Секция «Биоинформатика»

# Предсказание качества филогенетической реконструкции методом машинного обучения

**Научный руководитель – Спирин Сергей Александрович**

***Никитин Иннокентий Дмитриевич***

*Студент (специалист)*

Московский государственный университет имени М.В.Ломоносова, Факультет биоинженерии и биоинформатики, Москва, Россия

*E-mail: nikitinkesha94@gmail.com*

Филогенетическое дерево, построенное по заданному множественному выравниванию, далеко не всегда точно описывает реальную филогению соответствующих белков. Встаёт задача оценки вероятного качества филогенетической реконструкции на основании данного выравнивания. Машинное обучение применяется в широком спектре задач, в частности и в биоинформатике. Ранее уже была доказана возможность предсказания категории выравнивания (хорошее/плохое) и оптимального метода филогенетической реконструкции [1]. В данной работе задачей является оценка расстояния реконструированного дерева от реального методом машинного обучения.

Для работы используется 20 различных признаков, полученных из выравниваний. Для обучения и тестирования используется набор из нескольких тысяч семейств ортологичных белков, взятых из организмов с известной филогенией.

Для выбора основного алгоритма первоначально модели тренировались и тестировались на выборках выравниваний с постоянным числом последовательностей. Наилучший результат показали алгоритмы градиентного бустинга, однако алгоритм "случайный лес" также показал хорошие результаты. Для этих же условий был подобраны параметры для бустинга.

Была предпринята попытка обобщить работу предсказателя до выборок с выравниваниями с различным числом последовательностей. Для этого признаки были нормированы относительно числа последовательностей, а само число было добавлено в список признаков. Сама нормировка увеличивает качество предсказания. Была оценена способность таких предсказателей к обощению — насколько хорошо модель, натренированная на выравниваниях с малым числом последовательностей, оценивает выравнивания с большим числом. При этом наблюдался сдвиг в предсказаниях, по сравнению с моделью, натренированной на равномерной выборке.

В ближайших планах — создание практически применимого предсказателя качества реконструкции по выравниваниям с произвольным числом последовательностей.

Выражается благодарность научному руководителю Сергею Александрович Спирину (НИИ ФХБ им. Белозерского).

## Источники и литература

1) Krivozubov M., Goebels F., Spirin S. Estimation of relative effectiveness of phylogenetic programs by machine learning // Journal of Bioinformatics and Computational Biology. 2014. 12(2).