Секция «Биоинженерия и биоинформатика»

Нормализация выборок нуклеотидных и аминокислотных последовательностей и оценка влияния состава и размеров выборок на результаты статистического анализа ДНК

Кветко Павел Юрьевич
Студент
БГУ - Белорусский государственный университет, Биологический факультет, Минск, Беларусь
E-mail: pavelkvetko@gmail.com

Анализ АК- и НК-последовательностей в различных таксонах сопряжён с погрешностями, связанными с нерепрезентативностью выборок. Выборки же являются нерепрезентативными вследствие неравновероятного нахождения в них филогенетически близких и удалённых видов. Т.к. статистические показатели, применяемые при анализе выборок, напрямую зависят от состава выборки, нашей целью было создание инструмента для нормализации выборок последовательностей.

С помощью средств языка программирования Python была создана программа, позволяющая в процессе статистического анализа АК- и НК-последовательностей избежать проблем, связанных с неоднородностью состава и размеров сравниваемых выборок. Данный инструмент создает заданное количество искусственных выборок (реплик) установленного исследователем размера путем случайного изъятия последовательностей из исходной выборки. Разработанное средство дополнительно позволяет провести сортировку последовательностей по названию вида, проверку принадлежности последовательности к желаемому гену или его продукту, удалить повторяющиеся последовательности. Эмпирически установленное время выполнения программы линейно возрастает на интервале входных данных от 10 до 100000 последовательностей.

Тестирование программы провели в рамках проверки гипотезы о возможности установления относительного времени дивергенции таксонов методом нахождения т.н. «точки насыщения ДНК» [1]. В данной работе возникла проблема сравнения выборок, существенно различающихся по размеру и составу. Требовалось установить, как повлияет изменение размеров и состава выборок на получаемые результаты для подтверждения их достоверности. Для этого использовали 13275 последовательностей насекомых из 7 семейств, полученных из BOLD. Для каждой выборки создавалось по 10 реплик с возрастающей долей случайно изымаемых последовательностей (от 5 до 25 с шагом в 5 %). Находимые статистические показатели для реплик существенно не отличались при варьировании размера выборок, что позволило достоверно подтвердить получаемые результаты [2].

Таким образом, был создан и применён эффективный инструмент для быстрой подготовки и нормализации наборов последовательностей для последующего статистического анализа.

Литература

1. Воронова Н.В., Ризевский С.В., Курченко В.П., Буга С.В. Оценка уровня насыщения последовательности гена субъединицы I цитохромоксидазы с у тлей

(Hemiptera: Sternorrhyncha: Aphidoidea) как метод определения относительного возраста таксонов / Сборник научных трудов «Факторы экспериментальной эволюции организмов», Т. 12. – Киев, Логос, 2013. – С. 102–106.

2. Кветко П.Ю. Установление относительного времени дивергенции таксонов на основе оценки степени насыщения гена COI / Сборник материалов IV Международной научно-практической молодежной конференции «Научные стремления – 2013». – Минск, 3–6 декабря, 2013. – С. 44–47.