УДК 519.6, 577.214

Селиверстов А.В. [1], *Любецкий В.А.* [1]

[1] Институт проблем передачи информации РАН

Алгоритмы для поиска много боксовых сигналов и их применение

Постановка задачи. Дан *набор* дискретных последовательностей в фиксированном алфавите, например, в алфавите из четырех букв $\{A,C,G,T\}$. В эти последовательности разрешается добавить произвольное число знаков «-» пробела, так что новые последовательности получаются одинаковой длины; в отличие от исходных они называются «выравненными». Результат такого выравнивания удобно представлять матрицей, строки которой – выравненные последовательности. Ниже говорится о столбцах этой матрицы. С точки зрения одного из возможных приложений наших алгоритмов эти буквы удобно называть «нуклеотидами». Сайт в последовательности из этого набора определяется как сочетание нескольких слов («боксов») с определёнными длинами, определённым образом расположенных относительно друг друга и имеющих в определённых позициях определённые нуклеотиды. Упомянутая «определённость» понимается как интервально заданная, т.е. как один из некоторого списка допустимых вариантов; иными словами, – как условие, выполняемое с некоторой достоверностью. Задача состоит в поиске сайта в каждой последовательности из данного набора, так чтобы упомянутая матрица имела как можно более консервативные столбцы; и особенно, столбцы, в которых располагаются найденные сайты. Проще говоря, это значит, что почти одни и те же сайты должны располагаться в матрице друг под другом. «Консервативность» понимается как возможно большее превалирование одной из букв в столбце. Математически задача состоит в максимизации некоторого описанного выше словами функционала, аргументом которого является упомянутая матрица, а значение функционала называется «качеством выравнивания». Мы не выписываем здесь этот сложный функционал. Так полученный набор сайтов называется ***сигналом***, а сама матрица – ***множественным выравниванием***.

Описание алгоритмов. Для сигнала из двух боксов $\langle w', w''\rangle$ с длинами l', l'' и неизвестным буквенным составом, находящихся на расстоянии $d \in [d_{\min}, d_{\max}]$,

функционал качества сигнала описывает среднюю величину сходства сайтов сигнала:

$$Q = \frac{1}{P(P-1)} \sum_{k=1}^{n} q_k \to \max , \qquad (1)$$

где q_k – качество сайта $\langle w'_k, w''_k \rangle$, найденного в k-й последовательности, и

$$q_k = \sum_{\substack{i=1 \\ i \neq k}}^{n} \left(l' + l'' - H(w'_i, w'_k) - H(w''_i, w''_k) - D(d_i) - D(d_k) \right), \qquad (2)$$

а P – число найденных (т.е. непустых) сайтов во всех последовательностях («мощность» сигнала). Функция $D(d)$ в (2) определяет величину штрафа за отклонение расстояния между боксами сигнала от некоторого «наилучшего» значения в интервале от $d_{\min}$ до $d_{\max}$. Вид этой функции является параметром программы; в частности, она может тождественно равняться 0. Алгоритм строится по оптимизационно-комбинаторной схеме, т.е. сочетает в себе направленный частичный перебор с поиском локального квазиоптимума. Это достигается путём поочерёдного применения двух этапов алгоритма: «расстановки» и «сборки».

Программная реализация алгоритма изначально ориентирована на параллельную вычислительную установку с межпроцессорным обменом информацией средствами протокола MPI. Число процессоров кластера может быть любым; программа в состоянии задействовать все доступные процессоры; при этом общее время счёта снижается за счет распараллеливания приблизительно в $s-1$ раз, где s – число процессоров.

Второй алгоритм. Изложим алгоритм для случая бинарного филогенетического дерева видов, хотя он рассчитан и на случай политомического дерева. Сначала алгоритм работает от листьев к корню. Две последовательности распределений для пары сыновей некоторой вершины выравниваются, и их отцу приписывается полусумма выравненных последовательностей. Если для корня дерева таким образом построена последовательность, то выполняется обратный ход алгоритма, при котором пробелы, вставленные в последовательности у прикорневых вершин, опускаются вниз дерева вплоть до его листьев. Затем то же самое делается с пробелами, вставленными при выравнивании на предыдущем уровне дерева (третьем, считая от корня), и т.д. вплоть до уровня, расположенного непосредственно

над листьями дерева. Последовательности, вообще говоря, с многочисленными пробелами, полученные таким образом в листьях дерева, – *искомое множественное выравнивание* и результат работы алгоритма, при данном бинарном дереве.

Само парное выравнивание выполняется очень быстро: например, алгоритм находит множественное выравнивание для 16 последовательностей с длиной 120–223 нуклеотида за менее, чем 1 сек на Pentium-4 PC. Быстродействие алгоритма остаётся достаточно высоким и в случае политомического исходного дерева. Например, для дерева с 12 листьями, в котором у корня имеется 5 сыновей и ещё одна политомическая вершина является родительской для 4 листьев, построение множественного выравнивания последовательностей длиной 115-229 нт заняло 193 с на Pentium-4 PC (3 GHz). При этом было опробовано 1575 различных по топологии деревьев, и одинаковое качество выравнивания достигается только для трёх из них.

Результат работы алгоритмов на одной биоинформатической задаче. Алгоритмы применялись для поиска потенциальных PEP-промоторов перед всеми белок-кодирующими генами пластид из растений и водорослей. В результате перед геном *psbA* найден потенциальный промотор, восходящий по крайней мере к предку Streptophyta. Перед геном *psbB* найден потенциальный промотор, восходящий к предку Streptophyta, который исчез у всех мохообразных. Перед геном *psbE* найден промотор, восходящий к предку Streptophytina, который значительно изменился у водоросли *Chara*. Перед генами *rbcL* и *psaA* найдены промоторы, восходящий к предку Streptophytina. Перед геном *rps20* найден потенциальный промотор, восходящий к предку хлоропластов красных и криптофитовых водорослей. Аналогичные результаты получены для промоторов генов *psbN* и *ndhF*, которые испытали более сложную эволюцию. Предложены некоторые регуляторные механизмы, связанные с этими генами и промоторами. Метод показал свою пригодность для поиска древних промоторов, относящихся сразу к большинству видов из Streptophyta или сразу ко всем красным и криптофитовым водорослям, и т.п. В транскрибируемой нетранслируемой области гена *psbB* найден консервативный у всех Streptophyta регуляторный участок, позиционно связанный с расположенным выше сайтом процессинга мРНК. В области промотора гена *rps20* предположено присутствие репрессии транскрипции гена *rps20* и активации гена *glnB*.