Projekt 02: Odkrywanie konserwatywnych wzorców w rezultatach próbkowania miejsc wiązań RNA-białko

Zespół:

Max Adamski
Sławomir Gilewski
Patryk Jedlikowski
Mikołaj Sienkiewicz

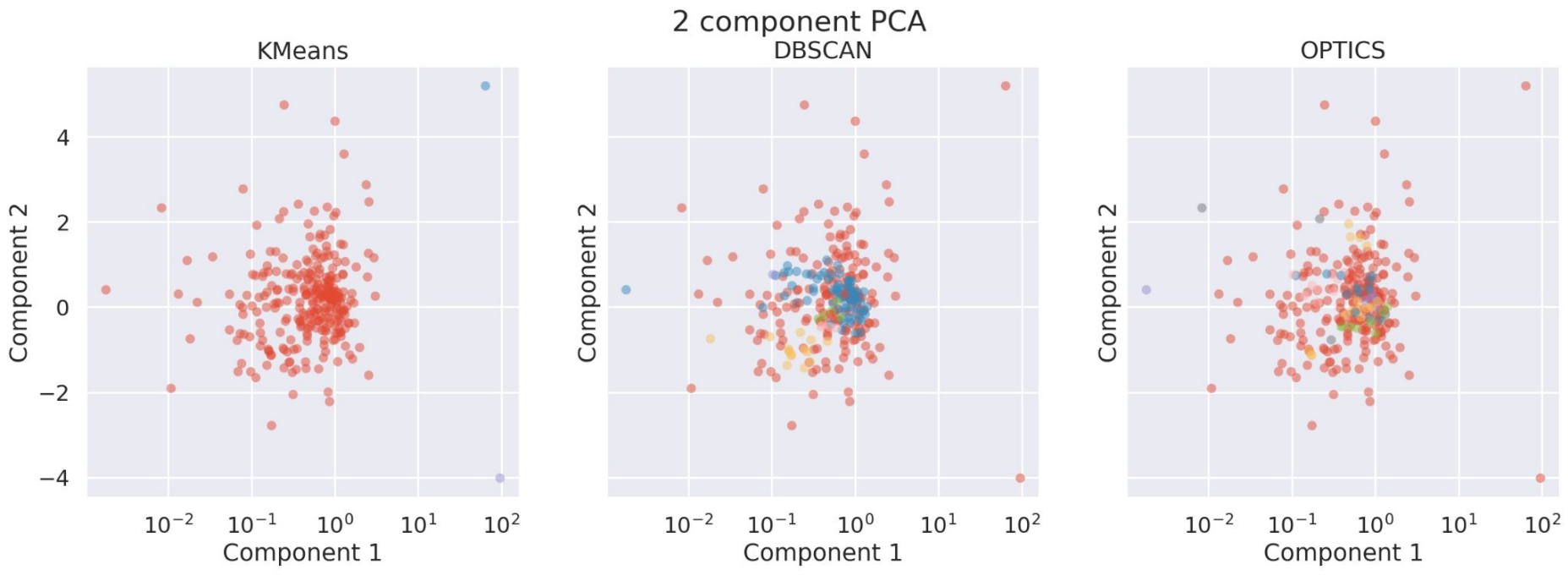
Wybrane białko:

hnrnpc

Zaproponowane rozwiązanie:

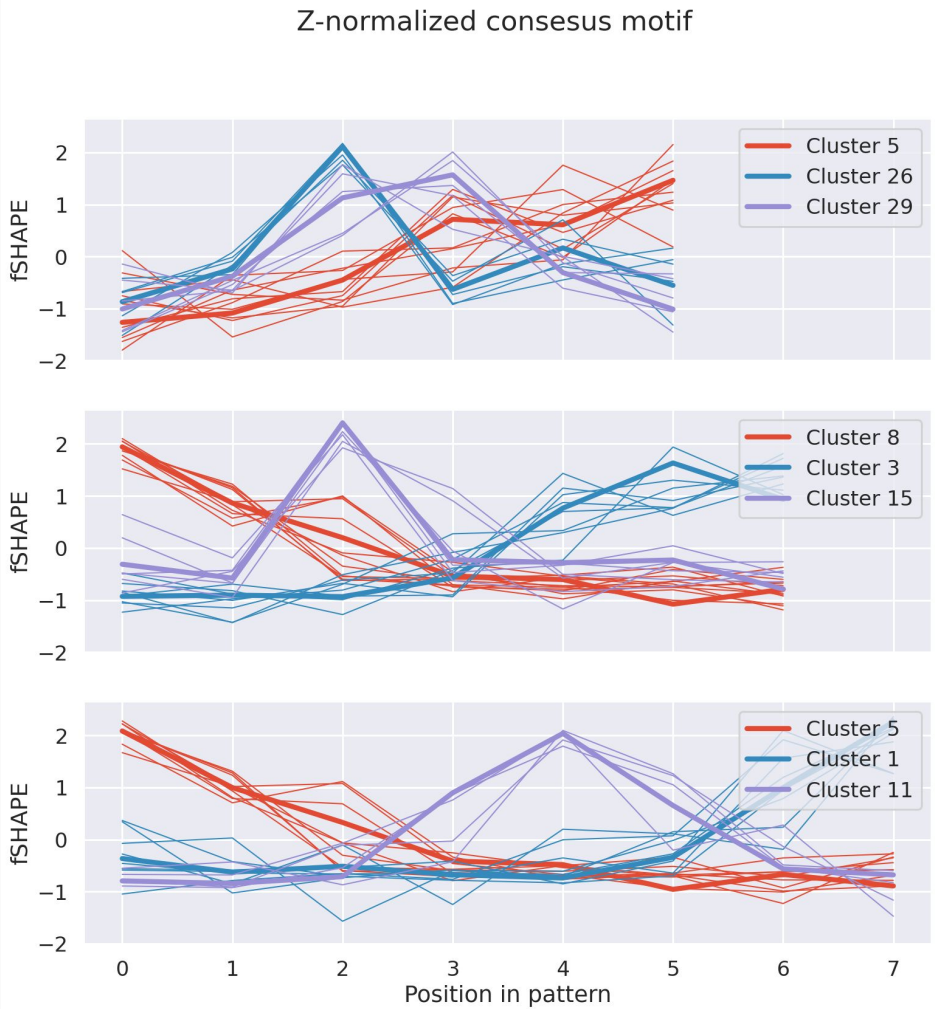
- **Wybór białka**
- Przygotowanie różnych **wielkości okna** na bazie szablonu białka
- Wyciągnięcie obiecujących motywów (żadna wartość w oknie nie może być **NaN** i co najmniej jeden gen musi mieć **fSHAPE > 1.0**
- Analiza skupień - Wizualizacja PCA algorytmów **KMeans, DBSCAN, OPTICS**
- Użycie ostinato ze stumpy do znalezienia **motywów konsensusowych** dla 3 największych klastrów wygenerowanych algorytmem OPTICS
- Porównanie motywów konsensusowych z danymi sekwencjami genów

Efekty analizy skupień



Wizualizacja klastrów (motyw konsensusowy grubszą linią)

```
Top clusters for length 6: [ 5 26 29]
Consensus motif in cluster 05: CATTTT
Consensus motif in cluster 26: TAAAAT
Consensus motif in cluster 29: ATTGTT
Top clusters for length 7: [ 8  3 15]
Consensus motif in cluster 08: AAACAGA
Consensus motif in cluster 03: CCGGTGT
Consensus motif in cluster 15: TCACCTC
Top clusters for length 8: [ 5  1 11]
Consensus motif in cluster 05: AAACAGAC
Consensus motif in cluster 01: AATGGTAA
Consensus motif in cluster 11: AGCTGAGC
```



Wyniki:

Dla CATTTT (konsensus)

1	sequence	file	start	end	znED	ssf	aS
2	GGACTT	hnrnpc_NM_001632_ALPP.txt	2267	2272	0.1604	1.1667	0.4369
3	GTTTCG	hnrnpc_NM_005627_SGK1.txt	1658	1663	0.2949	1.8333	1.1159
4	TTCTTT	hnrnpc_NM_001136025_PLS3.txt	2560	2565	0.3333	1.8333	1.4999
5	GGCTTA	hnrnpc_NM_001631_ALPI.txt	1114	1119	0.3330	1.5000	1.8299
6	CCTGGT	hnrnpc_NM_001195129_PRSS56.txt	535	540	0.4312	1.1667	3.1454
7	CTCATT	hnrnpc_NM_054012_ASS1.txt	1172	1177	0.4681	1.5000	3.1811
8	GCTGAA	hnrnpc_NM_001875_CPS1.txt	4267	4272	0.4541	1.1667	3.3745
9	GGGCCA	hnrnpc_NM_001311_CRIP1.txt	385	390	0.4511	1.0000	3.5107
10	CTTTAA	hnrnpc_NM_001136025_PLS3.txt	129	134	0.5194	1.6667	3.5274
11	GCCCTG	hnrnpc_NM_001632_ALPP.txt	2109	2114	0.5053	1.5000	3.5528
12	TCCTCT	hnrnpc_NM_000067_CA2.txt	780	785	0.5070	1.5000	3.5702
13	CCTTCT	hnrnpc_NM_001195129_PRSS56.txt	1236	1241	0.5295	1.6667	3.6287

https://docs.google.com/spreadsheets/d/1J3y-B-xJkl1StjOB3KmFFLQu7cVzBMbfSOyOGx6i5Ok/edit?usp=sharing