Секция «Вычислительная математика и кибернетика»

Формирование непротиворечивого представительного подмножества прецедентов для распознавания вторичной структуры белка

Солодкин Дмитрий Леонидович
Студент
Московский государственный университет имени М.В. Ломоносова, Факультет вычислительной математики и кибернетики, Москва, Россия
E-mail: rooney74@mail.ru

Задача прогнозирования свойств белка по его первичной структуре является одной из ключевых в биоинформатике. Она разбивается на несколько подзадач, первой из которых является распознавание вторичной структуры белка по его аминокислотной последовательности.

Существующие методы распознавания существенно ограничены по точности, главным образом, из-за неполноты и неточности исходных данных. Имеющиеся в свободном доступе база данных белков PDB (Protein Data Base) содержит большое количество «неточных» дубликатов одних и тех же белков, при этом данные о вторичных структурах получены по различным методикам, имеющим различные погрешности, и потому могут противоречить друг другу. Количество «неточных» дубликатов в базе существенно неравномерно (от 1 до 10000 для гемоглобина) и составляет в среднем около 9 дубликатов на каждый белок. Различия во вторичных структурах дубликатов составляет около 20 процентов от общего числа символов. В данной работе предложен и реализован метод формирования представительной подвыборки белков, пригодной для обучения распознаванию вторичной структуры белков.

Для формализации сходства белков как аминокислотных последовательностей использовалась нормированное расстояние Левенштейна. Для того, чтобы проверить качество работы алгоритма был предложен следующий тест: для большого количества произвольно выбранных белков были выделенным кластеры «схожих», для каждого представителя кластера были выделены свои кластеры «схожих» и показано, что эти кластеры совпадают на 97 процентов, что говорит о получении достаточно четкой кластеризации. В качестве алгоритмов выделения наиболее представительных прецедентов из PDB рассмотрено несколько известных алгоритмов кластеризации (Форэль, Complete-link) и предложен один новый, основанный на принципе ранжирования: каждый следующий выделяемый из общей базы представительный прецедент должен быть как можно менее «схож» со всеми выделенными ранее. Для того, чтобы сравнить качество алгоритмов выделения представительного подмножества белков, было предложено несколько критериев, а именно алгоритм считается тем лучше, чем больше он набрал 1) выделенных белков 2) символов в общей длине выделенных белков 3) «сегментов» вторичной структуры в выделенных белках. Предложенный алгоритм показал наилучшие результаты по всем приведенным критериям, превысив показатели остальных алгоритмов на 5-10 процентов.

Литература

1. Загоруйко Н. Г., Елкина В. Н., Лбов Г. С. Алгоритмы обнаружения эмпирических закономерностей. Новосибирск: Наука, 1985.

2. Кристофер Д. Маннинг, Прабхакар Рагхаван, Хайнрих Шютце. Введение в информационный поиск. МоскваВильямс, 2011. 528 с.

3. Torshin I.Yu. Bioinformatics in the Post-Genomic Era: The Role of Biophysics, 2006 Nova Biomedical Books, NY, ISBN 1-60021-048-1.

4. Torshin I.Yu. Bioinformatics in the post-genomic era: physiology and medicine. Nova Biomedical Books, NY, USA (2007), ISBN 1-60021-752-4.

Слова благодарности

Выражаю благодарность своему научному руководителю д.ф-м.н. Воронцову Константину Вячеславовичу и эксперту к.х.н. Торшину Ивану Юрьевичу