Conference track «Bioinformatics»

Предсказание качества филогенетической реконструкции по характеристикам входного выравнивания

Научный руководитель – Спирин Сергей Александрович

Коростина Мария Алексеевна
Student (specialist)
Московский государственный университет имени М.В.Ломоносова, Факультет биоинженерии и биоинформатики, Москва, Россия
E-mail: masha.korostina@mail.ru

Филогенетическая реконструкция — важная задача биологической науки, которая заключается в построении связного ациклического графа (филогенетического дерева), отражающего эволюционные связи между биологическими объектами, такими как белковые домены, гены или виды организмов. Программы реконструкции деревьев допускают ошибки, и актуальной задачей является оценка уровня ошибок в каждом конкретном случае. Нормализованное расстояние Робинсона-Фоулдса (nRFD) [3] — это число нетривиальных разбиений, различающихся между двумя деревьями, деленное на суммарное количество нетривиальных разбиений в этих деревьях. nRFD принимает значения от 0 (если деревья идентичны) до 1 (если деревья полностью различны). RFD позволяет оценить структурные различия между деревьями, а нормирование позволяет сравнивать расстояния деревьев разного размера.

В исследованиях, связанных с оценкой точности филогенетических деревьев [1, 2], была предпринята попытка предсказывать расстояние Робинсона-Фоулдса между реконструированным и эталонным деревом по свойствам выравнивания последовательностей. Однако осмысленная реализация такой задачи затрудняется тем, что обучение модели проводилось на деревьях фиксированного размера.

Целью данной работы является решение задачи оценки деревьев произвольного размера. Для достижения этой цели было исследовано несколько подходов: предсказание расстояния между полными деревьями с помощью значений nRFD между поддеревьями; экстраполирование значений расстояния между деревьями на деревья большего размера, экстраполирование значений свойств выравниваний, матриц расстояния и деревьев и последующего предсказания по ним nRFD. Для обучения предсказания была применена модель градиентного бустинга XGBoost на выборке деревьев размера 30.

References

1) Ефремов А.А. Программа для предсказания точности филогенетической реконструкции методами машинного обучения, дипломная работа.
2) Krivozubov, M., Goebels, F., & Spirin, S. (2014). Estimation of relative effectiveness of phylogenetic programs by machine learning. Journal of Bioinformatics and Computational Biology, 12(02), 1441004.
3) Robinson DR, Foulds LR. Comparison of phylogenetic trees. Math Biosci. 1981:53(1–2):131–147. https://doi.org/10.1016/0025-5564(81)90043-2.