Отзыв научного руководителя
на курсовую работу
студента 344 группы П. С. Луниной
“Использование нейронных сетей для распознавания 16s РНК по вторичной структуре”

Классификация (микро)организмов —- одна из часто встречающихся задач биоинформатики. Для её решения часто используются так называемые маркерные гены (например транспортная РНК или 16s рибосомальная РНК). Однако, прежде чем использовать маркерные гены для классификации, их необходимо выделить из биологических материалов (например, результатов секвенирования генома, метагеномной сборки и т.д., в рамках данной работы будет предполагаться, что сборка генома уже проведена).

С точки зрения биоинформатики геном может быть рассмотрен как строка над алфавитом из 4 символов, что позволяет использовать для поиска маркерных генов различные алгоритмы текстового поиска. Вместе с этим, показано, что многие участки “генетического текста” обладают некоторой дополнительной структурой, которая называется вторичной и является, по сути своей, синтаксической структурой этого “текста”. Задача осложняется тем, что в реальных данных всегда присутствуют различного рода “шумы” и мутации, из-за чего не приходится ожидать точного совпадения, что не позволяет использовать точные алгоритмы и приводит к использованию различных вероятностных подходов, таких как скрытые модели Маркова, ковариационные модели и т.д.. Было показано, что учёт вторичной структуры в вероятностных моделях для задач поиска маркерных генов позволяет улучшить качество работы соответствующих инструментов.

В рамках данной работы перед П. С. Луниной была поставлена задача проверки гипотезы о том, что вторичная структура 16s рРНК достаточно богата для того, чтобы использоваться в качестве основной информации для обнаружения этого маркерного гена.

В ходе работы П. С. Лунина смогла собрать воедино такие далёкие на первый взгляд области, как синтаксический анализ и теория формальных языков, нейронные сети, обработка генома. В итоге был предложен набор и разработан набор инструментов, который позволяет с помощью синтаксического анализа выявить особенности вторичной структуры маркерных генов и оформить их набор данных для обучения нейронной сети, которая должна выявить наиболее существенные особенности. Далее были поставлены эксперименты по оценке качества полученной сети, которые показали, что детектирование маркерных генов на основе особенностей вторичной структ извлечённых предложенным образом, возможно, а значит гипотеза состоятельна и необходимы дальнейшие исследования в данном направлении.

Выполнение работы потребовало от П.С. Луниной не только хороших исследовательских навыков, которые были продемонстрированы, но и определённых инженерных, необходимых для создания комплексного решения, включающего модуль обработки данных (загрузка данных и метаданных, синтаксический анализ), реализованный на F#, нейронную сеть (загрузка данных, их перд-и постобработка, архитектура сети), реализованную на Pytjon, которые также были продемонстрированы на должном уровне.

Связь с руководителем поддерживалась хорошо, поставленные задачи решались в

установленные сроки, замечания исправлялись оперативно.

Текст отчета грамотно структурирован, содержит все основные разделы, в достаточной мере раскрывает содержание проделанной работы. Необходимо отметить, что некоторые части текста выглядят слабо связанными между собой, а в некоторых местах ощущаются пропуски в изложении, однако, вероятно, это связано с тем, что работа велась на стыке слишком большого количества областей и действительно связанное её представление —- весьма нетривиальная задача.

Считаю, что работа П. С. Луниной заслуживает оценки “отлично”.

Руководитель курсовой работы,
К.ф.-м.н., доцент кафедры информатики СПбГУ,
. …… /С. В. Григорьев /