

РОЗРОБКА СИТЕМИ КЛАСТЕРИЗАЦІЇ АНТИТІЛ НА ОСНОВІ КОЕФІЦІЄНТУ ПЕРЕХРЕСНОГО ЗВ'ЯЗУВАННЯ

Олександр Зелінський, Віталій Горлач, Юрій Лебедін
Факультет прикладної математики та інформатики
Львівський національний університет імені Івана Франка
Oleksandr.Zelinskyi@lnu.edu.ua

В умовах пандемії надзвичайно важливими є дослідження які несуть безпосередню користь для виявлення, запобігання та лікування вірусних захворювань, а якщо точніше 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 два круги однакового розміру, що приєднуються до меншого круга який представляє молекулу вірусу.

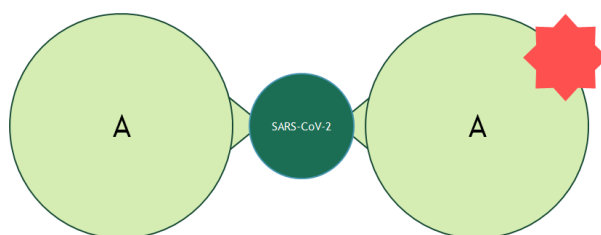


Рис. 1 Модель приєднання антитіл до вірусної молекули

Задача полягає у знаходженні двох антитіл (можуть бути однаковими), таких що 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 для відповідного міченого антитіла.

Labelled	NP1501*	NP1502*	NP1503*	NP1508*	NP1510*	NP1514*	NP1516*	NP1517*
blank	1.089	1.067	1.3664	1.412	1.67	1.07	1.1704	1.11
NP1501	0.449	0.715	1.0664	1.0248	1.136	0.26	0.4172	0.268
NP1502	0.893	0.425	0.336	0.196	0.349	0.3625	0.665	0.45
NP1503	0.768	0.309	0.068	0.052	0.095	0.305	0.7126	0.408
NP1507	0.422	0.856	0.7216	0.6088	0.785	0.1825	0.4256	0.154
NP1508	0.732	0.388	0.1456	0.0848	0.19	0.345	0.8456	0.411
NP1510	0.781	0.382	0.2152	0.1056	0.233	0.495	0.7826	0.527
NP1512	0.79	0.789	0.876	0.7504	0.979	0.735	1.015	0.737
NP1514	0.448	0.822	1.0968	1.0088	1.189	0.2475	0.4858	0.253
NP1516	0.385	1.034	0.9832	0.9304	1.053	0.1775	0.3052	0.152
NP1517	0.425	0.644	0.7952	0.7304	0.885	0.155	0.2758	0.079

Рис. 2 Фрагмент початкових даних

Для подальшої роботи з даними їх позначають за наступним алгоритмом:

1. Цифрою 3 (темно-зеленим кольором) антитіла з хорошим зв'язуванням, якщо

$$\frac{-(cell[i][j]-blank[j])}{blank[j]} > 0.75 \quad (1)$$

2. Цифрою 2 (світло-зеленим кольором) антитіла з середнім зв'язуванням, якщо

$$0.5 < \frac{-(cell[i][j]-blank[j])}{blank[j]} \leq 0.75 \quad (2)$$

3. Цифрою 1 (білим кольором) антитіла майже без зв'язування, у всіх інших випадках.

label	NP1501*	NP1502*	NP1503*	NP1508*	NP1510*	NP1514*	NP1516*	NP1517*
NP1501	2	1	1	1	1	3	2	3
NP1502	1	2	3	3	3	2	1	2
NP1503	1	2	3	3	3	2	1	2
NP1507	2	1	1	2	2	3	2	3
NP1508	1	2	3	3	3	2	1	2
NP1510	1	2	3	3	3	2	1	2
NP1512	1	1	1	1	1	1	1	1
NP1514	2	1	1	1	1	3	2	3
NP1516	2	1	1	1	1	3	2	3
NP1517	2	1	1	1	1	3	3	3

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-means і використовується для кластеризації даних на основі якісних змінних. k-modes визначає кластери на основі відповідності категорій між точками даних. В даному алгоритмі відстань між двома точками даних X та Y описується як сума не схожих елементів:

$$d_1(X, Y) = \sum_{i=1}^n \delta(x_i, y_i), \text{ де } \delta(x_i, y_i) = \begin{cases} 0, & x_i = y_i \\ 1, & x_i \neq y_i \end{cases} \quad (3)$$

Для визначення оптимальної кількості кластерів використовується elbow метод, який для різних значень k буде вибирати значення k у тій точці, де значення істотно не зменшується зі збільшенням значення k.

Для обробки даних та кластеризації використовувалась бібліотека kmodes, pandas, matplotlib та kneed з Python. В результаті виконання програми отримано оптимальне розбиття на 9 кластерів.

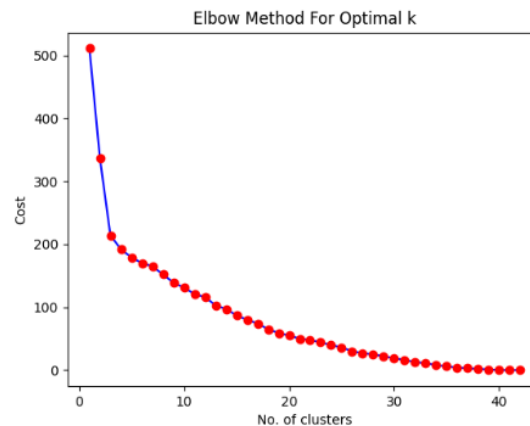


Рис. 4 Фрагмент позначених даних

Antigen by groups					
Group I	Group II	Group III	Group IV	Group V	Group VI
NP1501, NP1512, NP1514, NP1516, NP1517, NP1521, X211, X215, X218, X220, X275	NP1528	NP1502, NP1503, NP1508, NP1510, NP1520, NP1522, NP1525, NP3708, X190, X221, X271	NP1524, NP3715, X212, X217, X223, X224	NP1526, NP3701, X200, X201	NP1507
Group VII	Group VIII	Group IX			
X155, X202, X213, X233, X32, X41	NP1518	NP1527, NP3706			

Рис. 5 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. 1 Очікуваний результат розбиття

З результатів видно, що група I майже відповідає групі 1A в об'єднанні з 1B/2 та 4B, 4C, 5 і 6. Група III майже відповідає групі 2, Група II відповідає групі 2B/3. Група 3B відповідає групі VII та групі IV. Зважаючи на те що кількість елементів які мають бути в однакових групах 30 з 43 елементів то можна вважати, що похибка становить близько 30%.

На рисунках 6 та 7 наведена візуалізація реального та очікуваного розбиття на групи за допомогою кольорів.



Рис. 6 Візуалізація результатів розбиття

Labelled	NP1501	NP1514	NP1516	NP1517	X200	NP1521	X201	X217	X221	NP1520	NP1522	NP1502	NP1525	NP1503	NP1508	NP1510	NP1527	X202	NP1518	NP3715	X217	X155	X213	NP1524	X223	X233	X41	X32	NP3706	X211	X215	X220	
Group 1A	1.089	1.087	1.364	1.412	1.07	1.07	1.1704	1.11	1.1816	1.007	1.004	1.3702	1.3708	1.302	1.417	1.0303	1.1395	1.672	1.36	1.4525	1.4364	1.789	1.1844	1.422	1.3617	1.105	1.3671	1.404	1.3036	1.4688	1.234	1.412	
NP1501	0.440	0.38	0.4172	0.38	0.6978	0.596	1.095	1.228	0.8956	0.814	0.7326	0.715	1.061	1.0604	1.0248	1.16	0.852	0.5316	0.8512	1.0706	1.0224	1.3395	1.2519	1.1392	1.4616	0.769	0.7912	1.303	0.914	1.2287	0.64	1.189	
NP1514	0.446	0.4976	0.4052	0.29	0.891	0.625	1.038	0.564	0.8712	1.04	0.7821	0.822	1.136	1.0960	1.0095	1.180	1.024	0.6816	1.0306	1.1725	1.0413	1.4196	1.224	1.2512	1.4616	0.772	0.9076	1.364	0.84	1.2364	0.975	1.332	
NP1516	0.385	0.1779	0.5032	0.175	0.6006	0.405	1.063	0.874	0.7392	0.782	0.6399	1.034	0.99	0.9832	0.9304	1.053	0.752	0.5016	0.9504	1.0682	1.0116	1.3013	1.1844	1.0396	1.4544	0.878	0.8792	1.485	0.7833	1.2078	0.9502	1.281	
NP1517	0.425	0.159	0.2878	0.675	0.1103	0.416	1.121	1.011	0.7209	0.836	0.5427	0.844	0.92	0.7392	0.7304	0.885	0.71	0.5016	1.1504	1.0563	1.0179	1.3006	1.1801	1.0824	1.4376	0.882	0.8504	1.368	0.8673	1.1645	0.9075	1.245	
NP1507	0.425	0.4976	0.4052	0.675	0.6459	0.503	0.983	0.847	0.6024	0.462	0.5335	0.558	0.9336	0.9726	0.8698	0.785	0.839	0.6426	1.0332	1.0914	1.0215	1.3111	1.2872	1.1104	1.4168	1.018	0.8992	1.388	0.8862	1.2213	0.9126	1.254	
Group 1B	X150	0.442	0.3375	0.4432	0.239	0.6846	0.349	0.188	0.385	0.5818	0.937	0.46	0.36	0.568	0.6976	0.6445	0.58	0.62	0.4344	0.364	0.8638	0.9054	1.2005	0.999	0.9232	1.3182	0.895	0.696	1.011	0.7381	1.2811	0.539	1.188
NP1528	0.415	0.415	0.4074	0.484	0.394	0.445	0.404	0.404	0.348	0.404	0.404	0.404	0.404	0.404	0.404	0.404	0.404	0.404	0.404	0.404	0.404	0.404	0.404	0.404	0.404	0.404	0.404	0.404	0.404	0.404	0.404	0.404	
X200	0.369	0.39	0.4464	0.417	0.716	0.346	0.401	0.401	0.401	0.401	0.401	0.401	0.401	0.401	0.401	0.401	0.401	0.401	0.401	0.401	0.401	0.401	0.401	0.401	0.401	0.401	0.401	0.401	0.401	0.401	0.401	0.401	
X201	0.445	0.3375	0.4432	0.435	0.1817	0.389	0.389	0.512	0.3545	0.481	0.512	0.421	0.512	0.512	0.512	0.512	0.512	0.512	0.512	0.512	0.512	0.512	0.512	0.512	0.512	0.512	0.512	0.512	0.512	0.512	0.512	0.512	
NP1521	0.538	0.4307	0.5342	0.574	0.7209	0.58	1.287	1.281	0.8514	0.851	0.729	0.628	1.058	0.9084	0.9016	1.057	0.754	0.544	1.0248	1.0556	0.972	1.2529	1.2168	1.0872	1.456	0.964	0.8184	1.786	1.071	1.2006	0.9775	1.209	
Group 2	X217	0.489	0.5375	0.4844	0.538	0.6938	0.399	0.361	0.111	0.4422	0.108	0.6711	0.573	0.884	0.6968	0.6502	0.168	0.81	0.6784	0.5472	0.8666	0.8386	1.2565	1.0233	0.8736	1.412	0.815	0.6292	1.335	0.8987	1.2942	0.58	1.174
X221	0.646	0.16	0.8814	0.603	0.488	0.446	0.33	0.2975	0.2644	0.237	0.1719	0.264	0.289	0.2773	0.1916	0.268	1.348	0.2646	0.9332	0.9744	1.0206	1.2095	1.1801	0.9912	1.336	0.954	0.78	1.44	0.8274	1.7937	0.9915	1.166	
NP1520	0.685	0.315	0.5886	0.436	0.134	0.431	0.168	0.109	0.1179	0.505	0.505	0.505	0.505	0.505	0.505	0.505	0.505	0.505	0.505	0.505	0.505	0.505	0.505	0.505	0.505	0.505	0.505	0.505	0.505	0.505	0.505	0.505	
NP1522	0.689	0.472	0.7476	0.374	0.154	0.641	0.177	0.375	0.162	0.205	0.1143	0.384	0.108	0.1668	0.0912	0.38	0.795	0.87	1.0064	1.0633	1.0125	1.3265	1.1916	1.0872	1.3704	0.832	0.8528	1.809	0.9093	1.1529	0.430	1.187	
NP1502	0.891	0.403	0.588	1.44	0.299	0.53	0.488	0.447	0.2979	0.986	0.119	0.421	0.6268	0.409	0.196	0.149	0.822	0.4942	1	1.081	1.0393	1.2024	1.2753	1.1584	1.4184	0.809	0.9182	1.456	0.9323	1.2687	0.36	1.203	
NP1507	0.968	0.31	0.4844	0.437	0.1103	0.268	0.421	0.545	0.4311	0.32	0.47	0.402	0.362	0.3094	0.3196	0.387	0.646	0.442	1.0706	1.018	1.1493	1.3053	1.1448	1.0072	1.388	1.129	0.8024	1.722	0.8715	1.1467	0.435	1.248	
NP1525	1.003	0.765	0.9506	0.67	0.166	0.677	0.362	0.11	0.18	0.186	0.128	0.388	0.103	0.1372	0.0872	0.148	0.872	0.8168	0.9728	0.9821	0.8684	1.309	1.1178	0.892	1.2152	0.604	0.7928	1.984	0.9429	1.2915	0.7625	1.313	
NP1503	0.793	0.382	0.7126	0.448	0.059	0.57	0.144	0.108	0.555	0.575	0.061	0.558	0.076	0.068	0.0302	0.089	0.855	0.442	0.9889	1.1179	1.0791	1.3114	1.2392	1.1234	1.4704	0.93	0.8336	1.712	0.8165	1.197	0.6075	1.216	
NP1508	0.732	0.382	0.8436	0.411	0.376	0.36	0.17	0.333	0.287	0.144	0.161	0.198	0.195	0.1498	0.0644	0.18	0.858	0.6482	1.1344	1.1074	1.0095	1.3356	1.2899	1.1632	1.378	1.052	0.9202	1.771	0.8736	1.2474	0.9075	1.181	
NP1510	0.781	0.448	0.7626	0.527	0.1636	0.661	0.163	0.23	0.2381	0.227	0.1335	0.382	0.146	0.2152	0.1006	0.233	0.836	0.8744	1	1.0542	0.9513	1.3503	1.1934	1.1216	1.42	0.842	0.9192	2.086	0.8505	1.1547	0.975	1.117	
NP3708	0.678	0.38	0.392	0.575	0.444	0.624	0.261	0.195	0.2653	0.257	0.1618	0.177	0.168	0.1676	0.1104	0.245	0.735	0.9522	0.9568	0.9695	0.8982	1.2891	1.1718	0.9704	1.306	0.952	0.8332	1.875	0.8484	1.0599	0.645	1.249	
Group 2A	NP1512	0.78	0.785	1.015	0.737	0.641	0.633	0.639	0.637	0.6384	0.633	0.7181	0.789	0.836	0.876	0.7954	0.876	0.841	0.801	0.818	0.9975	0.8332	1.4441	1.0281	1.1086	1.3544	1.018	0.88	1.032	0.7623	1.1826	0.7435	1.123
Group 2B3	NP1528	1.101	0.3075	0.5874	0.442	0.169	0.275	0.364	0.266	0.266	0.266	0.266	0.266	0.266	0.266	0.266	0.266	0.266	0.266	0.266	0.266	0.266	0.266	0.266	0.266	0.266	0.266	0.266	0.266	0.266	0.266	0.266	
Group 3A	NP1527	0.719	0.605	0.59	0.594	0.6273	0.488	1.392	1.254	1.0125	0.771	0.7303	0.707	1.101	1.012	0.9168	1.023	0.588	0.596	0.808	0.8008	0.8739	0.9359	0.9331	0.8792	1.2168	0.735	0.9688	1.494	0.8547	0.941	0.9029	1.148
X218	0.686	0.4737	0.4956	0.623	0.7719	0.18	1.038	1.407	1.1110	0.968	0.9672	0.729	1.279	1.1096	1.1584	1.228	0.447	0.16	0.8466	0.8208	0.9207	0.9016	0.8631	0.8215	1.2742	0.468	0.3468	0.733	0.6025	1.4623	0.4875	1.241	
X202	0.745	0.6025	0.6776	0.692	0.7614	0.506	1.419	1.171	0.9627	0.982	0.864	0.733	1.386	1.178	1.088	1.054	1.178	0.288	0.344	0.8544	0.9384	0.9488	0.9397	0.873	0.8284	0.9068	0.928	0.344	0.614	0.6216	0.9064	0.5395	0.975
NP1518	0.517	0.544	0.5052	0.34	0.6174	0.358	1.527	1.121	0.8894	0.787	0.753	0.638	1.051	1.0272	0.9424	1.107	0.934	0.413	0.5718	0.6355	0.5188	0.8643	0.7096	0.548	0.8336	0.345	0.150	1.452	0.8132	0.9297	0.4975	1.083	
Group 3B	NP3715	0.603	0.41	0.7685	0.554	0.4437	0.612	1.164	0.896	0.8905	0.855	0.533	0.533	0.729	0.7096	0.6112	0.848	0.768	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76
X217	0.476	0.5325	0.4844	0.409	0.6988	0.399	0.186	0.243	0.7985	0.733	1.124	0.978	0.956	1.072	0.738	0.3016	0.308	0.6053	0.1608	0.168	0.1987	0.9846	0.9376	0.9846	0.9376	0.9846	0.9376	0.9846	0.9376	0.9846	0.9376	0.9846	0.9376
X155	0.364	0.6828	0.325	0.562	0.6616	0.397	1.132	1.124	0.8721	0.763	0.6948	0.718	1.059	0.9904	0.8832	1.049	0.592	0.1644	0.5963	0.9791	0.1163	0.9742	0.9490	0.9344	1.083	0.9795	0.474	0.4688	0.9748	0.44	1.026		
X213	0.646	0.16	0.8814	0.603	0.488	0.446	0.33	0.2975	0.2644	0.237	0.1719	0.264	0.289	0.2773	0.1916	0.268	1.348	0.2646	0.9332	0.9744	1.0206	1.2095	1.1801	0.9912	1.336	0.954	0.78	1.44	0.8274	1.7937	0.9915	1.166	
NP1524	0.604	0.7425	0.9028	0.813	0.6055	0.675	1.358	0.84	0.7317	1.185	0.8163	0.948	1.123	1.128	1.0584	1.194	0.712	0.168	0.438	0.9416	0.9484	0.1005	0.1044	0.9725	0.159	0.975	0.9488	0.344	0.6022	0.8008	0.63	0.888	
X212	0.712	0.87	0.9436	0.872	1.0279	0.664	1.739	1.114	1.0827	1.268	0.972	0.805	1.408	1.2036	1.2192	1.414	1.54	0.1644	0.2224	0.9489	0.9463	0.1623	0.9463	0.9728	0.9568	0.161	0.9689	0.916	0.8085	1.9539	0.81	0.952	
X223	0.694	0.748	0.7616	0.649	0.7628	0.539	0.846	0.35	0.8253	1.113	0.7488	0.79	1.14	1.0044	1.0096	1.103	0.619	0.1193	0.3036	0.9416	0.9999	0.9998	0.9998										