УДК 004.942

Байесовский подход к построению филогенетических сетей

Новик Д.С., Университет ИТМО, г. Санкт-Петербург
Научный руководитель – Алексеев Н.В., к.ф.-м.н., ведущий научный сотрудник ФИТиП, Университет ИТМО, г. Санкт-Петербург

Настоящая работа посвящена разработке методов построения филогенетических сетей близкородственных штаммов вирусов. В качестве подзадачи рассматривается построение непротиворечивых сетей для небольших входных данных. Также рассматриваются подходы построения модели, учитывающей возможные рекомбинации.

Введение. Филогенетическая сеть — граф, отражающий эволюционные взаимосвязи между различными видами или другими сущностями, имеющими общего предка. Каждой вершине такого дерева соответствует некоторый вид живых организмов или какой-то другой объект (например, штамм вируса). Также каждый узел представляет собой некоторое эволюционное событие, при котором произошло разделение предкового вида на два или более, которые далее эволюционируют независимо. Веса на ребрах соответствуют эволюционному расстоянию между вершинами.

Существующие методы построения филогенетических сетей могут выдавать сети, которые не могли существовать в реальном эволюционном процессе, потому что они не проверяют непротиворечивость своих промежуточных результатов. Также не существует подходов, которые во время построения сетей, учитывали бы существование такого эволюционного события, как рекомбинация. Зачастую это не мешает применять на практике такие подходы, однако в случае работы с вирусами, которым свойственны частые рекомбинации, полученные результаты могут абсолютно не соответствовать тому, как эволюция происходила на самом деле.

Основная часть. Построение филогенетических сетей зачастую выполняется с помощью Markov Chain Monte Carlo симуляций, во время которых выполняются случайные изменения в структуре сети, после чего вычисляется функция правдоподобия и решается хорошая ли была произведена модификация сети. Сами же симуляции необходимо начинать с какой-то уже построенной филогенетической сети.

Вместо того, чтобы подавать на вход симуляциям какую-то случайную филогенетическую сеть, предлагается сначала построить сеть с помощью метода, который не будет выполнять симуляции, а будет смотреть только на расстояние Хэмминга между разными геномами. На основе алгоритма TCS был разработан эвристический метод, который во время построения сети выполняет перебор с отсечениями возможных геномов в добавляемых промежуточных вершинах. Перебор выполняется до тех пор, пока расстояние Хэмминга между объединяемыми геномами не становится большим. При этом получается непротиворечивая филогенетическая сеть, для которой можно сделать функцию правдоподобия, которая максимально уверена в корректности некоторых подграфов, полученной сети.

Рекомбинации в свою очередь происходят не в случайных местах генома, а в так называемых хотспотах, поэтому можно разбить все геномы по этим хотспотам, посчитать для каждого независимого участка генома филогению обычным способом, а потом слить получившиеся графы в один. Также из-за того, что рекомбинации происходят несколько реже, чем мутации, можно установить им некоторую цену, и, если в какой-то момент построения сети алгоритм хочет добавить мутаций больше, чем цена рекомбинации, можно проверить наличие такой тройки геномов, которые могли получиться с помощью рекомбинации.

Выводы. Разработанный на основе алгоритма TCS эвристический подход позволяет для небольших входных данных построить непротиворечивую филогенетическую сеть. Также

предложены способы того, как можно учесть во время построения сетей информацию о возможных рекомбинациях.

Новик Д.С.

Алексеев Н.В.