Секция «Биоинформатика»

Кластеризация последовательностей с помощью дифференцируемого редакционного расстояния

Научный руководитель – Горбачев Дмитрий Викторович

Офицеров Евгений Петрович
Аспирант
Тульский государственный университет, Тула, Россия
E-mail: eofitserov@gmail.com

Сравнение аминокислотных и нуклеотидных последовательностей является ключевой составляющей многих задач биоинформатики. Наиболее естественным способом сравнения двух строк различной длинны является редакционное расстояние или расстояние Левенштейна, которое определяется как минимальное количество вставок, замен и удалений символов, необходимых для преобразования одной строки в другую. Такая метрика является оправданной с биологической точки зрения, а также для её вычисления существует эффективный алгоритм динамического программирования [1]. Благодаря этому расстояние Левенштейна часто используется в задачах, связанных со сравнением биологических последовательностей. Однако, редакционное расстояние является дискретной функцией, которую сложно оптимизировать. Этот важный недостаток ограничивает применение метрики Левенштейна в задачах машинного обучения. Например, крайне сложно применить классический алгоритм кластеризации K-средних к большому набору последовательностей переменной длинны. Возможным решением является использование K-меров или других признаков для получения числовых представлений последовательностей, которые можно сравнивать как многомерные вектора [2][3]. Однако, такой подход требует дополнительных предположений о свойствах сравниваемых строк в каждой отдельной задаче.

В работе предлагается новая метрика для сравнения последовательностей - дифференцируемое редакционное расстояние, которая является гладкой аппроксимацией расстояния Левенштейна. Данная метрика, также как и оригинальное редакционное расстояние, позволяет сравнивать последовательности различной длинны с точки зрения количества вставок, замен и удалений символов, однако является непрерывной функцией и может быть продифференцирована по параметрам входных последовательностей, что позволяет оптимизировать её с помощью градиентных методов. Для расчёта значений предлагаемой метрики и её производных, так же, как и в случае классического расстояния Левенштейна, используются алгоритм на основе рекуррентных формул, позволяющий проводить вычисления за полиномиальное время.

Для демонстрации эффективности подхода, в работе предлагается реализация простого алгоритма для кластеризации нуклеотидных или аминокислотных последовательностей, позволяющего эффективно кластеризовать большие множества строк переменной длинны, а также находить центроиды или консенсусные последовательности для каждого кластера.

Источники и литература

1) Jurafsky, D., Martin, J. Speech and Language Processing. –Pearson Education International, 107–111.
2) Hsu J. L., Yang H. X. A modified K-means algorithm for sequence clustering //2009 Ninth International Conference on Hybrid Intelligent Systems. – IEEE, 2009. – Т. 1. – С. 287-292.

3) James B. T., Luczak B. B., Girgis H. Z. MeShClust: an intelligent tool for clustering DNA sequences //Nucleic acids research. – 2018. – Т. 46. – №. 14. – С. e83-e83.