Секция «Биоинформатика»

PQ: новая программа филогенетической реконструкции

Пензар Дмитрий Дмитриевич

Студент (специалист)

Московский государственный университет имени М.В.Ломоносова, Факультет биоинженерии и биоинформатики, Москва, Россия

E-mail: dmitry_penzar_1996@mail.ru

Филогенетические деревья – деревья, наглядно представляющие процесс происхождения видов от общего предка. Так как достоверно происхождение известно лишь для малого числа видов, то прибегают к «реконструкции филогенетических деревьев». Для этого используют выровненные последовательности характерных для организмов белков или нуклеиновых кислот. Известно большое число методов оценки качества гипотетического дерева по заданному выравниванию, самые популярные из них: максимальная экономия (парсимония), максимальное правдоподобие, ряд дистанционных методов, таких как метод наименьших квадратов и метод минимальной эволюции. Имея оценку качества, можно выбрать из возможных деревье наилучшее. Так как построение дерева полным перебором уже для деревьев из 15 видов представляется затруднительным, то были разработаны многочисленные методы, помогающие в большинстве случаев найти оптимальное дерево или близкое к оптимальному.

В данной работе предлагается новая оценка качества филогенетических деревьев. А именно, пусть нам дано **выравнивание последовательностей** гомологичных белков разных видов (s_i — последовательность с номером i) длиной L и матрица замен M, тогда качество дерева определяется по формуле:

$$\sum_{i,k,l,m} W_{iklm} \ ,$$

где $i,\ k,\ l,\ m$ пробегает все четвёрки последовательностей входного выравнивания такие, что пара (i,k) отделена от пары (l,m) хотя бы одной ветвью входного дерева.

$$W_{iklm} = \sum_{p=1}^{L} w_{p\ :\ i,k,l,m} \ ,$$

где

$$w_{p\ :\ i,k,l,m} = max(M(s_{ip}, s_{kp}) - penalty, 0) + max(M(s_{lp}, s_{mp}) - penalty, 0) \ ,$$

где

$$penalty = max(M(s_{ip}, s_{lp}), M(s_{ip}, s_{mp}), M(s_{kp}, s_{lp}), M(s_{kp}, s_{mp}))$$

Были реализованы следующие методы поиска оптимального дерева:

1) **Выращивание и множественное выращивание** со случайным перемешиванием входных последовательностей
2) **Улучшение имеющегося дерева**.Данный подход реализован в двух вариантах: детерминированный **NNI** (Nearest neighbor interchange) и случайное блуждание с отжигом на основе **NNI**
3) **Branch and bound** – переборный алгоритм, который в типичных случаях позволяет перебирать не очень большое число вариантов и при этом гарантированно дает глобальный максимум.
4) **Множественное выращивание до насыщения** – повторение выращивания до стабилизации функции распределения качества полученных деревьев

5) **Ограниченный branch and bound** – алгоритм, осуществляющий перебор до достижения определенной «глубины» в дереве построения филогенетического дерева.

На основании анализа результатов предложен протокол выбора метода поиска в зависимости от объёма входных данных.

Создан также веб-интерфейс к программе, доступный по адресу:
http://mouse.belozersky.msu.ru/tools/pq/pq.html

Тестирование программы на наборах белковых последовательностей с известной филогенией показало хорошее качество её работы, ни одна из нескольких протестированных программ не превзошла нашу по качеству реконструкции филогении на контрольной выборке.

Слова благодарности

Выражаю благодарность своему научному руководителю С.А. Спирину.