1) generalization of all the conservative regions (they cannot be present in all the aligning sequences).

2) optimal arrangement of these regions using two criteria — maximization of the total power of the conservative regions and minimization of the total number of spaces.

This algorithm has at least two advantages over traditional algorithms (such as Needleman-Wunsch's one) : no penalty for insertion / deletion; not subsequent pair aligning procedure. The efficiency of the algorithm is shown at model example.

СПИСОК ЛИТЕРАТУРЫ

1. *Needleman S. B., Wunch C. D.* A general method applicable to the search for similarities in the amino acid sequence of two proteins // J. Mol. Biol.— 1970.— **48**, N 2.— P. 443—453.
2. *Gotoh O.* Alignment of three biological sequences with an efficient traseback procedure // J. Theor. Biol.— 1986.— **121**, N 1.— P. 327—337.
3. *Sobel E., Marttinez H. M.* A multiple sequence alignment program // Nucl. Acids Res.— 1986.— **14**, N 2.— P. 363—374.
4. *Bacon D. J., Anderson W. J.* Multiple sequence alignment // J. Mol. Biol.— 1986.— **191**, N 1.— P. 153—161.
5. *Леонтович А. М., Бродский Л. И., Горбаленя А. Е.* Построение полной карты локального сходства двух полимеров (программа DotHelix пакета GenBee // Биополимеры и клетка.— 1990.— **6**, № 6.— С. 14—21.
6. *Dayhoff M. O., Barker W. C., Hunt L. T.* Establishing homologies in protein sequences // Meth. Enzymol.— 1983.— **91**.— P. 524—545.

Науч.-произв. кооператив «Комби», Москва
Межфакультет. н.-и. лаб. им. А. Н. Белозерского, МГУ

Получено 28.06.90

УДК 576.315.42

В. А. Шепелев

АЛГОРИТМ УСКОРЕННОГО ПОСТРОЕНИЯ ТОЧЕЧНЫХ МАТРИЦ ГОМОЛОГИИ

Метод анализа гомологичных участков с помощью точечных матриц гомологии заключается в нахождении и отображении на прямоугольной матрице общих для двух последовательностей слов, в которых совпадает определенное количество букв. Предложен алгоритм ускоренного построения таких матриц с различными параметрами фильтрации.

Введение. Метод исследования гомологичных участков с использованием точечных матриц гомологии [1] состоит в том, чтобы на прямоугольной матрице найти общие для двух последовательностей слова, то есть подпоследовательности длиной W, в которых совпадает не менее M букв. Параметры W и M называются параметрами фильтрации, а величина W — размером окна. Такое построение обладает большой наглядностью, так как гомологичные участки выявляются в виде диагональных линий. При $W=M=1$ матрица для двух случайных последовательностей оказывается забитой случайными (с вероятностью около 1/4) точками, отмечающими совпадение отдельных букв. Вообще число точек на матрице для двух случайных последовательностей можно оценить по формуле

$$Z = P \cdot L_1 \cdot L_2,$$

где

$$P = \sum_{k=M}^{W} C_w^k p^k (1-p)^{w-k}.$$

© В. А. ШЕПЕЛЕВ, 1991

p — вероятность появления данной буквы; L_1 и L_2 — длины последовательностей; C_W^k — биномиальные коэффициенты. Величина W обычно выбирается в диапазоне 7—60, а величина M — так, чтобы P было равно 10^{-4}—10^{-5}. При сравнении последовательностей необходимо выполнить расчеты с разными значениями параметров фильтрации, так как даже при фиксированном P, но различных W точечные матрицы гомологии могут существенно отличаться. Трудоемкость построения одной такой матрицы T оценивается как $k_1L_1L_2W$, где k_1 — постоянная для данного алгоритма. Для гена среднего размера $T \sim 10^7$—10^9 шагов, или от десятков секунд до десятков минут. Таким образом, уменьшение времени счета является актуальной задачей. В работе [2] предложен алгоритм (назовем его алгоритмом 0) с обходом матрицы по диагоналям и подсчетом входящих в окно и выходящих из окна нуклеотидов. Трудоемкость этого алгоритма $k_2L_1L_2$. Иногда, особенно для анализа аминокислотных последовательностей, используется следующее обобщение. Совпадение букв a_i и a_j берется с весом q_{ij}, M — пороговый уровень для общего слова. В настоящей работе предложен алгоритм ускоренного построения точечных матриц гомологии для случая единичной матрицы $\{q_{ij}\}$, в котором удалось несколько уменьшить коэффициент k_2.

Методы. Описание алгоритма (алгоритм 1). При построении матрицы, образованной последовательностями A(1) ... A(L_1) и A(1) ... A(L_2), ее обход ведется вдоль диагоналей, параллельных главной диагонали. На первом этапе строится вспомогательный одномерный массив R(i) следующим образом. На каждом шаге вдоль диагонали в R(i) записывается величина j, если A(j)=A(k), в противном случае происходит только сдвиг по диагонали на одну букву. На втором этапе последовательно просматривается массив R(i). Если R=W—1—R(i+M—1)+ +R(i)>0, то на данной диагонали номерам букв от R(i)—R до R(i) соответствуют искомые общие слова.

Алгоритмы 0 и 1 были реализованы на языке ассемблера для ПЭВМ Искра-226 в виде одного блока каждый. Ниже приведена существенная часть блока для алгоритма 1.

```
2153 % — — — Цикл образования вспомогательного буфера R(i) — — —
2154 % DIAGO.+Б1*Б12
2155 % LOOPPOS.+(БТ)Б4*Р11, +(БТ)Б5*Р12 Извлечение очередных букв
2156 % +Р11—Р12, (УП40)SLIP Сравнение букв
2157 % +Б4—Б2, *(СЛ)Б12, +Б12+2* Запись R(i) в буфер
2158 % SLIP. +Б4+1*, +Б5+1*, +Р14+1* Шаг вдоль диагонали
2159 % (УП40)LOOPPOS, +Б12*Р9
2160 % (ВВ)
2161 % — — — — —
2162 % — — — Цикл обработки буфера R(1) для 0—3 окон — — —
2163 % DIAGI. +Б1 Б12, Р10
2164 % +Б12+Р3, *Б11, +Б11—Р9, (УП31)ESAVE
2165 % (КОН)О*Р3, В; 177+377 Р3=М—1
2166 % +Р8—Р3, (УП30)RR, —Р8 Р4=W—1
2167 % RR. —Р6, +1, *Р11, +Р11+1, *Р12
2168 % LOOP. +(СЛ)Б11*РО
2169 % +(СЛ)Б12
2170 % L. +Р4—РО* Вычисление R
2171 % (УП30)NEXTP
2172 % — — — Запись точки матрицы в буфере матрицы — — —
. . . .
2180 % NEXTP. +Б12+2*, +Б11+2*, +Б11—Р9, (УП30)LOOP
2181 % ESAVE. +Р10, (УП41)EDIAGI
```

```
2182 % (ПВ)SAVE, (ПВ)RESAVE
2183 % EDIAGI. (ВВ)
  . . .
```

Блоки были вставлены в программу Q>REVN пакета SEQBUS [3]. В блоке одновременно проводятся вычисления для четырех наборов параметров фильтрации. Особенностью реализации является также способ хранения результатов поиска. Запоминаются не отдельные общие слова, а целые диагональные отрезки из общих слов.

Результаты и обсуждение. Очевидно, что алгоритм 1 имеет ту же зависимость трудоемкости от длины последовательности, что и алгоритм 0. Уменьшение коэффициента k_2 достигается в основном за счет того, что на каждом шаге извлечение букв и их сравнение происходит один раз вместо двух. Хронометрирование показало, что выигрыш во времени составляет 6,8 раза для одновременного просчета четырех окон размером 7, 15, 31 и 63 нуклеотида. При этих условиях и быстродействии процессора (300 тыс. операций в секунду) время счета в секундах для алгоритма 1 можно оценить по формуле $T=0{,}0000391 \cdot L_1 \cdot L_2$. В программном блоке номер диагонали размещен в одном 16-разрядном регистре, откуда следует предельный размер анализируемой последовательности около 32 тыс. нуклеотидных пар.

Резюме

Метод аналізу гомологічних ділянок за допомогою точкових матриць гомології складається із знаходження відображення на прямокутній матриці загальних для двох послідовностей слів, цебто послідовностей довжиною W, в яких співпадають не менше M літер. Запропоновано алгоритм прискореного створення таких матриць з різними параметрами фільтрації.

Summary

The method for analysis of homologous regions using dot matrices consists in finding and graphical representation of words common for two sequences on a rectangular matrix, i. e. subsequencies of length W with at least M coinciding letters. The algorithm is suggested for accelerated calculation of dot matrices with the different filtration parameters. The time saving equals 6.8 times for parallel calculation of four windows.

СПИСОК ЛИТЕРАТУРЫ

1. *Gibbs A. J., McIntyre G. A.* The diagram, a method for comparing sequences, its use with amino acid and nucleotide sequences // Eur. J. Biochem.— 1970.— **16**, N 1.— P. 1—11.
2. *Staden R.* An interactive graphics program for comparing and aligning nucleic acid and amino acid sequences // Nucl. Acids Res.— 1982.— **10**, N 9.— P. 2951—2961.
3. *Голованов Е. И., Шепелев В. А., Александров А. А.* Система программ для исследования нуклеотидных последовательностей на микро-ЭВМ Искра-226 // Теорет. исследования и банки данных по молекуляр. биологии и генетике.— Новосибирск, 1986.— С. 59—60.

Ин-т молекуляр. генетики АН СССР, Москва

Получено 29.05.90