УДК 577.112

Л. И. Бродский, А. Л. Драчев, Р. Л. Татузов, К. М. Чумаков

ПАКЕТ ПРОГРАММ ДЛЯ АНАЛИЗА ПОСЛЕДОВАТЕЛЬНОСТЕЙ БИОПОЛИМЕРОВ: GenBee

В работе описан пакет программ GenBee, предназначенный для анализа биологических последовательностей. Пакет ориентирован главным образом на задачи теоретической молекулярной биологии и совмещает в себе удобный пользовательский интерфейс с развитыми современными алгоритмами (включая оригинальные). Он написан на языке Си и пригоден для работы на компьютерах типа IBM PC.

Сравнительный анализ «биологических текстов», т. е. первичных структур белков и нуклеиновых кислот и выделение из них максимальной информации, является одной из основных задач современной молекулярной биологии. Все вновь секвенированные последовательности вносятся в основные международные банки данных (EMBL, GENBANK, PIR, SWISSPROT) и становятся доступными пользователю на магнитных дисках. Таким образом, важной задачей является создание пакета программ для работы с этими банками данных и отдельными последовательностями.

В настоящее время существует уже довольно много таких пакетов программ как в нашей стране, так и за рубежом. В основном это коммерческие пакеты, ориентированные главным образом на задачи генной инженерии и другие массовые операции, необходимые при экспериментальной работе. Наш пакет, направленный, как правило, на решение задачи теоретической молекулярной биологии, совмещает в себе удобный пользовательский интерфейс с развитыми современными алгоритмами (включая оригинальные).

Пакет GenBee имеет модульную структуру и состоит из следующих модулей:

— быстрый поиск подобий по банку — QUICK SEARCH;

— поиск функциональных сайтов и мотивов — SITE;

— поиск открытых рамок трансляции, предсказание кодирующих областей в ДНК/РНК и трансляция их в белок — PROTMAKE;

— составление полной карты локального сходства двух последовательностей — DOTHELIX;

— множественное выравнивание последовательностей — H-ALIGN;

— расчет матриц сходства последовательностей и построение эволюционных деревьев — TREE;

— построение вторичной структуры РНК — RNA2;

— предсказание вторичной структуры белка PROTEIN2.

Каждый модуль GenBee может обращаться к встроенной базе данных, имеющей собственный формат. Информация в базу либо заносится пользователем вручную, либо переносится специальной программой FileBase из стандартных банков данных EMBL, PIR, SWISSPROT и GenBank. Формат встроенной базы данных пакета рассчитан на максимально сжатое хранение последовательностей, что удобно для пользователей IBM PC, как правило, имеющих твердый диск небольшого объема. Такое сжатие достигается, во-первых, выбрасыванием из стандартных форматов информации вторичной важности и, во-вторых, побитной записью собственно последовательностей.

База данных GenBee организована по иерархическому принципу — ее элементами являются не только последовательности, но и другие подбазы. Одна последовательность может входить в несколько подбаз, однако физически хранится ее единственная копия.

Работа пользователя с подбазами и последовательностями в базе данных GenBee (GenBee Commander) внешне имитирует работу с директориями и файлами в широко распространенном пакете Norton Com-

mander. Пользовательский интерфейс на двух панелях экрана позволяет следить, во-первых, за директориями и файлами операционной системы MS DOS, в которых содержатся наши подбазы, и, во-вторых, входить внутрь подбаз, получая информацию об их содержимом. На нижнем уровне высвечиваются идентификаторы последовательностей. Как и в Norton Commander, можно отметить последовательности, скопировать их в другую подбазу или создать из них новую подбазу, переименовать подбазу и т. п. Кроме того, работа с базами данных в GenBee Commander позволяет осуществлять поиск и выделять последовательности по ключевым словам, содержащимся как в списке ключевых слов, так и в описательной части записей. Из выделенного набора последовательностей можно создать подбазу и повторить поиск (например, по другому ключевому слову).

Для ввода, просмотра и редактирования отдельных последовательностей и их наборов в пакете имеется специальный редактор. Помимо обычных операций работы с текстом он позволяет одновременно вставлять символы в заранее выбранный набор последовательностей, реверсировать последовательности, вручную выравнивать их, экспортировать последовательности из формата хранения в базе данных в текстовый формат и импортировать файлы последовательностей из ASCII кодов во внутренний формат базы.

Модулям Site, DotHelix и H-Align посвящены работы [1, 3, 4]. Кратко опишем другие модули нашего пакета.

Очень распространенной задачей при анализе первичных структур биополимеров является выделение из банка всех тех последовательностей, хотя бы один фрагмент которых достаточно сходен с каким-то фрагментом данной «поисковой» последовательности. Поскольку объем банка велик, необходимо проделывать эту операцию быть может грубо, но быстро. В модуле QUICK SEIRCH поисковую последовательность поочередно сравнивают с последовательностями банка (или его части), при этом сравнение происходит в два этапа.

На первом этапе алгоритмом типа Липмана — Пирсона [5] выделяют те сдвиги обрабатываемой пары последовательностей друг относительно друга, для которых наблюдается достаточно много точных совпадений символов (или пар символов, или троек, в зависимости от желания пользователя). Более того, если для всех сдвигов данной пары пользователей количество точных совпадений ниже ожидаемого для пары случайных последовательностей, данная последовательность банка считается несходной с поисковой и процедура переходит к следующей последовательности банка. Если же для какого-то количества сдвигов порог по точным совпадениям преодолен, то из этого набора сдвигов выделяют несколько лучших (их количество задается из меню) и для них выполняют второй этап поиска.

На втором этапе обработки данного сдвига текущей пары последовательностей происходит выделение наилучшим образом соответствующих друг другу фрагментов, при этом качество соответствия оценивается с учетом матрицы весов за замены остатков. Реально на втором этапе проводится процедура, являющаяся упрощением программы модуля DotHelix [3].

Время счета по этой программе на компьютерах типа IBM PC сравнительно небольшое и измеряется десятками минут. Так, для длины поисковой последовательности из 400 аминокислотных остатков время просмотра всего банка SwissProt, содержащего 12 305 аминокислотных последовательностей, составляет 17 мин на компьютере с процессором Intel 80 286.

Модуль PROTMAKE предназначен для нахождения кодирующих белки регионов в последовательностях ДНК и РНК. Обычно это делается так: в нуклеотидной последовательности находят все открытые рамки считывания во всех шести фазах, т. е. участки без терминирующих кодонов; затем для каждой такой рамки строят функцию (потенциал кодирования), интеграл от которой характеризует вероятность для

данной рамки быть кодирующим регионом. В описываемом пакете такую функцию вычисляют по алгоритму Трифонова [2], хорошо зарекомендовавшему себя в приложениях. Он основан на подсчете доли кодонов типа G-nonG-N. Практически эту долю в каждой открытой рамке подсчитывают для перекрывающихся окон определенного размера (например, размера 20) и график такой величины откладывают вдоль последовательности. Чем выше в данной рамке считывания среднее значение потенциальной функции, тем больше вероятность того, что эта рамка кодирующая. В качестве результата работы модуля, пользователь

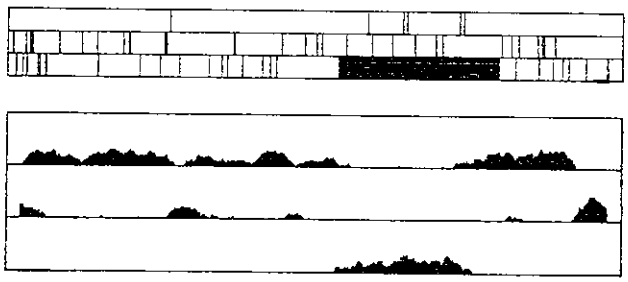

Рис. 1. Графическое представление кодирующего потенциала нуклеотидной последовательности по Трифонову в трех рамках считывания. Вертикальными линиями обозначены терминирующие кодоны

Fig. 1. A graphic representation of Trifonov coding potential of the nucleotide sequence at three reading frames. The vertical lines represent stop codons

видит картину типа той, что изображена на рис. 1. Здесь же, в этом модуле, можно оттранслировать выбранную рамку считывания и сохранить получившийся «белок» в базе данных.

Разумеется, алгоритм Трифонова не является панацеей и не всегда достаточно правильно предсказывает кодирующие белки регионы. В связи с этим в рамках развития пакета мы планируем добавить еще несколько методов предсказания кодирующих белки областей нуклеотидных последовательностей с тем, чтобы решение можно было принимать методом консилиума.

Знание пространственной структуры белка является одним из наиболее важных предварительных условий предсказания его функции. Поскольку реальная кристаллическая структура известна для очень небольшого числа протеинов, теоретическое предсказание вторичной структуры совместно со сравнительным анализом данной последовательности с другими образует необходимый инструмент получения выводов о пространственной структуре данного белка. К сожалению, существующие в настоящее время методы предсказания вторичной структуры обладают невысокой точностью: правильно предсказываются примерно 60 % остатков полипептидной цепи.

Посвященный данной теме модуль Protein2, разработчиком которого является Д. Р. Давыдов, сделан на основе сравнительно простого, но, быть может, наиболее точного алгоритма Гарнье — Робсона [9]. Этот алгоритм принадлежит к группе методов, опирающихся на статистический анализ базы данных о рентгеноструктурном анализе протеиновых структур. Для каждого i-го остатка в полипептидной цепи вычисляют его «склонности» находиться в альфа-спирали, бета-складке или в бета-повороте. Эти «склонности» получают суммированием информации, которую несут остатки от 8-го справа и до 8-го слева от i-го, о вероятности для него быть в данном конформационном положении. Информационные значения берут из специальных таблиц (таблиц сопряженности), полученных для конформаций из базы данных кристаллических структур протеинов. Для каждого остатка предсказывается та конформация, к которой у него наивысшая «склонность».

Модуль TREE позволяет строить филогенетические деревья для набора последовательностей. Эти последовательности предварительно вы-

равнивают, а затем вычисляют матрицу расстояний между ними [6]. В модуле реализованы два метода построения деревьев по вычисленной матрице расстояний: кластерный метод и метод максимального топологического подобия.

В широко известном кластерном методе [6] происходит восстановление хода эволюционного процесса в обратном порядке: от имеющихся последовательностей к предкам. При этом предполагается равномерность накопления мутаций во всех филогенетических линиях, т. е. постоянство скорости эволюции.

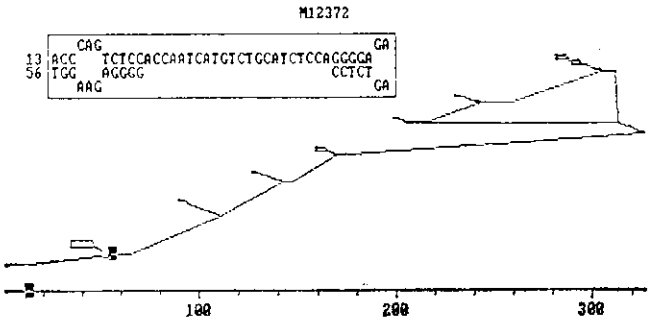

Рис. 2. Графическое представление вторичной структуры РНК. При движении курсора по последовательности (горизонтальная ось), второй курсор движется по изображению вторичной структуры, показывающей топологию свертывания молекулы. Одновременно в рамке вверху можно видеть нуклеотиды соответствующего одно- или двунитчатого участка

Fig. 2. A graphic representation of RNA secondary structure. When the pointer moves along the sequence (horizontal axis), the second pointer moves along the topology of RNA folding. The upper box shows nucleotides and Watson-Crick pairing

В том случае, когда предположение о постоянстве скорости эволюции не выполняется (что бывает, быть может, чаще, чем обратное), более предпочтительным является метод максимального топологического подобия [7]. Он наиболее полно представляет внутреннюю структуру матрицы расстояний между последовательностями из исследуемого набора и исходит из величины топологического отклонения — числа тех наборов из четырех последовательностей (i, j, k, l), для которых выполняются неравенства

$$d_{ij} + d_{kl} < d_{ik} + d_{jl} < d_{il} + d_{jk}$$

в матрице, но не выполняются в построенном дереве или наоборот; здесь d_{ij} — расстояние между последовательностями i и j. Алгоритм стремится построить такое дерево, которое минимизирует эту величину — дерево максимального топологического подобия. Хотя программа обычно не находит глобального минимума, получаемые деревья достаточно разумно аппроксимируют дерево максимального топологического подобия. По сравнению с методами максимума парсимонии [8], этот алгоритм слабочувствителен к малым локальным изменениям исходных данных; кроме того, правильность самого принципа парсимонии (экономии) неочевидна [8].

В модуле RNA2 строят вероятную вторичную структуру РНК по горитму, описанному в [10] (разработчиком этого модуля является А. П. Гультяев). В качестве результата работы модуля выступает список спиральных участков вторичной структуры и графическое изображение расположения двунитчатых и однонитчатых участков (рис. 2).

Пакет GENBEE написан на языке программирования Си и реализован для компьютеров типа IBM PC. За счет совмещения в одном пакете удобного пользовательского интерфейса и развитых алгоритмов пакет GenBee одинаково пригоден как для начинающего, так и для специалиста.

Резюме

В роботі описується пакет програм GenBee, призначений для аналізу біологічних послідовностей. Пакет орієнтований головним чином на задачі теоретичної молекулярної біології і поєднує зручний для користування інтерфейс з розвинутими сучасними алгоритмами (в тому числі оригінальні). Він написаний на мові Сі і пригодний для роботи на комп'ютерах типу IBM PC.

Summary

The package of programs GenBee intended to analyze nucleotide and amino acid sequences is described. This package combines high-level algorithms (including original ones) suitable for advanced theoretical studies with flexibility and user-friendly service typical of commercially available packages. The package is designed for IBM-compatible personal computers. Accordingly it will be useful both for researches and students with practically no background in computer methods, and for theoreticians.

СПИСОК ЛИТЕРАТУРЫ

1. *Кунин Е. В., Чумаков К. М., Горбаленя А. Е.* Метод поиска структурных мотивов в аминокислотных последовательностях. Программа Site пакета GenBee // Биополимеры и клетка.— 1990.— **6**, № 6.— С. 42—48.
2. *Trifonov E. N.* Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16S-rRNA nucleotide sequences // J. Mol. Biol.— 1987.— **194**, N 2.— P. 643—652.
3. *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.
4. *Бродский Л. И., Драчев А. Л., Леонтович А. М.* Новый метод множественного выравнивания последовательностей биополимеров. Программа H-Align пакета GenBee // Там же.— 1991.— **7**, № 1.— С. 14—22.
5. *Lipman D. J., Pearson W. R.* Rapid and sensitive protein similarity searches // Science.— 1985.— **227**, N 1.— P. 1435—1441.
6. *Hartigan J. A.* Clastering algorithms.— New York : John Wiley and Sons, 1975.— 256 p.
7. *Чумаков К. М., Юшманов С. В.* Принцип максимального топологического подобия в молекулярной систематике // Молекуляр. генетика, микробиология, вирусология.— 1988.— **3**, № 1.— С. 3—9.
8. *Эволюция* РНК-зависимых РНК-полимераз позитивных рибовирусов : сравнение филогенетических деревьев, построенных различными способами / Е. В. Кунин, К. М. Чумаков, С. В. Юшманов, А. Е. Горбаленя // Там же.— С. 16—19.
9. *Gibrat J.-F., Garnier J., Robson B.* Further developments of protein secondary structure prediction using information theory // J. Mol. Biol.— 1987.— **198**, N 4.— P. 425—443.
10. *Гультяев А. П., Монаков Ю. Н.* Метод построения вторичной структуры РНК на основе принципов самоорганизации // Биополимеры и клетка.— 1991.— **7**, № 1.— С. 31—36.

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

Получено 28.06.90

УДК 577.112

Л. И. Бродский, А. Л. Драчев, А. М. Леонтович

НОВЫЙ МЕТОД МНОЖЕСТВЕННОГО ВЫРАВНИВАНИЯ ПОСЛЕДОВАТЕЛЬНОСТЕЙ БИОПОЛИМЕРОВ (ПРОГРАММА H-Align ПАКЕТА GenBee)

В работе предложен новый алгоритм множественного выравнивания биологических последовательностей. В этом алгоритме вначале на основе метода DotHelix строятся консенсусные участки в данном наборе последовательностей разной толщины и раз-