*Во время работы над задачами в области структурной биоинформатики постоянно приходится работать с различными специфическими данными, которые требуют особого внимания и осторожности. Для погружения в предметную область предлагается в течение тестового задания познакомиться с базой данных Sabdab (The Structural Antibody Database), взять оттуда человеческие комплексы (антитело-антиген) и ответить на несколько вопросов.*

*1. Предложите способ кластеризации антител в таких комплексах:*

**Антитела являются одним из наиболее важных классов фармацевтических препаратов, которые в настоящее время используются против широкого спектра заболеваний. Однако процесс открытия лекарств для терапевтических кандидатов на антитела требует много времени и средств и в значительной степени зависит от высокопроизводительных скринингов in vivo и in vitro. Кластеризация антител способна ускорить процесс поиска кандидатов против интересующих антигенных мишеней. В литературе уже описано достаточно много подходов к этой кластеризации, однако их можно разделить на 2 типа:**

*- по последовательности антитела (или каких-то его регионов);  
- по структуре антитела в области связывания с антигеном.*

1. **В литературе широко представлены подходы к машинному обучению в области кластеризации терапевтических антител на их последовательности (Schneider et all., 2022). Было показано, что эти подходы машинного обучения очень эффективны для задач прогнозирования, которые зависят только от антитела (а не от интерфейса антитело-антиген) или специфичны для одного антигена. Современный инструмент прогнозирования паратопов - Parapred показал хорошие результаты кластеризации и основан на рекуррентном нейросетевом подходе (RNN) (Liberis et al., 2018). Метод требует только отрезка аминокислотной последовательности, соответствующей гипервариабельной области (CDR), в качестве входных данных без какой-либо информации об антигене. Тем не менее, для препроцессинга данных я бы выбрал подход, описанный в статье об афинности связывания антитело-антиген (Nathaniel, et all., 2023) как наиболее современный и всеобъемлющий. Он предполагает выбор комплексов антитело-антиген при условии, что последовательности составляют более 50 аминокислот, применяются так же шаги для избавления от избыточности и низкого «сродства», оставляя комплексы со средним сродством 3 нМ. Далее предварительная обработка комплекса антитело-антиген состоит из: 1) перенумерации последовательностей по схеме нумерации Chothia, 2) удаления молекул воды и лигандов, 3) замены и переупаковки отсутствующих атомов боковой цепи в PyRosetta. Затем на основе опубликованных алгоритмов были выделены признаки: dMaSIF, SIN, AIF (более подробно смотреть в статье и ноутбуке с реализацией). Для работы с такими данными (pdb) понадобятся пакеты, разработанные специально для биостатистики, такие как biopandas, biopython и другие. В результате мы получаем список словарей с тензорами (torch) из оценок для каждой аминокислоты в каждом комплексе. Далее я бы выбрал только последовательности CDR: H1, H2, H3, L1, L2, L3. Далее, поскольку их распределение приближено к нормальному, а последовательности различаются по длине, я бы не использовал заполнение, поскольку это нарушает структуру последовательностей, а использовал показатели распределения: такие как среднее, квантили, стандартное отклонение, коэффициент вариации, дисперсия. По этой последовательности одинаковых по длине массивов строил кластеризацию. В литературе предпочтение при расчете матрицы расстояния отдают алгоритму DTW, поэтому я бы остановился на нем, но попробовал разные алгоритмы.**
2. **Если рассматривать алгоритм кластеризации по структуре в области связывания антитела с антигеном, то в этом случае полезным будет так же определение паротопов. Таким образом, предобработка и выделение сайтов связывание будет той же самой. Методологию дальнейшего выделения фич я бы взял из статьи о кластеризации сайтов связывания на основе моментов Цернике (Rienzo et all., 2017). Алгоритм основан на расчете гауссовской поверхности, вокселизации и вычислении 3D коэффициентов Цернике, описывающих поверхность связывания антител и антигена. В результате получается, что каждый сайт описывается 121 дескриптором, т.е. длина тензоров равная. Далее можно применять различные методы кластеризации, например, можно применить метод иерархической кластеризации с выбранным алгоритмом и метрикой (к примеру, single linkage и евклидово расстояние)**

*2. Реализуйте один из предложенных способов кластеризации на выбор. Можно использовать любые open-source инструменты, но позаботьтесь о воспроизводимости ваших действий.*

**Я реализовал первый способ: на основе последовательности антител в регионах CDR по методике (Nathaniel, et all., 2023). Jupyter notebook с реализацией находится по ссылке:**

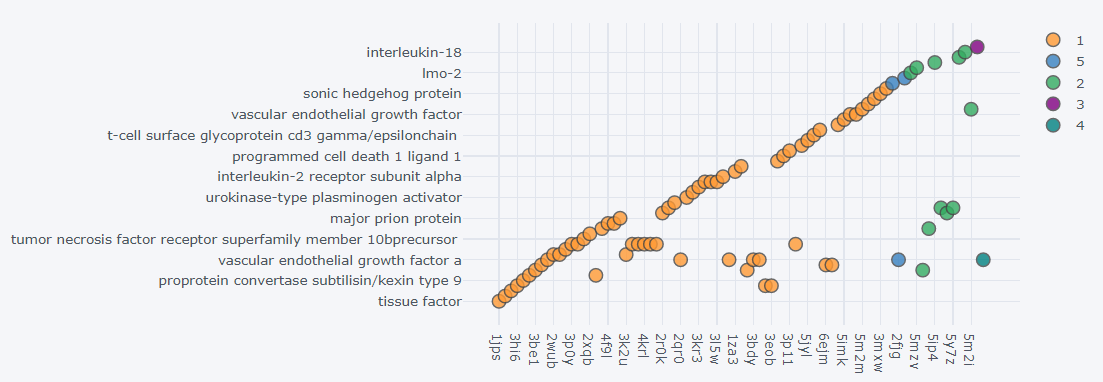
<https://www.kaggle.com/code/akscent/antibody-clustering/notebook>

*3. Проанализируйте полученные кластеры:*

*- хорошо ли данные делятся на кластеры, интерпретируем ли результат?*

*- приведите графики и таблицы, которые на ваш взгляд наиболее интересны.*

**В результате было выделено 5 кластеров. В целом, результаты не плохие, поскольку в одни и те же кластеры были объеденены антитела со схожими по своей природе антигенами.**



**Список источников:**

1. *C. Schneider, A. Buchanan, B. Taddese, C. M Deane, DLAB: deep learning methods for structure-based virtual screening of antibodies, Bioinformatics, Volume 38, Issue 2, January 2022, Pages 377–383,* [*https://doi.org/10.1093/bioinformatics/btab660*](https://doi.org/10.1093/bioinformatics/btab660)
2. *E. Liberis, P. Veličković, P. Sormanni, M. Vendruscolo, P. Liò, Parapred: antibody paratope prediction using convolutional and recurrent neural networks, Bioinformatics, Volume 34, Issue 17, 01 September 2018, Pages 2944–2950,* [*https://doi.org/10.1093/bioinformatics/bty305*](https://doi.org/10.1093/bioinformatics/bty305)
3. *H. Yan, Z. Ziding, Z. Yuan, AbAgIntPre: A deep learning method for predicting antibody-antigen interactions based on sequence information, Frontiers in Immunology, V. 13, 2022. doi: 10.3389/fimmu.2022.1053617*
4. *M. Nathaniel L., C. Thomas, R. Rahul, S. Ram, Learned features of antibody-antigen binding affinity, Frontiers in Molecular Biosciences, V. 10, 2023. doi: 10.3389/fmolb.2023.1112738*
5. *Di Rienzo, L., Milanetti, E., Lepore, R. et al. Superposition-free comparison and clustering of antibody binding sites: implications for the prediction of the nature of their antigen. Sci Rep 7, 45053 (2017). https://doi.org/10.1038/srep45053*