Секция «Биоинженерия и биоинформатика»

Программа для поиска гомологов нуклеотидных последовательностей

Пеков Юрий Алексеевич
Студент
Московский государственный университет имени М.В. Ломоносова, Факультет биоинженерии и биоинформатики, Москва, Россия
E-mail: yurapekov@gmail.com

Задача поиска гомологов нуклеотидных последовательностей в банке данных является одной из важнейших задач биоинформатики. В случае кодирующих последовательностей для проведения такого поиска успешно используется программа TBLASTN. Но для поиска гомологов некодирующих последовательностей удовлетворяющего все потребности инструмента нет. Для этой цели используется ряд программ, самые распространенные из них — FASTA [1], BLASTN [2] и discontiguous MEGABLAST [3], однако каждая из этих программ обладает каким-либо существенным недостатком.

FASTA при сравнении двух последовательностей выдает только одно выравнивание — то, которое имеет наивысший счет среди всех найденных выравниваний. Однако очень часто в последовательности из базы данных есть и другие значимые участки, гомологичные входной последовательности. Из-за того, что выравнивание с ними обладает чуть меньшим счетом, чем лучшее выравнивание, они не обнаруживаются программой.

BLASTN использует ускоренный алгоритм поиска, записывая положение в базе данных всех слов длиной N нуклеотидов (N = 7, 11 или 15). Но такой подход приводит к уменьшению чувствительности: если в гомологичной последовательности не встретится участок длиной N букв, полностью совпадающий с участком входной последовательности, то выравнивание этих двух последовательностей не обнаружится. Discontiguous MEGABLAST индексирует в банке данных не слова, а паттерны длиной t. Паттерн соответствует слову, если в них совпадает хотя бы W букв в определенных местах. Это призвано увеличить чувствительность, но на практике в большинстве случаев чувствительность в сравнении с BLASTN уменьшается.

Целью настоящей работы было создание компьютерной программы Nhunt для поиска гомологов нуклеотидных последовательностей, превосходящей по чувствительности как программу FASTA, так и программу BLASTN. В программе использован оригинальный алгоритм отбора диагоналей, позволяющий пользователю регулировать соотношение "скорость работы – количество отобранных диагоналей". Алгоритм построения выравниваний основан на алгоритме FASTA, но лишен ряда присущих ему недостатков.

Результаты тестирования представленной программы показали, что она заметно превосходит программу FASTA как в скорости, так и в чувствительности. На ряде примеров, таких как поиск гомологов рибосомальной РНК *E. coli* в геномах различных архей и поиск гомологов митохондриальной рибосомальной РНК человека в митохондриальных геномах различных грибов и насекомых, Nhunt в чувствительности превосходит и программу BLASTN.

Исполняемые файлы для Linux x86 и amd64 доступны в Интернете по адресу

http://mouse.belozersky.msu.ru/~bennigsen/nhunt.html

Литература

1. http://faculty.virginia.edu/wrpearson/fasta/
2. ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/user_manual.pdf
3. http://www.ncbi.nlm.nih.gov/blast/discontiguous.shtml

Слова благодарности

Автор выражает благодарность своему научному руководителю, к.ф.-м.н. Спирину Сергею Александровичу за чуткое руководство и всемерную помощь при создании программы.