УДК 621.3.049.77–048.24:537.2

ПРИЛОЖЕНИЕ ДЛЯ АНАЛИЗА ДАННЫХ МОЛЕКУЛЯРНОЙ ДИНАМИКИ БЕЛКА МЕТОДАМИ КЛАСТЕРНОГО АНАЛИЗА

Цацура Н.Ю.

Белорусский государственный университет информатики и радиоэлектроники, г. Минск, Республика Беларусь

Научный руководитель: Осипович В.С. – канд.техн.наук, доцент, доцент кафедры ИПиЭ

Аннотация. Экспериментально исследована эффективность применения методов кластерного анализа для изучения данных молекулярной динамики белка. Установлено, что кластеризация траектории конформаций белка позволяет обогащать выборку данных для повышения эффективности проведения молекулярного докинга.

Ключевые слова: анализ данных, кластеризация, кластерный анализ

Введение. Так как кластерный анализ или кластеризация предполагает разбиение выборки объектов (ситуаций) на группы, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались, методы кластеризации широко применяются в машинном обучении, позволяя получить больше информации об исследуемых данных, в частности, установление существования классов или кластеров данных и последующем применении прочих методов машинного обучения на полученных кластерах. Методы машинного обучения широко применяютсья в медицине, как, например, для кластеризации симптомов, заболеваний, препаратов.

Так как молекулярная динамика подразумевает моделирование движения множества частиц (в том числе атомов, ионов, молекул) и рассчёт коллективных свойств системы, зависящих от этого движения, результатом молекулярной динамики являются данные содержащие траекторию и топологию выбраного белка. Эти данные применяются для другого метода молекулярного моделирования, молекулярного докинга, который подразумевает предсказание наиболее выгодного для образования устойчивого комплекса ориентацию и положение одной молекулы по отношению к другой. Однако для докинга из всего набора данных нужно выбрать лишь несколько конформаций белка (состояние молекулы), следовательно они должны быть максимально репрезентативны. Используя методы кластерного анализа, можно разбить похожие состояния конформаций на кластере и выбрав в каждом кластере репрезентативное состояние молекулы провести молеклярный докинг на каждой выбраной конформации.

В данной статье автором показано, что применение кластерного анализа на наборе данных траектории изменения конформаций белка, полученных методом молекулярной динамики, позволяет получить репрезентативные конформации каждого кластера, обогащая данные для молекулярного докинга.

Основная часть. кластеризация данных требует решения следующих задач:

- выбор метода кластеризации;
- выбор репрезентативной конформаций;
- выбор количества кластеров.

Выбор метода кластеризации зависит от цели анализа. В таблице 1 представлены основные методы кластеризации, использующиеся для анализа.

Таблица 1 – Методы кластеризации

Название	Описание
Метод одиночной связи	Сходство определяется величиной расстояния между самыми близкими элементами. Есть тенденция к удлинению кластеров в одном направлении.
Метод полной связи	Сходство определяется величиной расстояния между самыми дальними элементами. Есть тенденция к избеганию удлинения кластеров в одном направлении

Продолжение таблицы 1.

Название	Описание
Центроидный метод	Сходство определяется расстоянием между центроидами кластеров. Тенденция к удлинению кластеров находится между одиночной и полной связью
Метод Уорда	После объединения кластеров вычисляется квадрат среднего расстояния между данными и центром кластера, из него вычитается эта величина до объединения. Часто получаются сравнительно хорошие кластеры.
Метод k-means (k-средних)	Метод k-средних – это метод кластерного анализа, цель которого является разделение m наблюдений (из пространства R^n) на k кластеров, при этом каждое наблюдение относится к тому кластеру, к центру (центроиду) которого оно ближе всего. В качестве меры близости используется Евклидово расстояние. Чувствителен к выбору начальных приближений центров.

Для примера кластеризации был выбран алгоритм кластеризации «Метод Уорда».

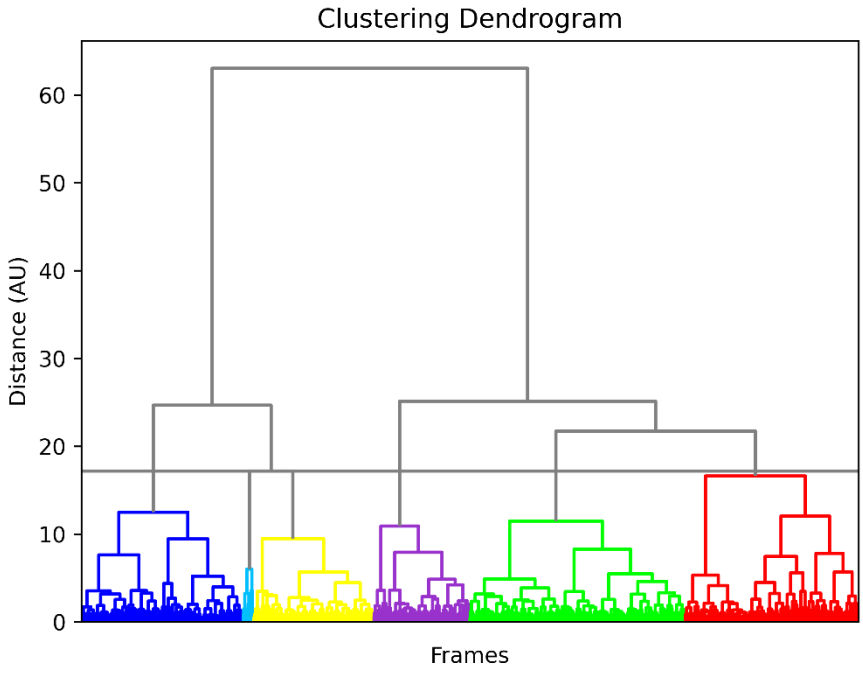

Рисунок 1 – Дендрограмма метода Уорда

Результатом кластеризации является 4 кластера с репрезентативными конформациями:
– 698 для первого кластера, количество членов 447;
– 457 для первого кластера, количество членов 416;
– 971 для первого кластера, количество членов 554;
– 175 для первого кластера, количество членов 309;
– 1063 для первого кластера, количество членов 247;
– 1140 для первого кластера, количество членов 27.
Общее количество конформаций: 2000.

Репрезентативные конформации выбирались как конформации, имеющие самое маленькое *RMSD* (*Root-mean-square deviation*, мера среднего расстояния между атомами) между всеми остальными конформациями в кластере.

Проблема выбора количества кластеров решается с помощью метода локтя. Следуя правилам этого метода, было выбрано n = 14: k>=2 и k<=14. Затем после применения метода кластеризации k-средних 13 раз, вычисляется значение сумм квадратов растояний выборок до их ближайших центроидов для каждого n, получая следующий масив значений инерций: 2492.623046875, 1961.738037109375, 1699.55224609375, 1524.90869140625, 1384.28515625, 1292.2786865234375, 1212.33203125, 1135.509521484375, 1078.6336669921875, 1034.5142822265625, 996.0445556640625, 956.3633422851562, 923.365966796875, 885.10693359375.

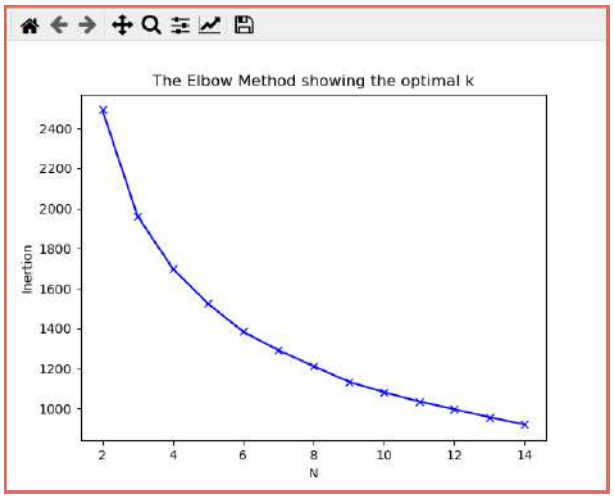

Рисунок 2 – Граф зависимости инерции от n

Для определения числа кластеров провёл условную прямую от точки начала графа к точке конца графа. После чего, программно вычислил по формуле нахождения расстояния от точки до прямой на плоскости $d = \frac{|ax_0+by_0+c|}{\sqrt{a^2-b^2}}$ где x_0 и y_0 – это координаты точки, а коэффициенты x, b, c – это уравнение прямой ax + by +c = 0, получаю длины всех отрезков: 0.0, 3.0552967276015157, 4.058082641942784, 4.392192001157954, 4.4663685393026284, 4.171345006515157, 3.771316487687424, 3.3721616816579045, 2.801264032177446, 2.1175035762185677, 1.4320385877141868, 0.7350251114767368, 0.0.

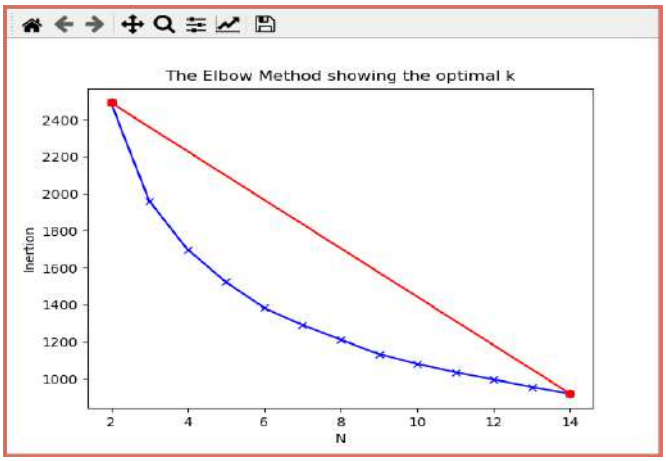

Рисунок 3 – Граф зависимости

Максимальное значение растояния 4.4663685393026284 указывает на оптимальное количество кластеров, равное 6.

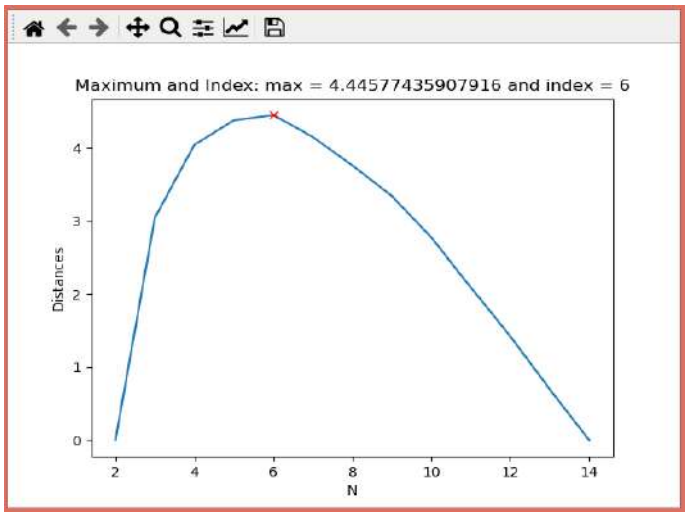

Рисунок 3 – Граф зависимости растояний отрезков от n

Заключение. Выполнен анализ методов кластеризации и метода локтя для определения количества кластеров метода *k*-средних. Продемонстрировано, что применение методов кластерного анализа на наборе данных траектории изменения конформаций белка, позволяет получить репрезентативные конформации каждого кластера.

Список литературы

1. *Воронцов К.В. Алгоритмы кластеризации и многомерного шкалирования. Курс лекций. МГУ, 2007.*
2. *Котов А., Красильников Н. Кластеризация данных. 2006.*
3. *Python Data Science Handbook by Jake VanderPlas Released November 2016*
4. Машинное обучение *с использованием* Python*. Сборник рецептов [2019] Крис Элбон*
5. *Python для сложных задач. Наука о данных и машинное обучение Дж. Вандер Плас, 2019*
6. *Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures.*

Alexander Wlodawer 1*, Wladek Minor, Zbigniew Dauter, Mariusz Jaskolski*

UDC 621.3.049.77–048.24:537.2

APPLICATION FOR THE ANALYSIS OF PROTEIN MOLECULAR DYNAMICS DATA BY CLUSTER ANALYSIS METHODS

Tsatsura N.Y.

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Osipovich V.S. – PhD, assistant professor, associate professor of the department of EPE

Annotation. The effectiveness of cluster analysis methods for studying protein molecular dynamics data has been experimentally investigated. It is established that clustering of the trajectory of protein conformations makes it possible to enrich the data sample to increase the efficiency of molecular docking.

Keywords: data analysis, clustering, cluster analysis