Секция «Биоинженерия и биоинформатика»

Сравнение устойчивости программ реконструкции филогении к шуму во входных данных

Галицына Александра Алексеевна
Студент
Московский государственный университет имени М.В. Ломоносова, Факультет биоинженерии и биоинформатики, Москва, Россия
E-mail: agalicina@fbb.msu.ru

Три программы филогенетической реконструкции (fitch, neighbor из пакета PHYLIP [3] и fastme [2]) восстанавливают дерево из матрицы попарных эволюционных расстояний с использованием различных алгоритмов. Выяснение зависимости устойчивости программы от топологии дерева представляет практический интерес, в частности позволит избежать ошибки притяжения длинных ветвей [1], выбрав наиболее подходящую программу.

Для исследования были взяты последовательности 16S и 23S рРНК 45 протеобактерий, по девять организмов из каждого из пяти классов. Программой dnadist пакета PHYLIP из выравниваний этих последовательностей были получены матрицы попарных эволюционных расстояний. Для получения «зашумленной» матрицы каждый элемент матрицы умножался на случайное число, распределённое логнормально со средним 0 и задаваемым стандартным отклонением – уровнем шума. Построение деревьев по исходной и зашумленной матрицам и их сравнение позволяет судить об эффектах шума заданного уровня. Этим методом проверялась устойчивость трех программ к шуму и выявлялись типичные ошибки каждой из программ. В качестве критериев устойчивости использовались расстояние между деревьями из реальных и зашумленных данных, вероятность исчезновения правильной ветви или появления ложной при добавлении шума, уровень полураспада ветви – уровень шума, при котором вероятность исчезновения ветви составляет 0,5.

Для исследования типичных ошибок использовались несбалансированные деревья. Для получения выравнивания, чьё дерево считалось несбалансированным, выбирались два из пяти классов. Затем эти два класса «прореживались» — в каждом из них оставлялся только один вид из девяти, дающий либо самую длинную, либо самую короткую ветвь в данном классе; в остальных трех классах оставалось по девять видов.

На сбалансированном дереве (45 видов) показано, что neighbor и fastme в равной степени уступают fitch по устойчивости к шуму. Все три программы восстанавливают деревья хуже из выравнивания 16S рРНК, чем из 23S рРНК и объединенного 16S-23S. На несбалансированных деревьях доказано, что neighbor и fastme, но не fitch, склонны к ошибкам по типу притяжения длинных ветвей.

Тем самым программа fitch более универсальна и точна, хотя и затрачивает больше времени на вычисления. Если в дереве предполагается наличие двух длинных ветвей, следует ожидать высокую вероятность их ошибочного объединения при использовании neighbor и fastme. Программа fitch позволяет избежать ошибки такого рода.

Литература

1. J. Bergsten. A review of long-branch attraction // Cladistics, 2005 21(2): 163–193

2. R. Desper, O. Gascuel. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle // Journal of Computational Biology. 2002 9(5):687–705

3. PHYLIP: http://evolution.genetics.washington.edu/phylip.html

Слова благодарности

Выражаю благодарность науному руководителю С.А.Спирину (НИИФХБ им. А.Н.Белозерского МГУ).