Секция «Биоинформатика»

Сравнение сложности натуральных и симулированных выравниваний белковых последовательностей

Научный руководитель – Спирин Сергей Александрович

Малышев Андрей Дмитриевич

Студент (специалист)

Московский государственный университет имени М.В.Ломоносова, Москва, Россия

E-mail: malyshew.a.d@yandex.ru

Филогенетическая реконструкция подразумевает установление родства между организмами или иными объектами, например, белками на основании данных о последовательностях биополимеров. Искомым результатом является двоичное дерево с некоторой топологией и длинами ветвей, листья которого соответствуют рассматриваемым объектам.

Для восстановления дерева по выравниваниям последовательностей существуют различные алгоритмы и их программные реализации. Зачастую для оценки точности их работы требуются искусственно создаваемые на основе деревьев выравнивания, поскольку информация о филогении реальных объектов бывает недоступна. С этой целью разрабатываются программы, симулирующие выравнивания. Важно, чтобы получаемые данные по сложности с точки зрения филогенетической реконструкции были сопоставимы с натуральными.

Целью данной работы является выявление критериев, которые позволяли бы сравнивать натуральные и симулированные выравнивания. Исходные данные представляют собой наборы натуральных выравниваний с разным числом белковых последовательностей и соответствующие им референсные деревья. На их основе с помощью программы INDELible [1] осуществлялась симуляция выравниваний с теми же параметрами длины и доли инвариантных сайтов. И натуральные, и искусственные выравнивания затем подавались на вход трём программам филогенетической реконструкции: FastME [2], RAxML-VIII [3], RAxML-NG [4].

Было выяснено, что натуральные выравнивания оказываются сложнее симулированных с точки зрения филогенетической реконструкции: в среднем программе RAxML-VIII требуется большее число шагов алгоритма SPR для поиска оптимального дерева, а при оптимизации 10 случайных деревьев программой RAxML-NG среднее относительное расстояние между результатами оказывается выше, если использовать натуральные выравнивания. Кроме того, обнаруживается разница в качестве реконструкции, осуществляемой программами RAxML и FastME на натуральных и искусственных выравниваниях. В работе также сравнивается эффективность моделей аминокислотных замен JTT и LG при реконструкции деревьев и обсуждаются различные аспекты применения гамма-распределения, отражающего разнородность сайтов с точки зрения скорости накопления мутаций.

Источники и литература

1) Fletcher, W., Yang, Z. INDELible: A Flexible Simulator of Biological Sequence Evolution // Molecular Biology and Evolution, 26(8), 2009, 1879–1888.
2) Lefort, V., Desper, R., & Gascuel, O. FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program // Molecular Biology and Evolution, 32(10), 2015, 2798–2800.
3) Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies // Bioinformatics, 30(9), 2014, 1312–1313.

4) Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference // Bioinformatics, 35(21), 2019, 4453–4455.