УДК 577.112

А. М. Леонтович, Л. И. Бродский, А. Е. Горбаленя

ПОСТРОЕНИЕ ПОЛНОЙ КАРТЫ ЛОКАЛЬНОГО СХОДСТВА ДВУХ БИОПОЛИМЕРОВ (ПРОГРАММА DOTHELIX ПАКЕТА GENBEE)

В работе предложен новый алгоритм построения полной карты локального сходства двух биополимеров, значительно более быстрый по сравнению с известным алгоритмом Альтшуля — Эриксона. Наш алгоритм реализован в программе DotHelix из пакета GenBee. Эффективность алгоритма по сравнению с традиционным методом Стадена и работа программы DotHelix проиллюстрированы на примере сопоставления полипротеинов двух штаммов вируса полиомиелита. Кратко обсуждены возможные области применения предлагаемой программы.

Введение. В ходе анализа первичных структур биополимеров (полипептидов и полинуклеотидов) одна из наиболее часто решаемых задач — сравнение двух биополимеров (линейных последовательностей мономеров) между собой. Можно выделить два подхода к решению этой задачи: в одном из них последовательности мономеров выравниваются друг относительно друга, в другом — строится так называемая «карта локального сходства» (Dot Matrix), с помощью которой в биополимерах выявляют пары сходных фрагментов. К принципиальным преимуществам второго подхода следует отнести отсутствие необходимости решать проблему оценки штрафа за введение пробелов в выравниваемые последовательности, а также возможность выявления повторов в структуре биополимеров. Вместе с тем наиболее распространенная среди исследователей программа «DIAGON», написанная Стаденом [1], либо ее модификации, с помощью которых строят карты локального сходства, несвободны от существенных недостатков. Они обусловлены наличием двух параметров сравнения, значения которых устанавливаются заранее, — размера сравниваемых фрагментов (так называемого размера «окна сравнения») и уровня (величины) сходства фрагментов. На конечном результате наиболее драматическим образом сказывается априорное определение величины «окна сравнивания». В итоге наблюдаются следующие искажения при анализе биополимеров: 1) выявляются не все пары сходных фрагментов, длина которых отличается от размера «окна сравнения»; 2) границы участков сходства определяются неточно и вследствие этого — 3) не определяется максимально возможный уровень сходства пар фрагментов.

Сравнительно недавно Альтшуль и Эриксон [2] предложили новый подход к построению карты локального сходства двух биополимеров, свободный от недостатков метода Стадена. В результате программной реализации их алгоритма удается получать так называемые полные карты локального сходства двух биополимеров с идентификацией всех пар сходных фрагментов вне зависимости от их длины и в их «естественных» границах, в которых между фрагментами наблюдается наибольшее сходство. Фактически в основу алгоритма было положено иерархически организованное сравнение двух белков по методу Стадена с использованием «окон сравнения» во всем возможном диапазоне их длин. Такой способ решения задачи привел к резкому увеличению времени счета на ЭВМ, что не позволяет использовать алгоритм в практической работе на персональных компьютерах.

В настоящей работе предлагается существенно более экономичная процедура построения полной карты локального сходства двух биополимеров, которая положена в основу программы DotHelix для компьютеров типа IBM PC. Работа программы рассмотрена на нескольких примерах, иллюстрирующих преимущества развиваемого подхода.

Метод. Определения. Оценку сходства пар фрагментов сравниваемых биополимеров проводили путем построения полной карты локального сходства двух биополимеров. Под такой картой мы подразумеваем прямоугольную матрицу, по двум сторонам которой отложены сопоставляемые последовательности длины m и n из мономеров, составляющих эти биополимеры. Каждая клетка (i, j) матрицы (карты) соответствует сопоставлению двух мономеров из сравниваемых биополимеров, которые являются проекциями этой клетки на стороны карты; здесь i и j — позиции мономеров в двух биополимерах. Начальным мономерам отвечает левый верхний угол карты. Любой паре сравниваемых фрагментов одинаковой длины L из сопоставляемых биополимеров соответствует диагональный отрезок карты той же длины, составленный из клеток (i, j), $(i+1, j+1)$,..., $(i+L-1, j+L-1)$, где $i+1$,..., $i+L-1$ и j, $j+1$,..., $j+L-1$ — позиции мономеров в двух сравниваемых фрагментах соответственно. Все такие диагональные отрезки располагаются на одной из $n+m-1$ диагоналей карты. Каждая диагональ карты отвечает определенному регистру сравнения двух полимеров, который характеризуется фиксированным сдвигом одной последовательности относительно другой. Таким образом, число возможных регистров сравнения также равно $n+m-1$.

В ходе построения карты в каждую клетку (i, j) записывается число a_{ij}, характеризующее сходство между парой соответствующих мономеров. Значения этих чисел берутся из специальных матриц сходства биологических мономеров (аминокислот или нуклеотидов). Например, при сравнении белков наиболее употребительна матрица Дейхофф [3], отражающая сходство между парами аминокислот. Таким образом, в заполненном виде карта состоит из $m \cdot n$ чисел a_{ij}, характеризующих сходство каждой пары мономеров из двух полимеров.

Пусть мы хотим выяснить, каково сходство между парой фрагментов сравниваемых полимеров. Как было изложено, этой паре отвечает диагональный отрезок карты, состоящий из L клеток, заполненных числами a_{ij}. В качестве меры сходства между этими фрагментами берется нормированная сумма этих чисел a_{ij} [2], то есть величина

$$A = \frac{S - M \cdot L}{\sqrt{D \cdot L}}, \qquad (1)$$

где S — сумма чисел a_{ij}, состоящих в клетках соответствующего диагонального отрезка; L — длина сравниваемых фрагментов (она равна длине отрезка); M и D — среднее значение и дисперсия всех $m \cdot n$ чисел a_{ij} в полностью заполненной карте соответственно. Вычисляемую таким образом меру сходства двух фрагментов будем называть также мощностью соответствующего диагонального отрезка. Как следует из формулы (1), мощность измеряется в единицах стандартного отклонения (SD).

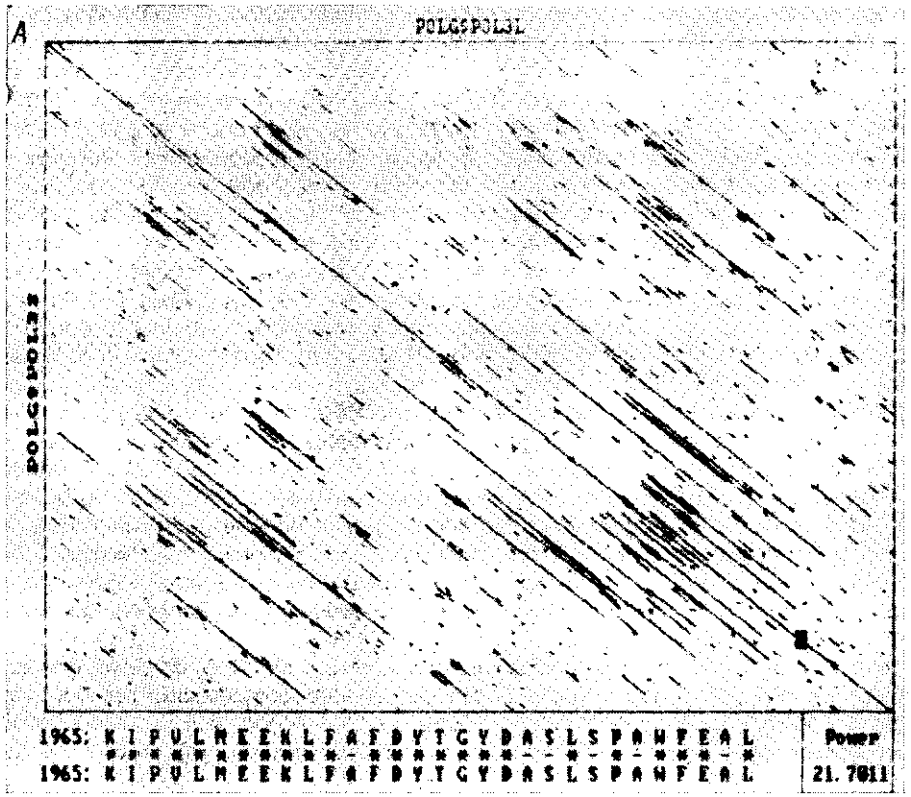
Более объективно, хотя более трудоемко, измерять сходство фрагментов, исходя не из мощности, а из вероятности случайного появления такого сходства в паре случайных бернуллиевских независимых между собой последовательностей той же длины, что и сравниваемые биополимеры, и с теми же частотами букв (мономеров); чем меньше эта вероятность (P), тем больше сходство. Например, в качестве меры сходства можно взять величину $\lg P$ [1, 2]. Для таких бернуллиевских последовательностей мера сходства сравниваемых фрагментов, измеряемая мощностью, является случайной величиной, и при увеличении длины L фрагментов ее функция распределения приближается к стандартной функции нормального распределения $N(0,1)$. Отсюда следует, что при больших L имеет место соответствие между мощностью A и вероятностью P, не зависящее от L, задаваемое формулой

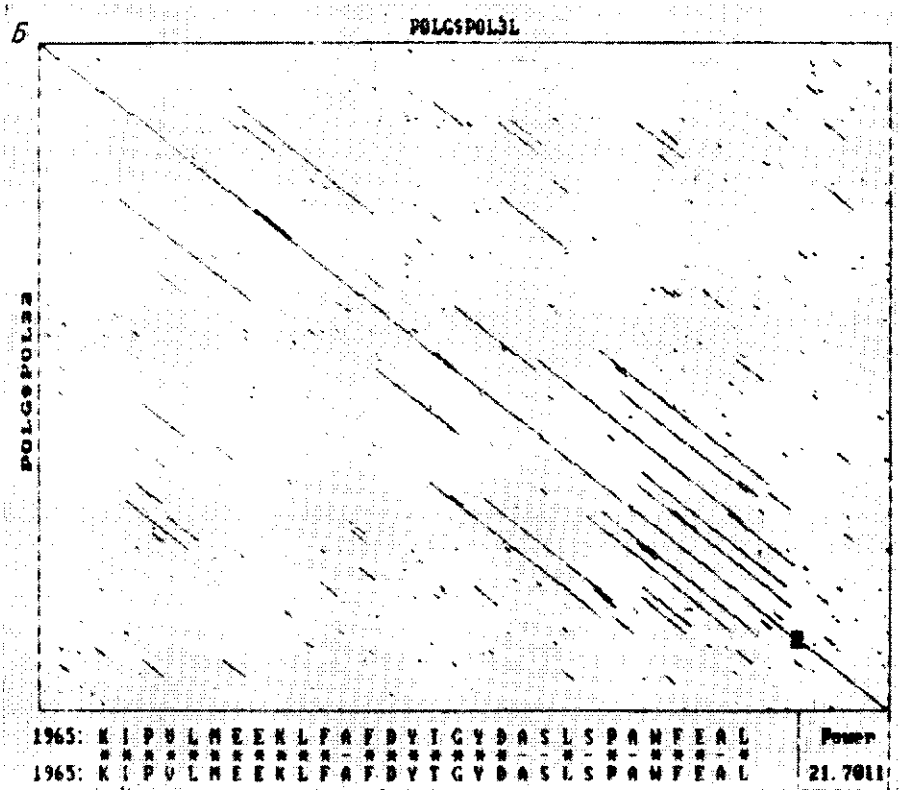
$$P \cong 1/\sqrt{2\pi} \int_{A}^{\infty} e^{-x^2/2} dx; \qquad (2)$$

это соотношение с увеличением длины L делается все более точным.

Алгоритм. В нашем алгоритме, как и у Альтшуля — Эриксона [2], на каждой диагонали находится следующая система отрезков: самый мощный отрезок; наиболее мощный отрезок среди отрезков, не пересекающихся с первым; наиболее мощный отрезок из непересекающихся с первыми двумя и т. д. до тех пор, пока мощности получаемых отрезков остаются больше некоторого порога. Нахождение такой системы отрезков для каждой диагонали происходит в описываемом алгоритме независимо от других диагоналей.

Процесс поиска указанной выше системы отрезков для каждой диагонали основан на следующих соображениях. Обозначим через A_{rs} мощность диагонального отрезка $[r, s]$, где r и s — начальная и конечная позиции этого отрезка в этой диагонали; в частности, A_{rr} — мера сходства между парой букв, соответствующих позиции r. Пусть


А
POLG:POL3L
POLG:POL3L
1965: K I P V L M E E K L F A F D Y T G Y D A S L S P A W F E A L
1965: K I P V L M E E K L F A F D Y T G Y D A S L S P A W F E A L
Power
21.7011


Б
POLG:POL3L
POLG:POL3L
1965: K I P V L M E E K L F A F D Y T G Y D A S L S P A W F E A L
1965: K I P V L M E E K L F A F D Y T G Y D A S L S P A W F E A L
Power
21.7011

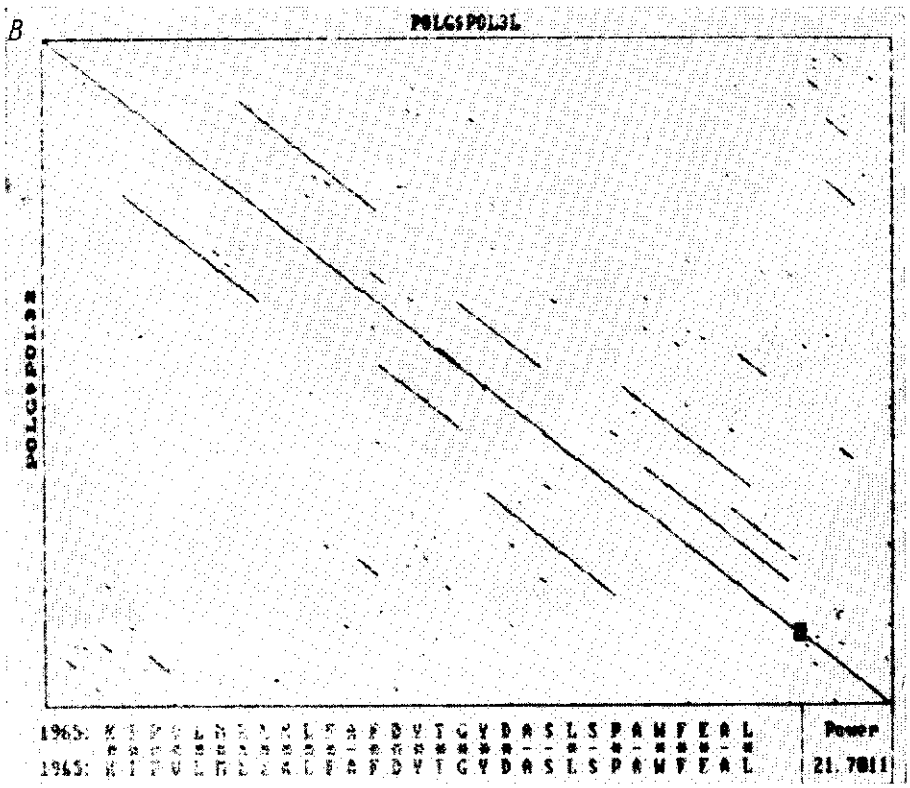


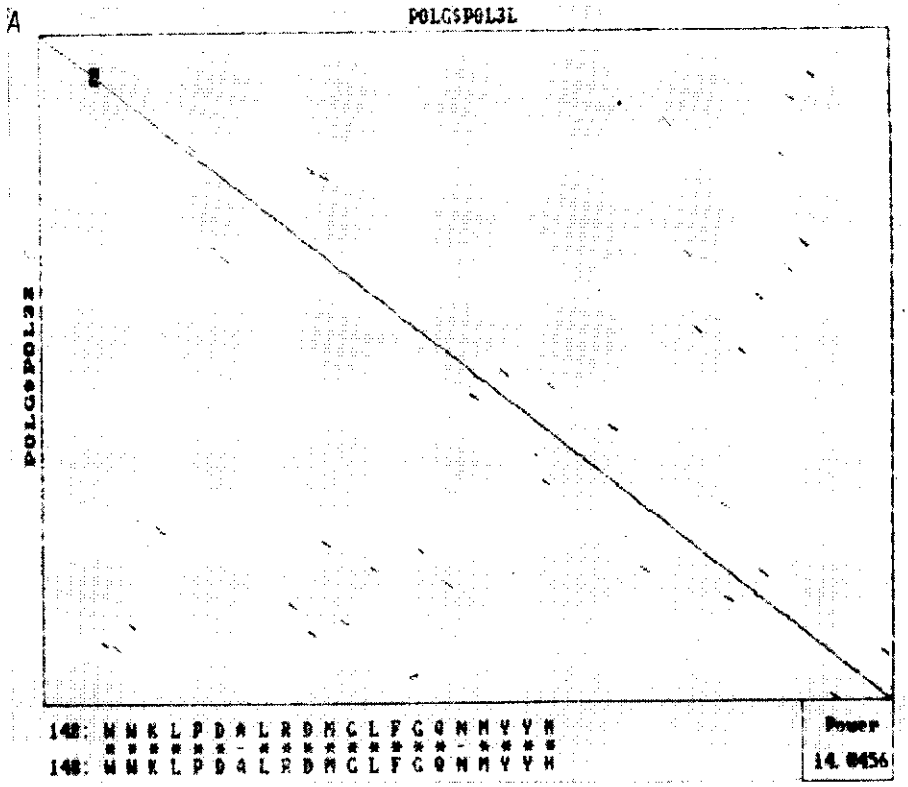
Рис. 1. Полные карты локального сходства полипротеинов двух штаммов полиовируса типа 3, полученные при трех значениях порогового уровня (*а* — 4 SD, *б* — 4,5 SD, *в* — 5,0 SD) в ходе работы программы DotHelix. Белки взяты из банка данных SWISSPROT, версия 10. На картах черным квадратом обозначена позиция диагонального сегмента с наибольшим значением мощности. Под картой сопоставлены последовательности аминокислот наиболее сходных фрагментов и обозначен уровень их сходства

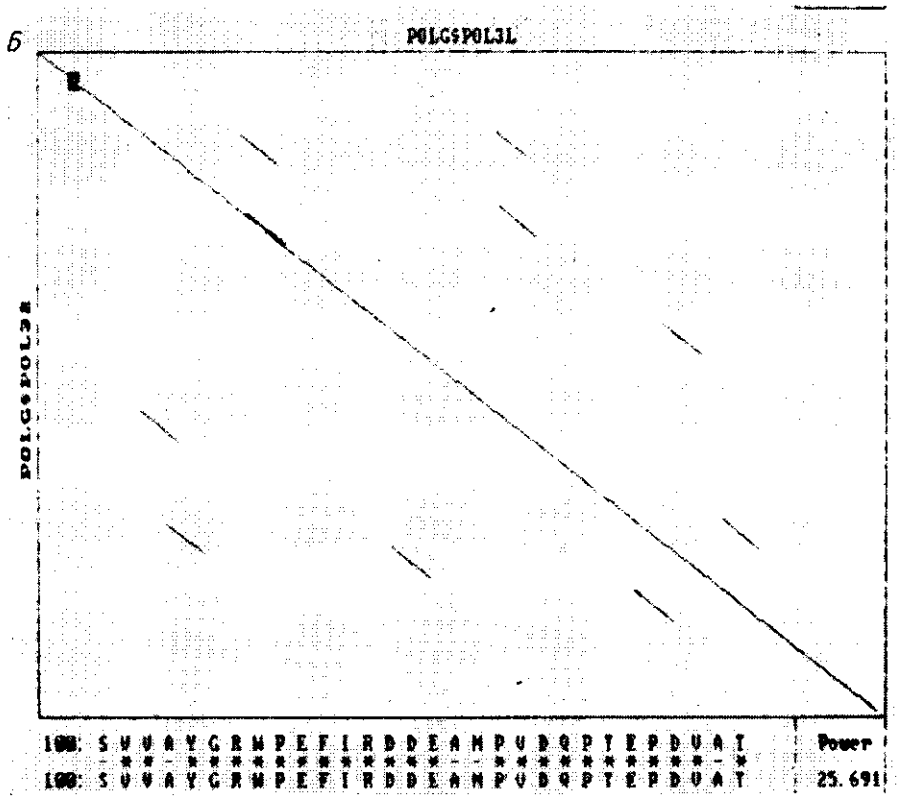
Fig. 1. The complete dot-matrices generated by the program DotHelix for the polyproteins of two strains of poliomyelitis virus type 3 obtained at 3 different values of the cut-off level: (*a*) — 4 SD, (*б*) — 4.5 SD, (*в*) — 5.0 SD. The proteins were from the SWISSPROT database, version 10. Black square marks the position of the diagonal segment with the highest power value. The alignment of the most similar segments and the similarity level (as the number of SD) are shown below the matrix.

имеется некоторый диагональный отрезок $[r_0, s_0]$ («зона просмотра»), и мы хотим найти отрезок максимальной мощности, лежащий внутри «зоны просмотра».

Очевидно, что при $A_{r_0 r_0} \leqslant 0$ отрезок максимальной мощности заведомо лежит строго правее позиции r_0, так что «зона просмотра» может быть уменьшена на одну позицию r_0. Пусть $A_{r_0 r_0} > 0$. Обозначим через s^* минимальное значение s, для которого мощность $A_{r_0 s}$ как функция от s меняет знак плюс на минус. Легко видеть, что отрезок максимальной мощности обязательно лежит либо левее позиции s^*, либо правее ее, но не может содержать внутри себя эту позицию s^*; таким образом, наша старая «зона просмотра» $[r_0, s_0]$ может быть разбита на две новые, меньшие «зоны просмотра»: $[r_0, s^*]$ и $[s^* + 1, s_0]$. Наконец, пусть при всех s $(r_0 \leqslant s \leqslant s_0)$ мощность $A_{r_0 s} > 0$. Пусть s^+ — то значение s, при котором мощность $A_{r_0 s}$ (как функция от s) принимает максимальное значение. Можно доказать, что отрезок максимальной мощности обязательно лежит либо левее, либо правее s^+, так что и в этом случае вся «зона просмотра» может быть разбита на две — $[r_0, s^+]$ и $[s^+ + 1, s_0]$.

Результаты и обсуждение. Описанный выше алгоритм был реализован нами в программе DotHelix, являющейся составной частью пакета GenBee, предназначенного для компьютеров класса IBM PC. При входе в программу, кроме двух последовательностей и матрицы сход-


А
P0LG$P0L3L
P0LG$P0L3L
Power
14.0456


Б
P0LG$P0L3L
P0LG$P0L3L
Power
25.691

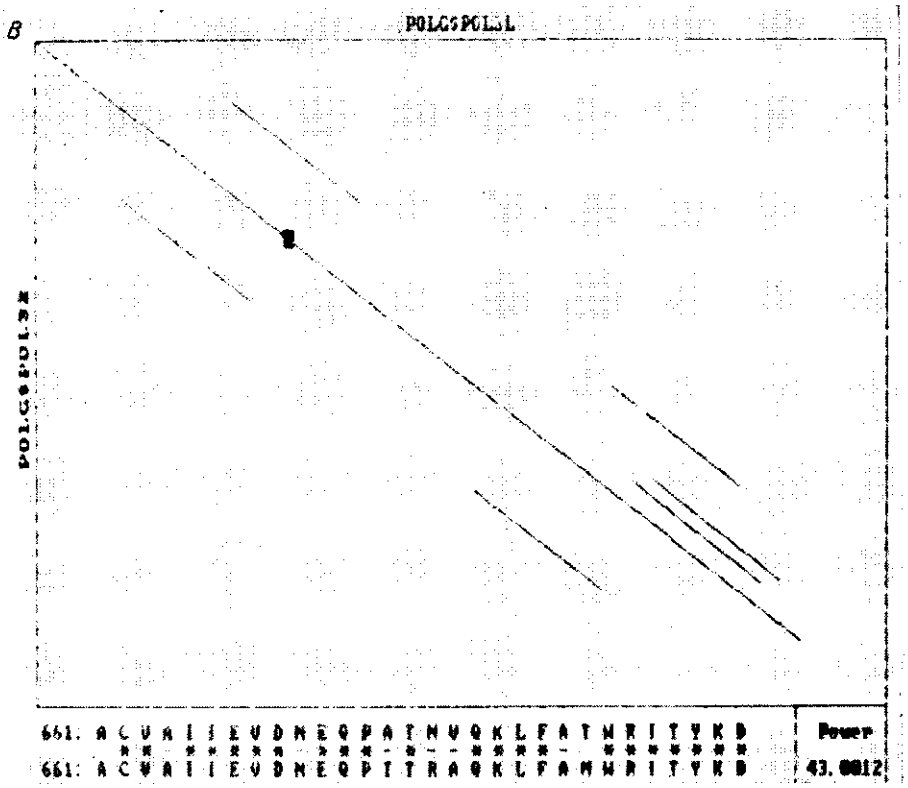


Рис. 2. Карты локального сходства тех же белков, что и на рис. 1, полученные в ходе работы нашей версии программы DIAGON при пороговом уровне сходства 4,5 SD и трех значениях размера «окна сравнения» (*а* — 21, *б* — 99, *в* — 330 аминокислотных остатков). Остальные детали см. подпись к рис. 1

Fig. 2. Dot-matrices for the proteins compared in Fig. 1 obtained by our version of the DIAGON program with the cut-off level of 4.5 SD and three values of the comparison window: 21 (*a*), 99 (*б*), or 330 (*в*) amino acid residues. For further details see the caption to Fig. 1.

ства аминокислот или нуклеотидов, пользователь задает значение порогового уровня для мощности диагональных отрезков, начиная с которого фрагменты последовательностей считаются сходными (run Level). После построения полной карты локального сходства для данного уровня пользователь имеет возможность проанализировать карты для более высоких уровней без повторения расчетов. Для этого требуется только определить интересующий вас порог (Draw Level). С помощью такого приема удается эмпирически подобрать уровень, при котором нужная информация наиболее полно отображается на карте.

На рис. 1 на примере сравнения полипротеинов двух штаммов вируса полиомиелита типа 3 проиллюстрирована работа программы Dot-Helix. Сравнение этих двух гигантских белков-предшественников (длина 2206 аминокислотных остатков) на IBM PC XT с тактовой частотой 8 МГц занимает около 45 мин. Представлены карты локального сходства, полученные при трех значениях порогового уровня. При рассмотрении этих карт видны следующие характерные черты: симметричность получаемых карт, наличие сгущений диагональных отрезков в определенных местах карт (особенно на рис. 1, *а*, *б*), а также присутствие отрезков существенно разной длины на одной и той же карте.

Аналогичный анализ для этих белков, выполненный с помощью собственной версии программы DIAGON, также выявил на картах симметрию и сгущения (рис. 2). Однако на этих картах было значительно меньше диагональных отрезков (для одного порогового уровня) и все они имели сходный, близкий к величине «окна сравнения» размер.

При обоих типах анализа симметричность карт объясняется очень высоким уровнем сходства двух белков, а наличие сгущений диагональных отрезков (особенно в правом нижнем углу карты) — присутствием тандемных повторов различной длины [4]. Наличие большего числа отрезков на картах (рис. 1) свидетельствует, что с помощью программы DotHelix можно более эффективно выявлять сходные фрагменты, чем при традиционном подходе. Действительно, при более детальном, выходящем за рамки настоящей работы, рассмотрении в областях сгущения расположения и размера диагональных отрезков удается в полипротеине полиовируса более полно выявить и охарактеризовать тандемные повторы (данные не приведены).

Следовательно, одна карта локального сходства, полученная по предлагаемому нами способу, содержит значительно больше информации, чем несколько карт локального сходства того же порогового уровня, построенных с помощью программы DIAGON. Что еще более важно — она не может быть получена механической суперпозицией сколь угодно большого числа традиционных карт. Таким образом, выигрыш во времени работы программы DIAGON (работа над картой, представленной на рис. 2, занимала около 7 мин) не может компенсировать потери существенной части информации.

В целом наиболее существенное достоинство развиваемого подхода по сравнению с традиционным состоит в следующем. Во-первых, метод имеет более высокую чувствительность: выявляются все сходные пары фрагментов с заданным уровнем сходства. Во-вторых, метод более точен: находятся точные границы пар сходных фрагментов и вычисляется мощность, характеризующая их сходство. Наиболее зримо эти преимущества проявляются при сравнении биополимеров большой длины и со слабым сходством. Как следствие, в случае сравнения биополимера себя с собой полная карта внутреннего локального сходства может быть успешно использована для обнаружения и анализа периодичностей в первичной структуре. Мы надеемся также, что на базе этой программы можно создавать эффективные методы выравнивания большого числа последовательностей. Один такой метод уже реализован в программе H-Align пакета GenBee (см. статью Бродского и др. в следующем выпуске данного журнала).

Сравним наш алгоритм с алгоритмом Альтшуля — Эриксона [2]. Выделенные в результате их работы пары сходных фрагментов должны совпадать между собой. Однако время работы, по-видимому, резко отличается. Поскольку в алгоритме Альтшуля — Эриксона производится полный перебор сравнений всех возможных пар фрагментов одинаковой длины, то его работа по существу эквивалентна суммарной работе алгоритма Стадена [1] при всех возможных длинах «окон сравнения». Исходя из этого можно оценить время его работы. В вышеприведенном примере сравнения двух полипротеинов время работы этого алгоритма будет приблизительно равно 2206·7 мин = 15 442 мин, что превышает время работы нашего алгоритма (45 мин) примерно в 343 раза.

Время работы алгоритма естественно оценивать через число осуществляемых при этом элементарных операций. Для алгоритма Альтшуля — Эриксона число таких операций при обработке одной диагонали длины N имеет порядок N^2 (см. [2]). К сожалению, для нашего алгоритма точной оценки получить не удалось. Пока возможны только нестрогие, качественные соображения. Они основана на том, что в нашем алгоритме каждая зона просмотра делится, по-видимому, на две примерно одинаковые новые зоны. Отсюда вытекает следующая гипотеза: если в качестве сравниваемых последовательностей взяты случайные, независимые между собой бернуллиевские последовательности, то число таких операций с вероятностью, стремящейся к 1 при $N \to \infty$, имеет порядок $N \cdot \lg_2 N$. (Более слабая гипотеза: $n^{1+o(1)}$.) Как нам кажется, и для биологических последовательностей справедлива аналогичная оценка. (Следует, однако, заметить, что можно построить

искусственные последовательности, для которых подобная оценка неверна; для них число операций имеет порядок N^2.)

Исходя из предложенной гипотезы, отношение времен работы алгоритмов Альтшуля — Эриксона и нашего имеет порядок $N/\lg_2 N$. В рассмотренном выше примере $N=2206$, $N/\lg_2 N=200$. Эта оценка, хотя и меньше таковой (343), полученной из реальных времен работы процедур DotHelix и DIAGON, но имеет тот же порядок.

В заключение укажем на один недостаток предложенной программы DotHelix. Он связан с оценкой сходства между короткими фрагментами. Как было сказано выше, соответствие между мощностью A и вероятностью P, задаваемое формулой (2), при небольших значениях L не вполне точное; как правило, при таких L измерение уровня сходства, основанное на оценке мощности, приводит к завышенным его значениям. В результате наша программа имеет тенденцию к выделению коротких фрагментов в качестве сходных и, наоборот, могут игнорироваться протяженные сходные участки. Для того чтобы смягчить отрицательный эффект, в программе есть возможность не рассматривать пары фрагментов с очень малой длиной L (например, $L < 3$ или $L < 4$). Такое решение — компромиссное. В настоящее время разрабатываются и другие, более эффективные пути решения этой проблемы.

COMPILE OF A COMPLETE MAP OF LOCAL SIMILARITY FOR TWO BIOPOLYMERS (DotHelix PROGRAM of the GenBee PACKAGE)

Leontovich A. M., Brodsky L. I., Gorbalenya A. E.

Coop 2 «ComBee» associated with Soviet-French-Italian
Joint Venture «Interquadro», Moscow;
A. N. Belozersky Laboratory of Molecular Biology and Bioorganic Chemistry,
M. V. Lomonosov State University, Moscow;
Institute of Poliomyelitis and Viral Encephalitides,
Academy of Medical Sciences of the USSR

Summary

A novel algorithm for generation of complete maps of local similarity for two biopolymers which is much faster than the similar Altschul-Erickson algorithm is described. This algorithm was implemented as the DotHelix program within the GenBee package. The algorithm effectiveness as compared to the Staden algorithm and the results obtained with the DotHelix program are illustrated by the comparison of the polyproteins of two strains of poliomyelitis virus type 3. The possible applications of the program are discussed in brief.

СПИСОК ЛИТЕРАТУРЫ

1. *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.
2. *Altschul S. F., Erickson B. W.* A nonlinear measure of subalignment similarity and its significance levels // Bull. Math. Biol.— 1986.— 48, N 5—6.— P. 617—632.
3. *Dayhoff M. O., Barker W. C., Hunt L. T.* Establishing homologies in protein sequences // Meth. Enzymol.— 1983.— **91**.— P. 524—545.
4. *Горбаленя А. Е., Донченко А. П., Блинов В. М.* Возможность общего происхождения полиовирусных белков с различными функциями // Молекуляр. генетика.— 1986.— № 1.— С. 36—41.

Науч.-произв. кооператив «Комби», Москва
Межфакультет. н.-и. лаб. им. А. Н. Белозерского, МГУ
Ин-т полиомиелита и вирус. энцефалитов
АМН СССР, Моск. обл.

Получено 28.05.90