АУГМЕНТАЦИЯ ДАННЫХ ДЛЯ ОБУЧЕНИЯ ОСНОВАННОГО НА ИСКУССТВЕННОЙ НЕЙРОННОЙ СЕТИ КЛАССИФИКАТОРА ПОСЛЕДОВАТЕЛЬНОСТЕЙ МИКРОБНЫХ РОДОПСИНОВ

Богданова Е.А., Шайтан К.В., Новоселецкий В.Н.

Московский государственный университет имени М.В.Ломоносова, Биологический факультет, Кафедра биоинженерии, Москва, Россия E-mail: elizawea@yandex.ru

Микробные родопсины - надсемейство фотоактивных ретиналь-связывающих мембранных белков, широко распространенных в микробном мире и среди низших эукариот [2]. Представители этого надсемейства обладают общим семиспиральным строением, но демонстрируют высокое разнообразие функций. Исследование микробных родопсинов расширило понимание структуры и функциональной активности мембранных белков, фотохимии, сенсорной сигнализации, эволюции белков, а также механизмов взаимодействия организмов со светом [1].

В литературе [1] описано несколько семейств микробных родопсинов (бактериородопсины, галородопсины и др.), причем для отнесения аминокислотной последовательности к тому или иному семейству требуется сравнить её со всеми гомологичными последовательностями. Методы машинного обучения способны классифицировать последовательности без такого сравнения, однако для корректного обучения этих алгоритмов требуются достаточно большие наборы данных, недоступные в настоящее время. Решением этой проблемы может оказаться генерация псевдопоследовательностей, обладающих свойствами природных последовательностей. Данная методика расширения набора обучающих данных из уже имеющихся получила название аугментация.

Целью настоящей работы являлась генерация искусственных последовательностей микробных родопсинов. Используя кластеризацию, мы разделили надсемейство микробных родопсинов на 14 семейств, и для каждого из них осуществлялась генерация псевдопоследовательностей. Особенностью нашего метода генерации является учет особенностей аминокислотного состава трансмембранных и немембранных участков белка, что позволяет максимально приблизить искусственные последовательности по структуре к природным. Так, расширение набора последовательностей природных микробных родопсинов псевдопоследовательностями может помочь в изучении структурных и функциональных особенностей различных родопсинов. Сгенерированные нами последовательности были использованны для классификатора, основанного на многослойном перцептроне. Так, нам удалось достичь точности предсказания 100% для 33 микробных родопсинов известных классов, которые не были использованы для аугментации.

1) 1. Govorunova E. G., Sineshchekov O. A., Li H., Spudich J. L. Microbial Rhodopsins: Diversity, Mechanisms, and Optogenetic Applications // Annual review of biochemistry. 2017. V. 86. P. 845–872.

2) 2. Spudich J. L., Yang C. S., Jung K. H., Spudich E. N. Retinylidene proteins: structures and functions from archaea to humans // Annual review of cell and developmental biology. V. 16. P. 365–392.