунов К.Ю.* [1]

[1] Институт проблем передачи информации РАН

Алгоритмы реконструкции эволюции регуляторных сигналов

Постановка задачи. В [1] формулируются понятия множественного выравнивания и сигнала, составленного из сайтов в данных последовательностях. Часто сигнал характеризуется дополнительными требованиями на искомые сайты, которые должны поддерживаться в ходе эволюции. Например, можно требовать, чтобы сигнал допускал определённую комбинацию спиралей, определённым образом расположенных относительно друг друга. Спиралью называется пара отрезков («плеч»), которые могут быть побуквенно спарены в соответствии с правилом G с C и A с T. В этом случае говорят о сигнале, основанном на вторичной структуре; при этом сами исходные последовательности называют первичными структурами. Поиск такого сигнала, т.е. в конечном счёте построение множественного выравнивания, о котором говорится в [1], представляет бóльшую трудность. Предлагается алгоритм восстановления эволюции такого сигнала. Алгоритм предполагает известным филогенетическое дерево видов и основан на предположении о консервативности вторичной структуры у рассматриваемых сигналов. Алгоритм получает на входе первичную структуру сигнала во всех листьях дерева и выдаёт первичную и вторичную структуры сигнала во всех вершинах дерева. Одновременно алгоритм строит множественное выравнивание современных (в листьях) сайтов сигнала с учётом его вторичной структуры вместе с сайтами, построенными во всех вершинах дерева. В биоинформатике исходные последовательности являются регуляторными участками генов, поэтому сигнал часто называют регуляторным сигналом. Получены результаты успешного тестирования алгоритма на трёх основных типах регуляции у бактерий: классической аттенюаторной, T-боксовой и на основе РНКовых переключателей. Далее рассматривается восстановление эволюции промотора, см. [1].

Описание алгоритма. Этап 1. Каждому листу дерева приписана одна последовательность распределений (т.е. наборов из пяти чисел – частот пяти возможных символов: A, C, G, T и ещё символа делеции), возникшая в ходе вычислений на предыдущей итерации (вначале она соответствует приписанной листу современной последовательности). Движемся от листьев к корню. Пусть двум сыновьям v_1 и v_2 вершины v уже приписаны последовательности распределений σ_1 и

σ_2. Алгоритм должен определить последовательность распределений σ в v. Для этого он сначала выравнивает вторичные структуры этих последовательностей, а затем их первичные структуры с учётом полученной общей вторичной структуры. Теперь для каждой позиции i в качестве распределения σ_i берём взвешенное среднее распределений σ_{1i} и σ_{2i} с весами, которые определяются отношением длин двух рёбер из v в v_1 и в v_2. Затем бывшие σ_1 и σ_2 в v_1 и v_2 заменяем на те, которые получены в результате выравнивания. После того, как дойдём до корня дерева, переходим к этапу 2 алгоритма и потом возвращаемся к этапу 1. Для каждой позиции полученных на этапе 1 изменённых нуклеотидных последовательностей (т.е. со вставленными на этапе 1 делециями) выполняем следующее. Каждой вершине v эволюционного дерева сопоставляется набор $\delta(v)$ частот пяти возможных символов в рассматриваемой позиции. На этапе 2 эти значения считаются переменными, и по всем по ним производится минимизация (поиск ближайшего локального минимума) для описанного ниже функционала F. Пусть ρ – это некоторая мера близости между векторами (в простейшем случае – сумма квадратов разностей компонент); для каждого листа v через $\sigma(v)$ обозначим константный вектор, в котором символу измененной нуклеотидной последовательности, приписанной этому листу, соответствует 1, а остальные элементы – 0; через e_H и e_K обозначим концы ребра e. Определим функционал F следующей формулой (первое суммирование происходит по всем рёбрам дерева, второе – по всем его листьям): $F = \sum_e \rho(\delta(e_н), \delta(e_к)) \cdot w(e) + \sum_v \rho(\delta(v), \sigma(v)) \cdot w(v)$. Здесь $w(e)$, $w(v)$ – весовые коэффициенты. Накладываются естественные ограничения: все переменные неотрицательны и в каждой вершине сумма соответствующих пяти переменных равна единице. В упомянутом простейшем случае функционал квадратичный с единственной точкой минимума, так что эта точка легко находится методом квадратичного программирования. Итерации заканчиваются, когда качество множественного выравнивания достигает максимума.

Результат применения этого алгоритма на примере RFN-регуляции экспрессии генов биосинтеза и транспорта рибофлавина у эубактерий. RFN-структура представляется как состоящая из спирали-черенка и четырёх спиралей в его петле, занумерованных по часовой стрелке (спирали 1, 2, 3, 4). Наш алгоритм выдаёт множественное выравнивание, приведённое на сайте http://lab6.iitp.ru (пункт 9). На нем

выделена искомая RFN-структура: лиловым цветом выделен черенок, жёлтым цветом – первая и третья спирали (для сравнения в современных последовательностях она же выделена зелёным), бирюзовым – вторая и четвёртая спирали, серым – вариабельные спирали. Две спирали – первая и вторая – могут замениться на одну спираль; две другие спирали – третья и четвертая – также могут замениться на одну спираль, эти альтернативные спирали показаны подчёркиванием. Отметим, что консервативные нуклеотиды, характерные для RFN-структуры, в основном выровнялись по столбцам. В предке 19 четвёртая спираль имеет альтернативу, показанную красными буквами, которая продолжается в потомках. На этом сайте приведены и другие примеры применения нашего алгоритма. Чтобы сравнить результаты нашего алгоритма с результатами других стандартных алгоритмов, мы применяли к тем же данным известные программы такие, как PAML, PAUP. Если на вход подавались лишь исходные первичные структуры (без выравнивания), то никакая из этих программ не могла реконструировать предковые регуляторные элементы того типа, который был представлен в листьях. Если на вход подавалось и выравнивание, выполненное с учётом фактически известной вторичной структуры, то результат зависел от используемой программы. PAML не смогла предсказать вторичную структуру требуемого типа в предковых последовательностях. PAUP смогла предсказать такую структуру, однако эта структура не была консервативной вдоль рёбер дерева.

Реконструкция эволюции промотора. Рассматривается ген *rps20* у красных и криптофитовых водорослей. Полученное множественное выравнивание приведено на указанном сайте. В корне дерева имеется бактериальный промотор TTGTCG.17н.TATAAT с TG-расширением, который сохранился в аутгруппе (цианобактерия) и в предке красных и криптофитивых водорослей. При переходе к предку криптофитовых водорослей произошло значительное изменение: промотор принял вид CTTATT.17-18н.TATAAT с TG-расширением, которое сохраняется по всему дереву. При переходе к красным водорослям шестая позиция левого («-35го») бокса сменилась на T, а в остальном произошло несколько замен, вставок и делеций.

СПИСОК ЛИТЕРАТУРЫ

1. *Селиверстов А.В., Любецкий В.А.* Алгоритмы для поиска много боксовых сигналов и их применение // Труды 51-й научной конференции МФТИ «Современные проблемы фундаментальных и прикладных наук». – 2008. (в данном сборнике)