

Поиск генов катаболизма стероидов и автоматизация работы программы BLAST+

Брагин Е.Ю.*, Фокина В.В., Донова М.В.

Институт биохимии и физиологии микроорганизмов им. Г.К. Скрыбина Российской академии наук, обособленное подразделение Федерального исследовательского центра «Пуцинский научный центр биологических исследований Российской академии наук», Пуцино, Россия

* bragory@yandex.ru

Ключевые слова: стероиды; Перл; программа; поиск генов; бактерии

Мотивация и цель: Развитие технологий полногеномного секвенирования открыло перспективу осуществления *in silico* поиска организмов, обладающих определенными свойствами. Такой поиск может осуществляться с помощью программы BLAST+. Работа с данной программой может быть относительно легко автоматизирована с помощью мини-программ, написанных на языке программирования Perl. Одной из областей, в которых подобный автоматизированный BLAST+ поиск может быть применим, является поиск бактерий, обладающих ферментами катаболизма стероидов. Стероидные соединения активно используются в различных областях медицины, а также ветеринарии. При получении подобных соединений активно используются биотехнологические методы. Штаммы микроорганизмов, применяемые в биотехнологических процессах, в свою очередь, могут быть получены с помощью методов геной инженерии из бактерий, обладающих ферментами катаболизма стероидов.

Методы и алгоритмы: Для автоматизации работы программы BLAST+ на языке Perl была написана программа blastconductor.pl. Данная программа использует нуклеотидные последовательности геномов и аминокислотные последовательности референсных белков. Она запускает выравнивание каждой геномной последовательности с каждой референсной последовательностью с помощью программы BLAST+ в режиме tblastn. Результат каждого такого выравнивания сохраняется в виде временного файла. Из каждого такого файла программа извлекает информацию о длине референсной последовательности; количество идентичных аминокислот для каждого выровнявшегося участка; e-value, и др. Далее происходит отбор и анализ выровнявшихся последовательностей. Результаты сохраняются в виде отдельных файлов. Для поиска генов катаболизма стероидов, помимо описанной выше программы blastconductor.pl, использовали программу BLAST+ версии 2.11.0, а также аминокислотные последовательности 230 референсных белков, которые связаны с катаболизмом стероидов (111 последовательностей), желчных кислот (46 последовательностей), а также диосгенина (73 последовательности). Поиск генов, ассоциированных с катаболизмом стероидов, проводили среди 567 известных геномов прокариот, депонированных во Всероссийской коллекции микроорганизмов (ВКМ) в ИБФМ РАН. Все референсные нуклеотидные и аминокислотные последовательности были получены из базы данных GenBank (NCBI).

Результаты: Поиск генов, связанных с катаболизмом стероидов, позволил выявить в 64 бактериальных геномах не менее 30 генов, схожих с генами, ассоциированными с катаболизмом стероидов не менее чем на 50 %. Из них 21 относились к роду *Rhodococcus*, 16 – к роду *Streptomyces*, 6 штаммов – к роду *Gordonia*, 4 – к роду *Actinomadura* и 3 штамма – к роду *Amycolatopsis*. В этой группе также присутствовали представители родов *Mycolicibacterium*, *Nocardia*, *Saccharopolyspora*, *Couchioplanes*, *Nocardioiodes*, *Actinocorallia*, *Actinoplanes* и *Crossiella*. От 20 до 29 таких генов было обнаружено у большого количества штаммов (табл. 1). Количество штаммов, у которых найдено значительное количество генов, схожих с генами катаболизма желчных кислот, было значительно меньше (см. табл. 1). Меньше всего было обнаружено штаммов, обладающих значительным количеством генов, схожих с генами кластера катаболизма диосгенина (см. табл. 1).

Таблица 1. Количество выявленных штаммов с различным количеством генов, ассоциированных с катаболизмом стероидных соединений

Количество генов, схожих с референсными генами не менее чем на 50 %	Количество бактериальных штаммов, у которых выявлены гены, схожие с генами катаболизма		
	стероидов	желчных кислот	диосгенина
10–19	20	3	23
20–29	98	18	3
>30	64	15	1

Выводы: Создана автоматизирующая работу BLAST+ программа *blastconductor.pl*, позволяющая быстро и просто осуществлять *in silico* поиск значительного количества генов в большом количестве геномов. Найдено значительное количество бактериальных штаммов, обладающих генами, связанными с катаболизмом стероидов, которые могут быть использованы при создании штаммов, задействованных в получении стероидных соединений.

Финансирование: Работа выполнена при финансовой поддержке Российского научного фонда (проект № 21-64-00024).

The search for steroid catabolism genes and automation of the BLAST+ program

Bragin E.Yu.*, Fokina V.V., Donova M.V.

G.K. Skryabin Institute of Biochemistry and Physiology of Microorganisms, Russian Academy of Sciences; Federal Research Center “Pushchino Scientific Center for Biological Research of the Russian Academy of Sciences”, Pushchino, Russia

* bragory@yandex.ru

Key words: steroids; Perl; program; gene search; bacteria

Motivation and Aim: The full-genome sequencing technologies development had opened up the possibility of *in silico* searching for organisms with certain properties. Such a search can be carried out using the BLAST+ program. Working with this program can

be relatively easily automated using mini-programs written in the Perl programming language. One of the areas in which such an automated BLAST+ search can be applied is the search for bacteria possessing steroid catabolism enzymes. Steroid compounds are actively used in various fields of medicine, as well as in veterinary. Biotechnological methods are actively used in the production of such compounds. Strains used in biotechnological processes can be obtained using genetic engineering methods from bacteria possessing steroid catabolism enzymes.

Methods and Algorithms: To automate the operation of the BLAST+ program, the program which was named *blastconductor.pl* was written in Perl language. This program uses nucleotide sequences of genomes and amino acid sequences of reference proteins. It starts the alignment of each genomic sequence with each reference sequence using the BLAST+ program in *tblastn* mode. The result of each alignment is saved as a temporary file. From each such file the program extracts the following information: length of the reference sequence; number of identical amino acids for each aligned region; *e*-value and others. The selection and analysis of aligned sequences are carried out. The results are saved as separate files. To search for steroid catabolism genes, in addition to the *blastconductor.pl* program described above, the BLAST+ program version 2.11.0 was used, as well as the amino acid sequences of 230 reference proteins, which associated with catabolism of sterols (111 sequences), bile acids (46 sequences) and diosgenin (73 sequences). An *in silico* search for genes associated with steroid catabolism was carried out among 567 known prokaryotic genomes deposited in All-Russian collection of microorganisms (VKM) IBFM RAS. All reference nucleotide and amino acid sequences were obtained from the GenBank database (NCBI).

Results: The search for genes associated with steroid catabolism made it possible to identify at least 30 genes in 64 bacterial genomes that are similar to known genes associated with sterol catabolism by at least 50 %. Of these, 21 belong to the genus *Rhodococcus*, 16 to the genus *Streptomyces*, 6 strains to the genus *Gordonia*, 4 to the genus *Actinomadura* and 3 strains to the genus *Amycolatopsis*. Representatives of the genera *Mycobacterium*, *Nocardia*, *Saccharopolyspora*, *Couchioplanes*, *Nocardioides*, *Actinocoralia*, *Actinoplanes* and *Crossiella* were also present in this group. From 20 to 29 such genes were found in a large number of strains (Table 1). The number of strains in which a significant number of genes similar to the genes of the bile acid catabolism were found was significantly smaller (Table 1). The least number of strains were found with a significant number of genes similar to the genes associated with the diosgenin catabolism (Table 1).

Table 1. The amount of founded strains with different numbers of genes associated with the catabolism of sterols, bile acids and diosgenin

The number of genes similar to the reference genes by at least 50 %	The number of revealed bacterial strains possessing genes similar to genes associated with the catabolism of		
	sterol	bile acids	diosgenin
10–19	20	3	23
20–29	98	18	3
> 30	64	15	1

Conclusion: The program, which automates the operation of BLAST+, has been created to quickly and easily perform *in silico* searches for a significant number of genes, in a large amount of genomes. A significant number of bacterial strains have been found that possessing genes associated with steroid catabolism, which can be used in the creation of strain-biocatalysts involved in the production of valuable steroid compounds.

Funding: This work was financially supported by the Russian Science Foundation (project No. 21-64-00024).