**Зелінський О. А.,**

студент факультету прикладної математики та інформатики

Львівський національний університет імені Івана Франка, м. Львів

**Горлач В. М.,**

кандидат фізико-математичних наук, доцент кафедри інформаційних систем

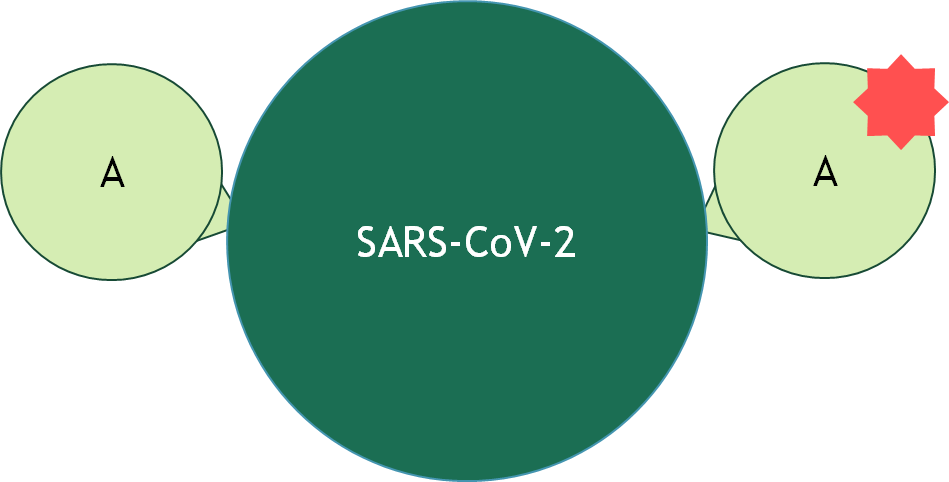
Львівський національний університет імені Івана Франка, м. Львів

**Лебедін Ю. С.,**

# Схематичне відображення 2D образів точок зв'язування антитіл на вірусному білку

В умовах пандемії надзвичайно важливими є дослідження які несуть безпосередню користь для виявлення, запобігання та лікування вірусних захворювань, а якщо точніше то вірусу Covid-19 або SARS-CoV-2. А в умовах поширення комп’ютерів та іншої потужної обчислювальної техніки зручним та важливим є використання комп’ютерних алгоритмів для виконання завдань пов’язаних з дослідженнями вірусів.

Дано молекулу вірусу SARS-CoV-2, до якої приєднуються два антитіла, для того щоб можна було їх відрізнити одне з них помічається \*. Для простоти вважатимемо, що все відбувається на площині, а антитіла це круги однакового розміру, що приєднуються до більшого круга який представляє молекулу вірусу.

  
Рис. 1 Модель приєднання антитіл до вірусної молекули

Задача полягає у знаходженні двох таких антитіл (можуть бути однаковими), таких що вони знаходяться на оптимальній відстані одне від одного (не перетинаються, не знаходяться занадто близько). Зручним теоретичним способом для цього є розбиття списку антитіл на групи (антитіла з різних груп взаємодіють краще ніж з однієї). Далі буде наведено опис алгоритму розбиття.

Дані з експерименту подані у вигляді таблиці, де кожна комірка це коефіцієнт перехресного зв’язування міченого антитіла зі стовпця та не міченого з рядка. В рядку позначеному як “blank” надані максимальні значення коефіцієнтів перехресного зв’язування.

Зображення, що містить стіл

Автоматично згенерований опис  
Рис. 2 Фрагмент початкових даних

Для подальшої роботи з даними їх позначають за наступним алгоритмом:

1. Темно-зеленим кольором (цифрою 3) антитіла з хорошим зв’язуванням, якщо
2. Світло-зеленим кольором (цифрою 2) антитіла з середнім зв’язуванням, якщо
3. Білим кольором (цифра 1) антитіла майже без зв’язування, у всіх інших випадках.

Зображення, що містить текст, зелений, надворі, вулиця

Автоматично згенерований опис  
Рис. 3 Фрагмент позначених даних

Тепер задача полягає у розбитті матриці перехресного зв’язування антитіл на групи за ознакою подібності раніше створеного показника зв’язування для полегшення виявлення оптимальних пар та приблизної локалізації місця зв’язування. Для цього використовують методи кластеризації, а саме k-modes.

k-modes – це алгоритм, який базується на алгоритмі k-середніх і використовується для кластеризації даних на основі якісних змінних. k-modes визначає кластери на основі відповідності категорій між точками даних.

Для обробки даних та кластеризації використовувалась бібліотека kmodes, pandas, matplotlib та kneed з Python. В результаті виконання програми отримано оптимальне розбиття на 6 кластерів.

|  |  |
| --- | --- |
| Кластери | Елементи |
| 0 | X201, X200, NP3701, NP1526 |
| 1 | NP1501, X215, X211, NP1512, NP1514, NP1521 |
| 2 | X221, X190, X271, NP1528, NP1525, NP3708, NP1522, NP1502, NP1520, NP1503, NP1508, NP1510 |
| 3 | X202 |
| 4 | X32, X233, X224, X223, X217, NP3715, X212, X213, NP1524, X155, X41 |
| 5 | NP3706, NP1518, X218, X220, NP1527, X275 |
| 6 | NP1517, NP1516, NP1507 |

Табл. 1 Результат розбиття на кластери

|  |  |  |
| --- | --- | --- |
| Групи | Під групи | Елементи |
| 1 | A | NP1501, NP1514, NP1516, NP1517, NP1507 |
| B | X190, NP1526, X200, X201 |
| 1B/2 | | NP1512, NP1521 |
| 2 | | NP1502, NP1503, NP1508, NP1510, NP1520, NP1522, NP1525, X221, X271, NP3701, NP3708 |
| 2B/3 | | NP1528 |
| 3 | A | X202, X218, NP1518, NP1527 |
| B | X32, X155, X41, X212, X213, X217, X223, X224, X233, NP1524, NP3715 |
| 4 | A | NP3706 |
| B | X211 |
| C | X215 |
| 5 | | X220 |
| 6 | | X275 |

Табл. 2 Очікуваний результат розбиття

Можна зауважити, що кластер 4 це те саме, що група 3B, кластери 0,1,6 утворюють групу 1 в об’єднанні з групою 1B/2 з точністю 10 з 11 елементів. Також група 2 утворю групу 2 в об’єднанні з групою 2B/3 з точністю 10 з 12 елементів. Також можна зауважити, що група 5 пов’язана з групою 3А.

В результаті кількість елементів що опинились в однакових кластерах буде 34 з 43, а це означає, що точність становить 79%.

**Список використаних джерел**

1. Satyam Kumar Clustering Algorithm for data with mixed Categorical and Numerical features URL: <https://towardsdatascience.com/clustering-algorithm-for-data-with-mixed-categorical-and-numerical-features-d4e3a48066a0>
2. Python: K-modes explanation

<https://stackoverflow.com/questions/42639824/python-k-modes-explanation>