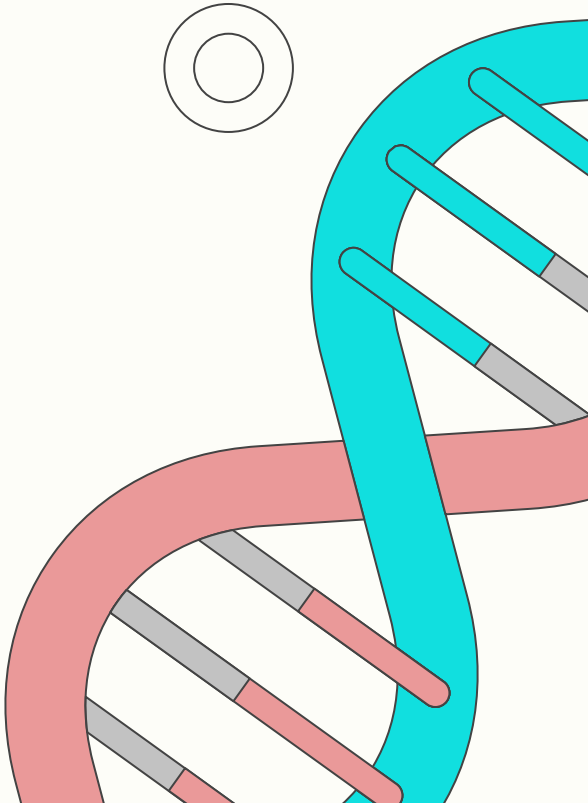
Podział genomu na domeny topologiczne

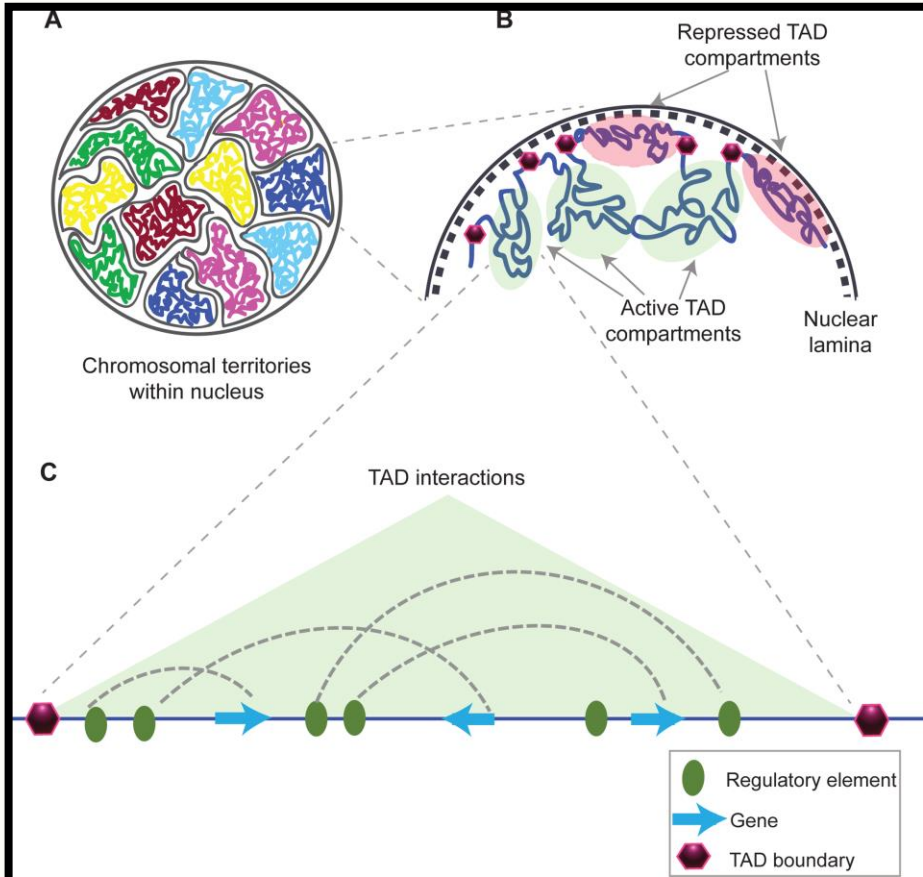
Autorzy: Sebastian Pergała, Aleksandra Samsel



Wprowadzenie

Domeny topologiczne (TADs - Topologically Associating Domains)

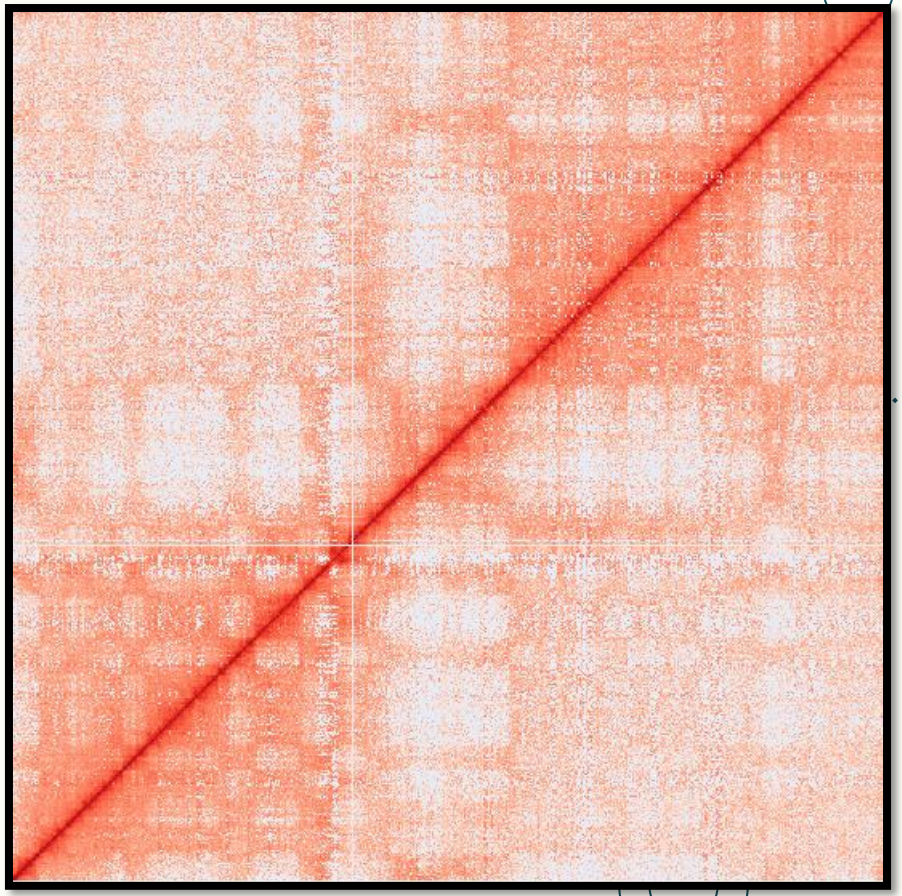
To regiony genomu, wchodzące w częste interakcje same ze sobą. Oznacza to, że sekwencje DNA znajdujące się wewnątrz domeny mają ze sobą więcej fizycznych interakcji, niż z sekwencjami poza nią.



Macierz interakcji

Macierz przedstawiająca interakcje między regionami chromatyny.

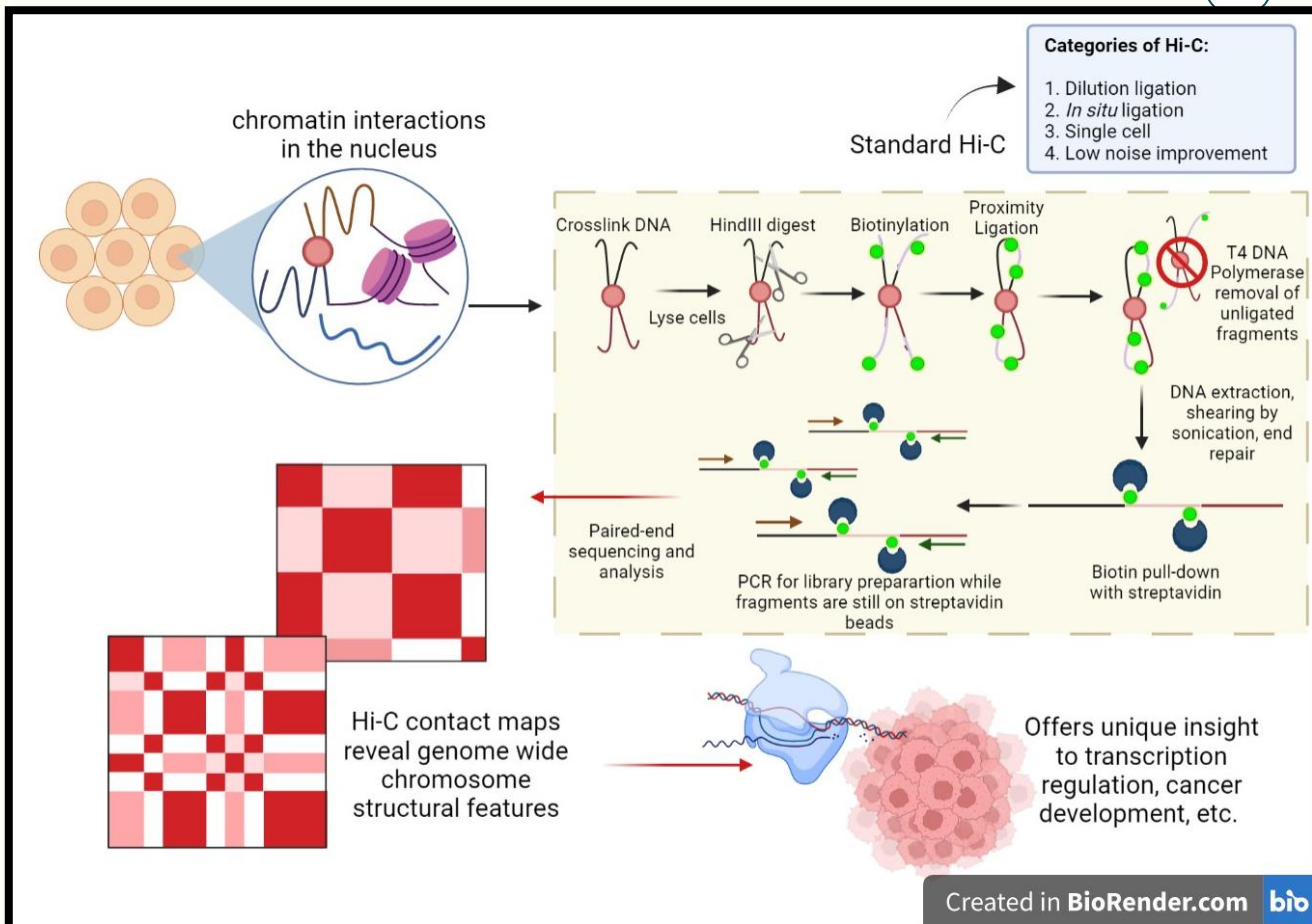
Fragmenty tworzące ciemniejsze kwadraty na macierzy to topologicznie sąsiadujące domeny. W związku z brakiem ogólnie przyjętej definicji domen, ich postać jest uznaniowa.



Hi-C

Technika zbierania informacji na temat konformacji chromatyny.

Zlicza częstotliwość (jako średnia na całą populację komórki) fizycznych kontaktów segmentów DNA, łącząc ze sobą trójwymiarową strukturę chromosomów i sekwencje DNA.

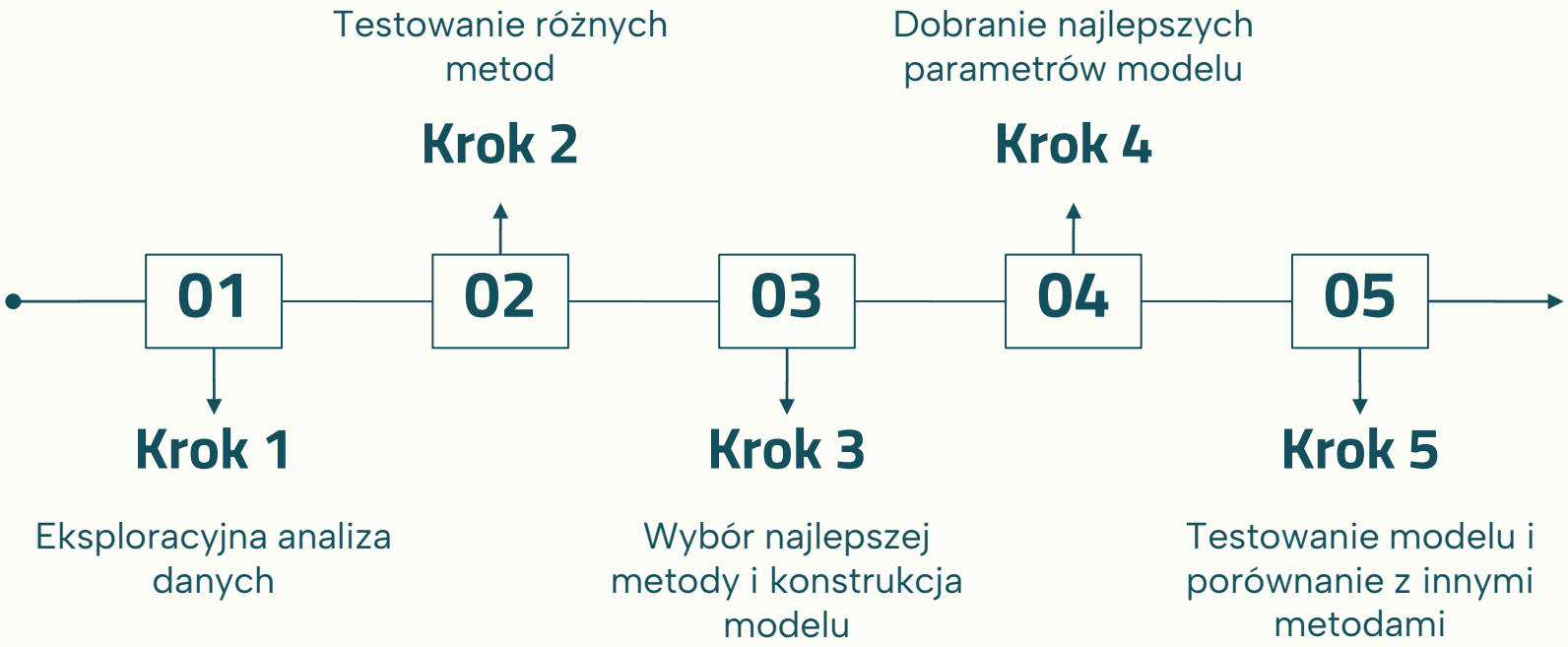


Cel biznesowy

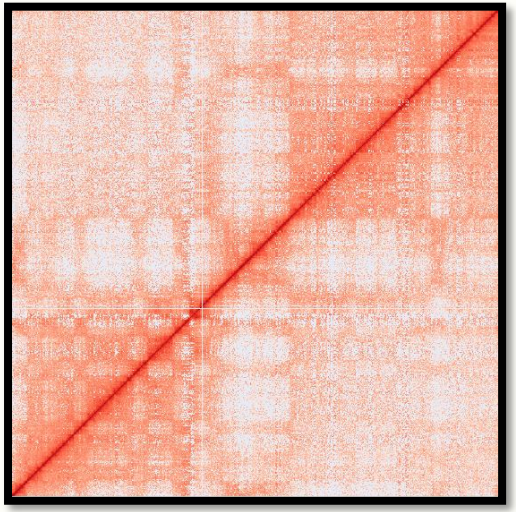
Stworzenie stabilnego (otrzymującego podobne wyniki na danych o różnej rozdzielczości) modelu do znajdywania domen topologicznych w macierzach interakcji z danych pozyskanych metodą Hi-C.

Metodologia

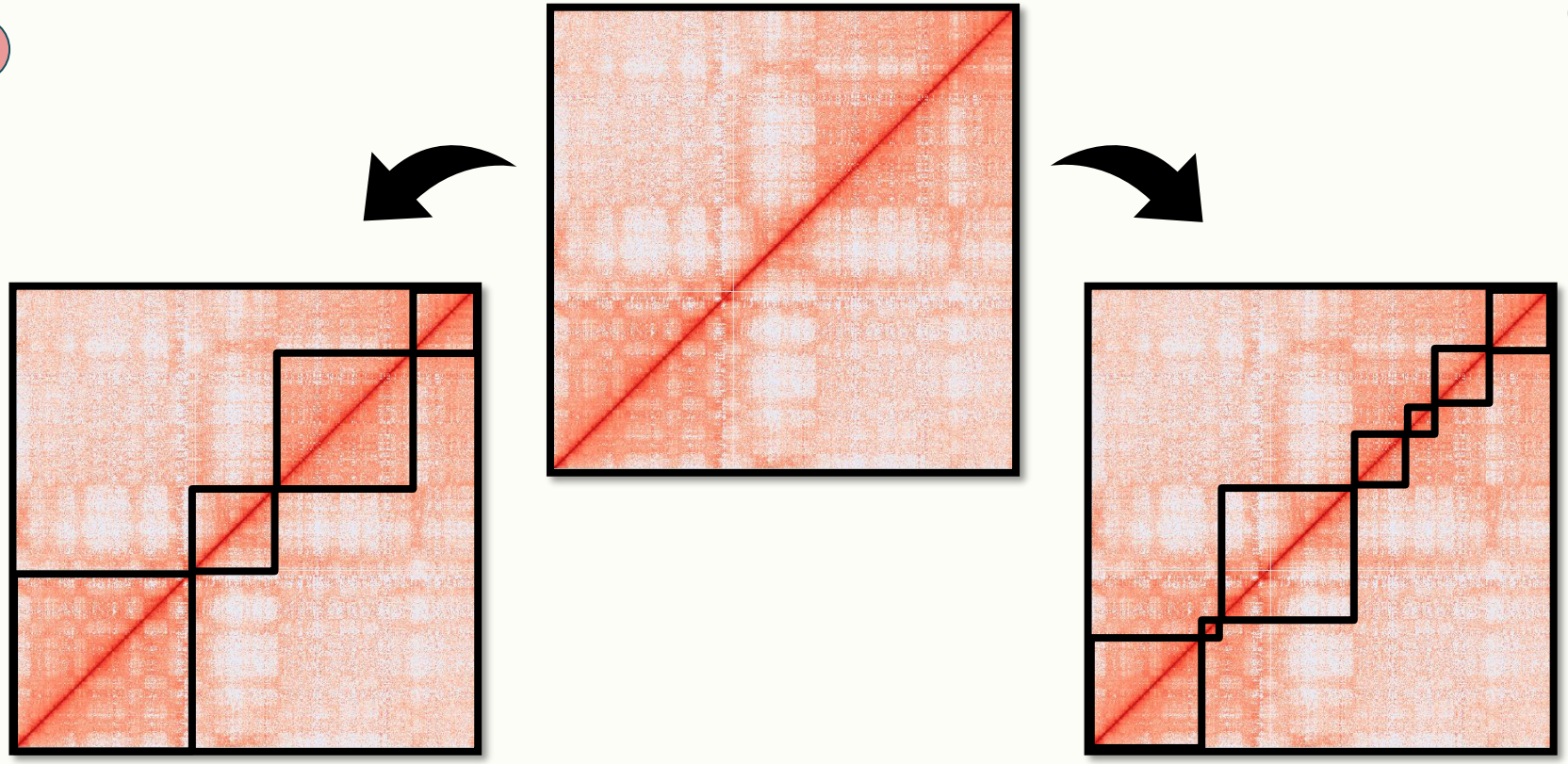
Plan działania



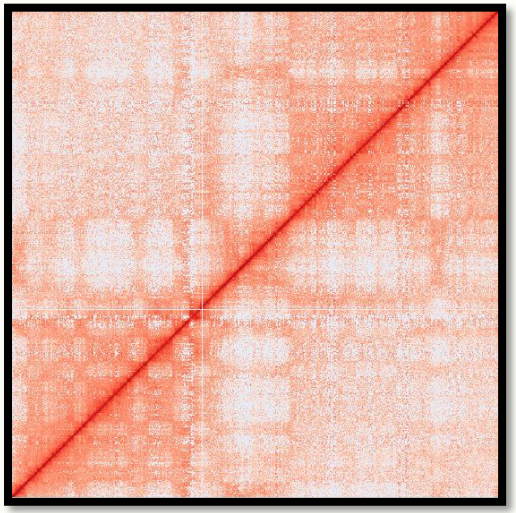
Problem: wykrycie domen topologicznych

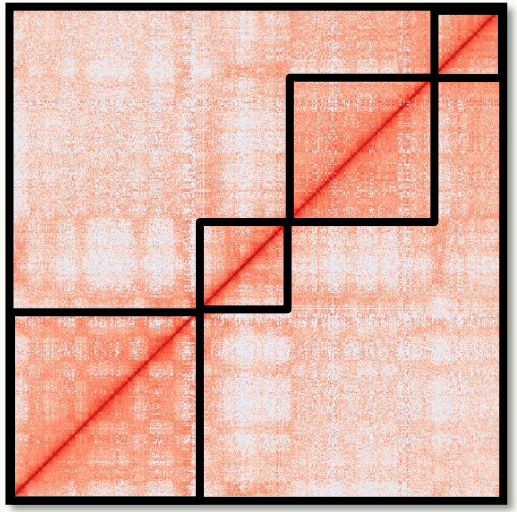


Problem: wykrycie domen topologicznych

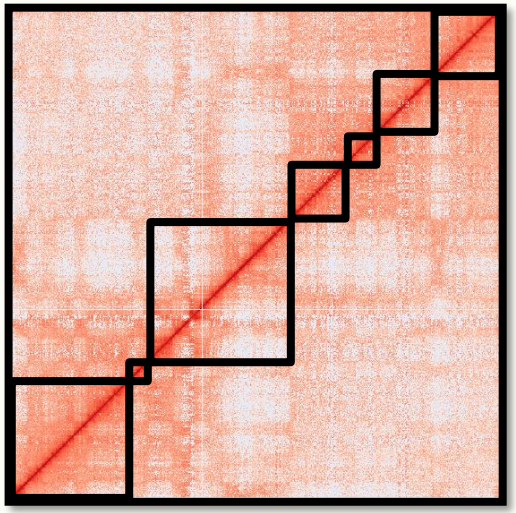


Jak ocenić jakość wyniku?

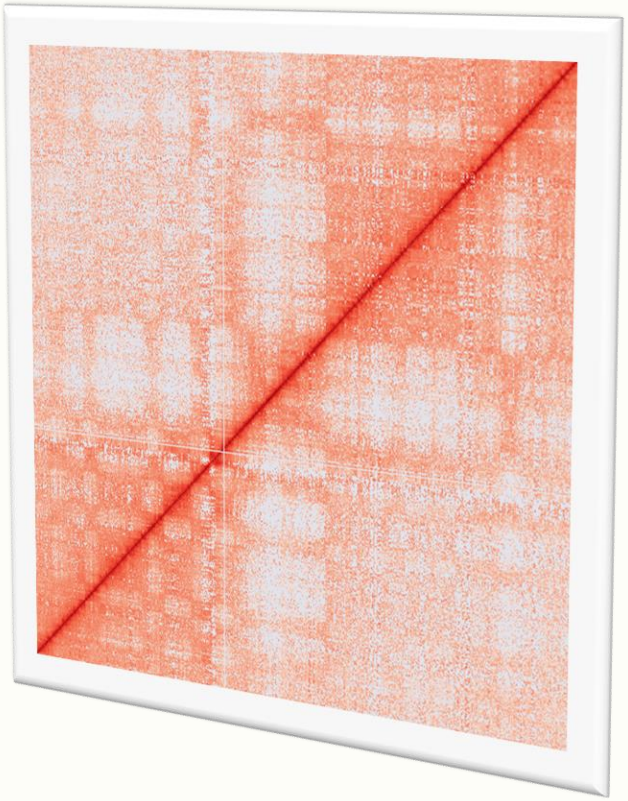




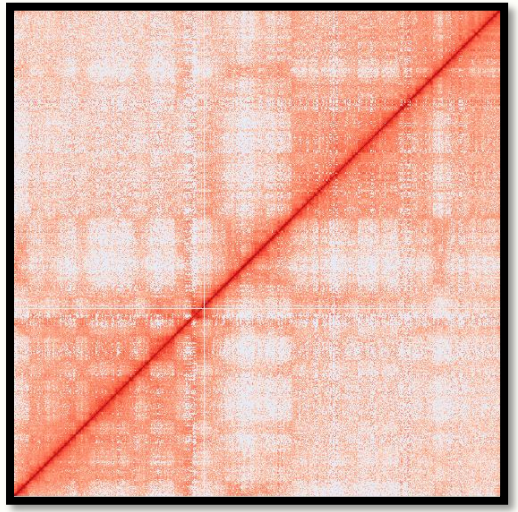
???



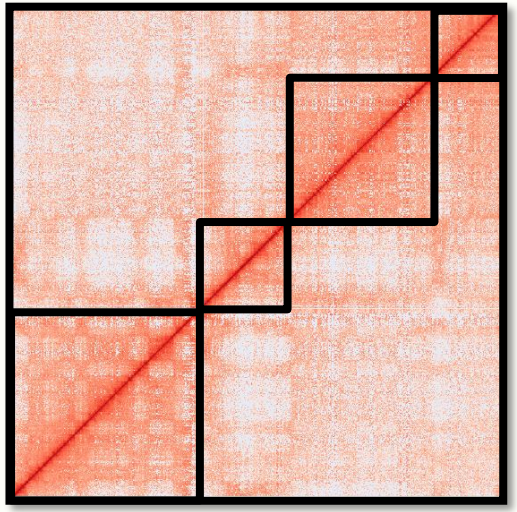
„Na oko”

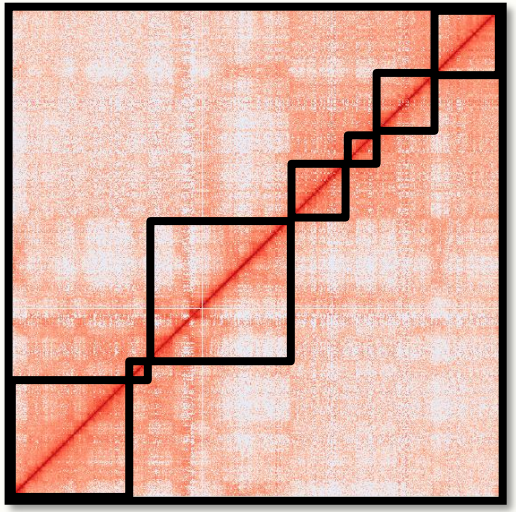


„Na oko”

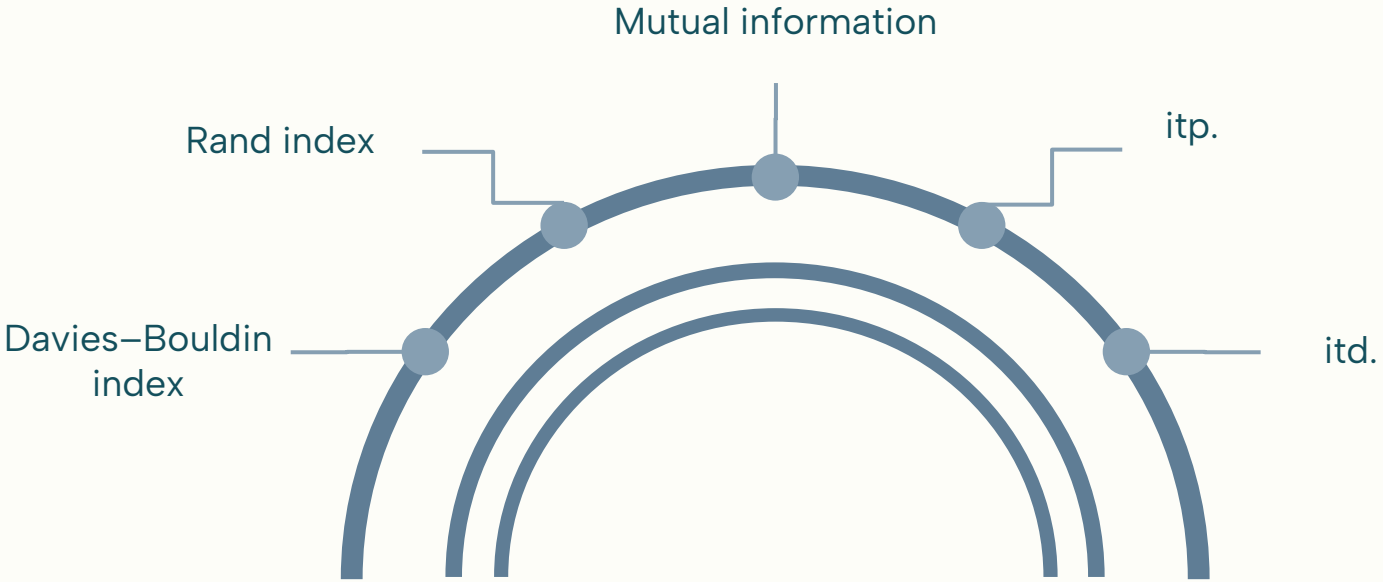


„To widać”





... oraz metrykami



Usprawnienie pracy

Jak było wcześniej wspomniane, celem jest zbudowanie metody stabilnej (nieczułej na rozdzielczość danych).

To oznacza, że model powinien dobrze działać na danych małego rozmiaru, jak i tych z większym detalem.

Tak więc budowa modeli i wstępne testowanie, w celu usprawnienia procesu twórczego, odbywały się na danych o małym detalu i metodą ,,na oko''. Następnie, jeśli wyniki były zadowalające, używane były dane dokładniejsze i stosowane były metryki.

Eksploracyjna analiza danych

Pobieranie danych

Dotychczas oglądana macierz ma następujące parametry: KR, chromosomy 11–11, resolution 100000.
Szczegółowe informacje o pochodzeniu danych (diploidalna komórka macierzyńska) można znaleźć na końcu prezentacji.

Parametry:

- Typ normalizacji (NONE, VC, VC_SQRT, KR, SCALE, etc.),
- Ścieżka do pliku,
- Chromosom 1,
- Chromosom 2 (powinien być taki sam, co 1, bo szukamy domen topologicznych),
- Typ interakcji fragmentów (BP – basic pair),
- Rozdzielczość (resolution: typically, 2500000, 1000000, 500000, 100000, 50000, 25000, 10000, 5000, etc.).

- Uwaga: rozdzielczość to parametr rozdrobnienia, ile elementów przypada na fragment – im mniejsza wartość, tym większa ilość danych.

Surowe dane

Tablica zawierająca 3 listy:

- Fragment A
- Fragment B
- Ilość interakcji fragmentu A z fragmentem B

[[0, 0, 0, 5000, 4000, 0, 420, ...],
[0, 1, 420, 69, 50000, 1100, ...],
[27759.044009373243,
5561.095348743029,
20778.80718787852,
882.7940292422061, ...]]

... i każda z pod-tablic ma długość 604977.

Ramka danych

Ramka danych zawiera:

- Fragment A –> x
- Fragment B –> y
- Ilość interakcji fragmentu A z fragmentem B –> count.

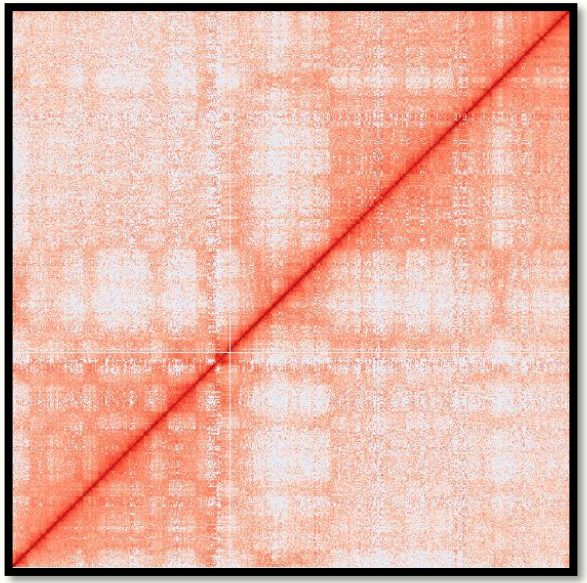
	x	y	count
0	0.0	0.0	27759.044009
1	0.0	500000.0	5561.095349
2	500000.0	500000.0	20778.807188
3	0.0	1000000.0	882.794029
4	500000.0	1000000.0	4594.032284
...	...	...	...
35204	132500000.0	134500000.0	928.262847
35205	133000000.0	134500000.0	1022.731802
35206	133500000.0	134500000.0	1542.685709
35207	134000000.0	134500000.0	5327.048897
35208	134500000.0	134500000.0	14060.778135

Macierz interakcji

Macierz, przedstawiana jako mapa ciepła o skali logarytmicznej:

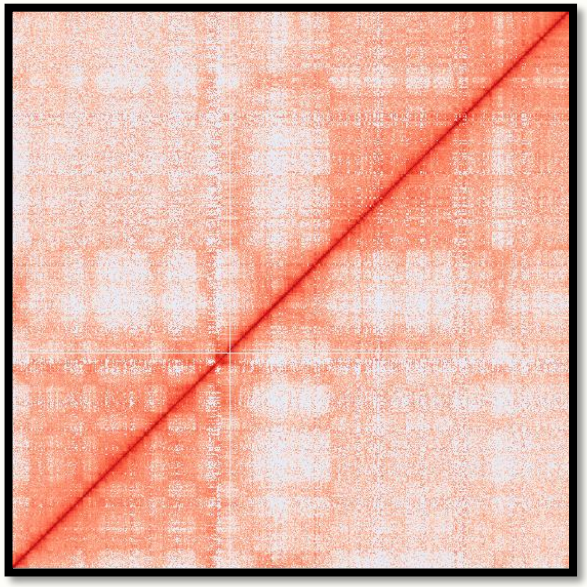
- Poziomo – x
- Pionowo – y
- Wartość pola w macierzy – count

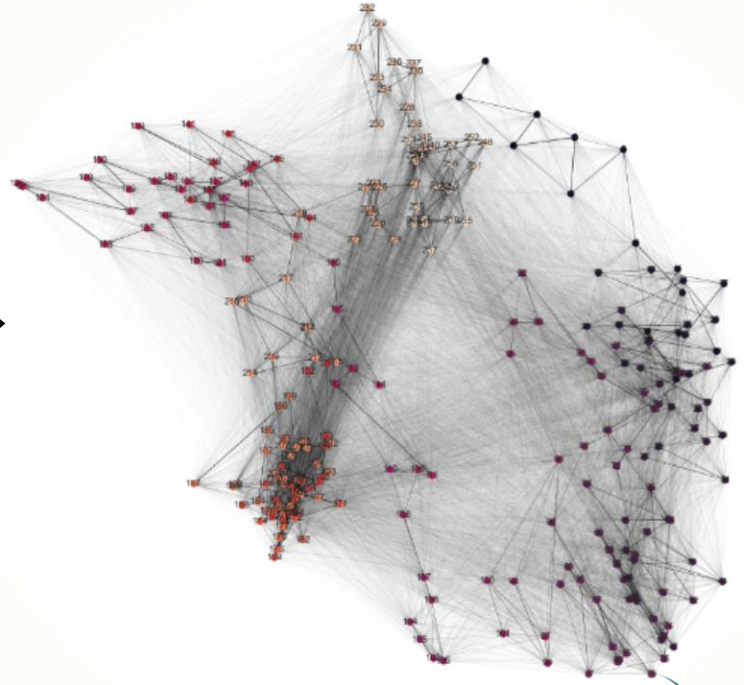
Braki danych zostały zastąpione 0.



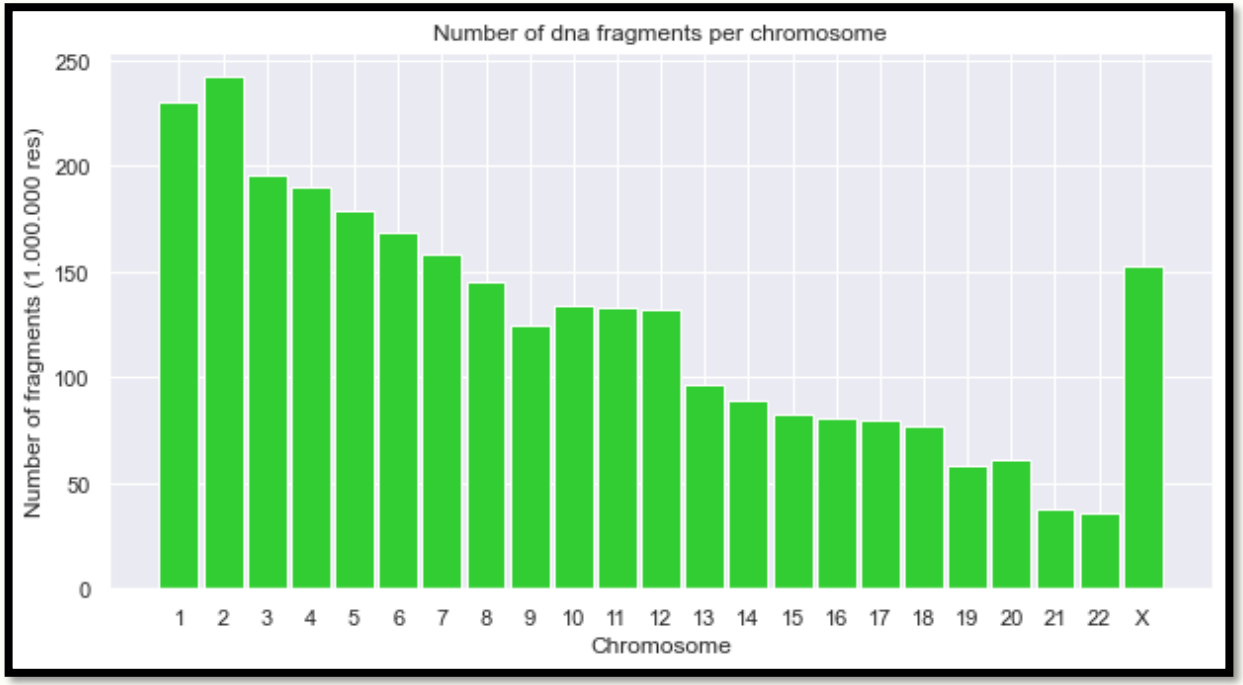
Wizualizacja klastrów w grafie

Wizualizacja macierzy kontaktów jako graf z rozmieszczeniem sprign_layout.



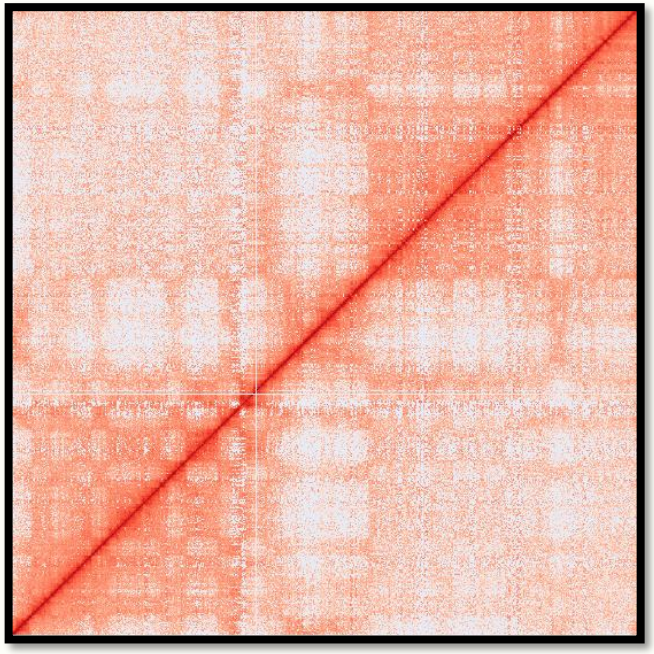


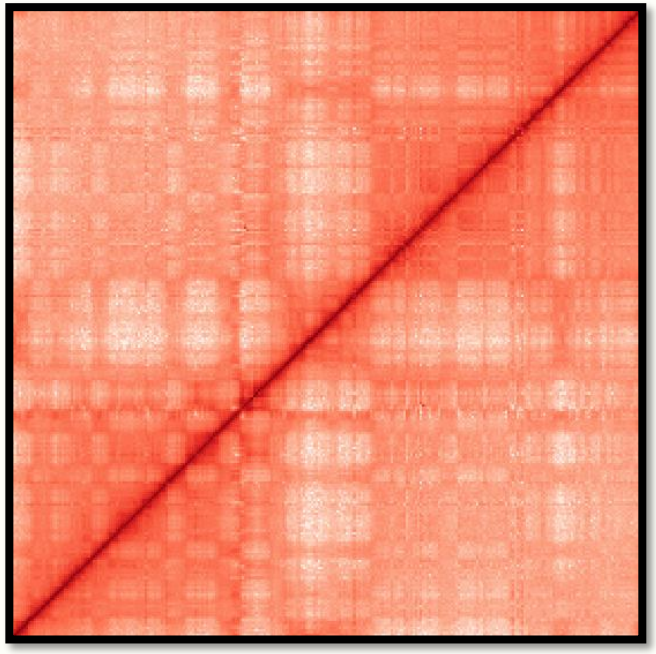
Rozmiary chromosomów



Porównanie rozdzielczości

100.000 vs 500.000





Wnioski

- Dane z większym detalem mają więcej braków danych (są rzadsze).
- Domeny topologicznie nie są wyznaczone jednoznacznie.
- Kolejność fragmentów ma znaczenie. Zwykłe algorytmy klasteryzujące grafy ważone nie dadzą poprawnych wyników.

Tworzenie modelu

Porównanie metod wygładzania macierzy

Wygładzanie w celu poprawy jakości danych. Szukamy takiego algorytmu, który nie zaburzy wyglądu macierzy, domeny topologicznie będą lepiej widoczne i macierz będzie mniej rzadka.

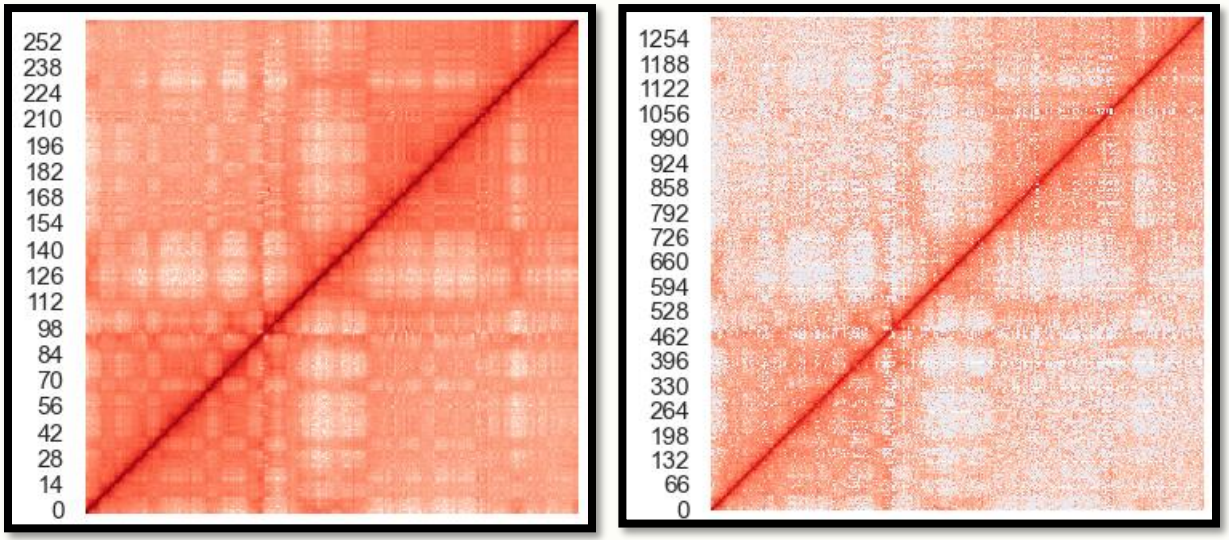
Rozmiary macierzy: 265 vs 1316 (rozdzielczość 500.000 vs 100.000).

Porównywane były:

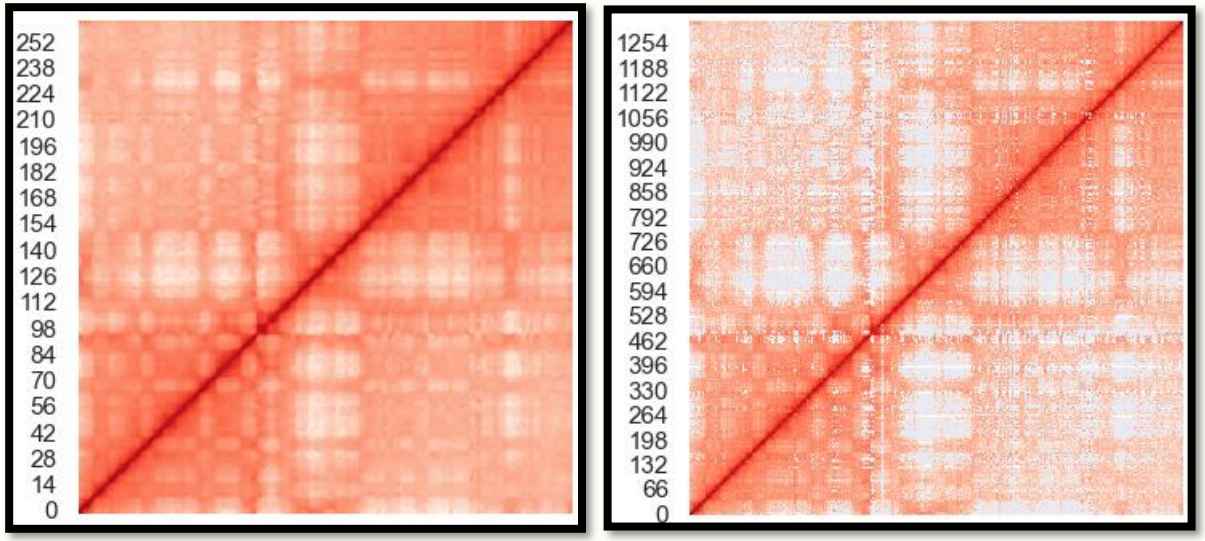
- Brak wygładzenia
- Filtr gaussowski, sigma=1
- Filtr średniej, rozmiar jądra=10
- Filtr mediany, rozmiar jądra=3

- Parametry były dobrane tak, aby wynik na mniejszej macierzy był jak najlepszy. Ten sam parametr jest dla większej macierzy, aby zaobserwować, jak rozmiar wpływa na wynik wygładzenia.

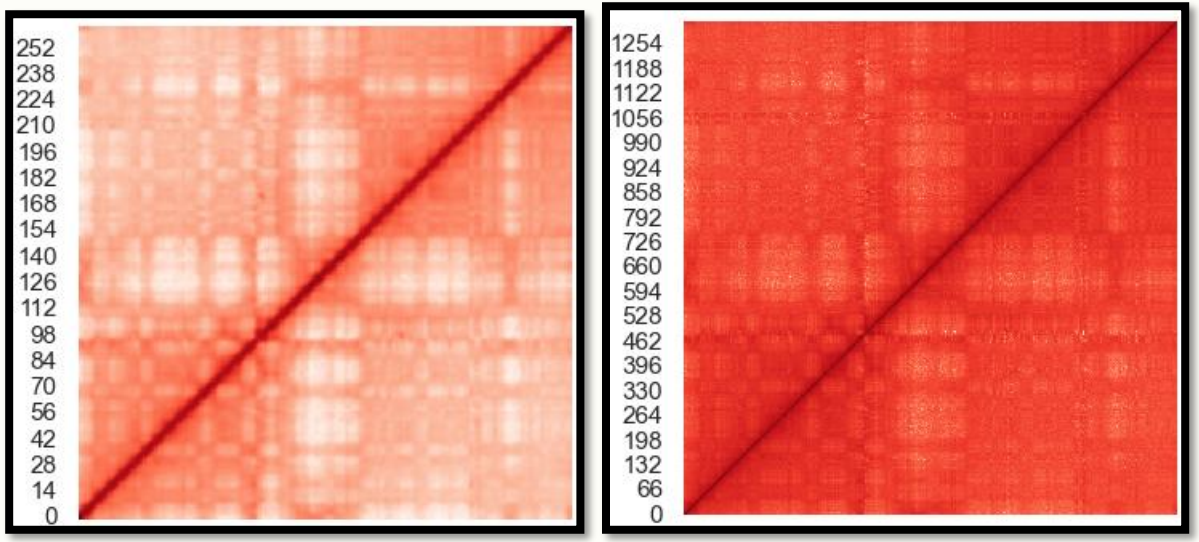
Brak wygładzenia



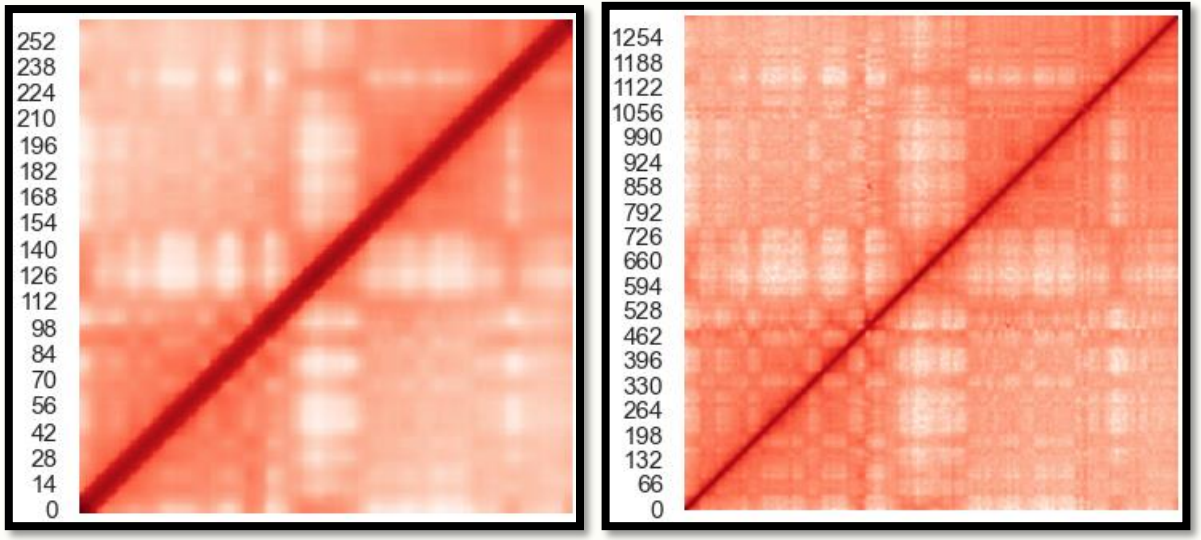
Filtr mediany



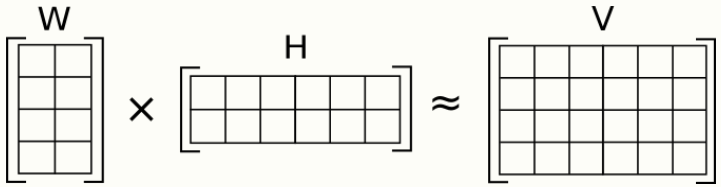
Filtr gaussowski



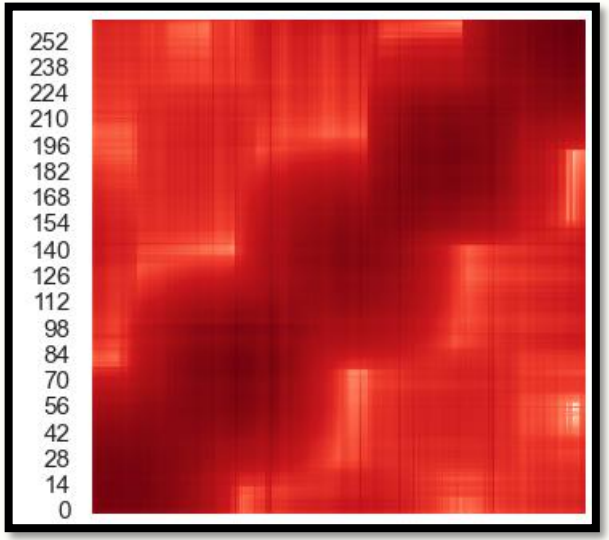
Filtr średniej



Regulowana grafami dekompozycja macierzy nieujemnej (graph regularized NMF)



Parametry były ręcznie dobrane tak (siła regularyzacji, promień sąsiedztwa i ilość składowych), aby efekt był jak najlepszy.



Walidacja i poprawa macierzy

Po wygładzeniu macierz może przestać być symetryczna.

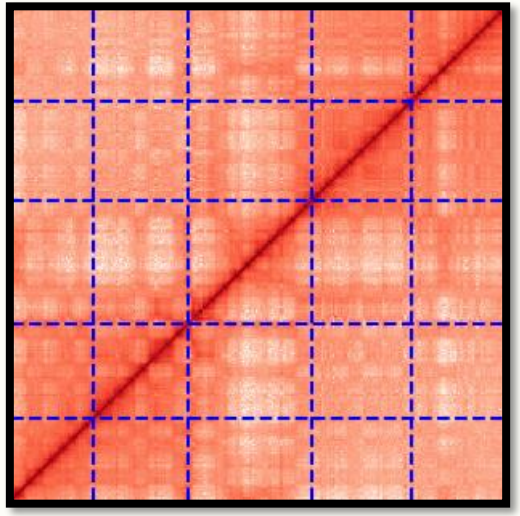
W takim wypadku dolna trójkątna połowa jest odbijana lustrzanie, przywracając symetrię.

Kreacja i testowanie metod

Następna część pokazuje rozpatrywane metody do znajdywania domen topologicznych i najlepsze rezultaty, jakie się udało uzyskać.

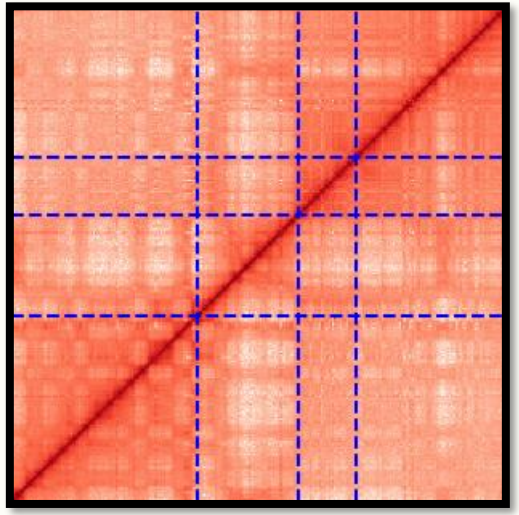
Metoda 1

Na podstawie insulation score.



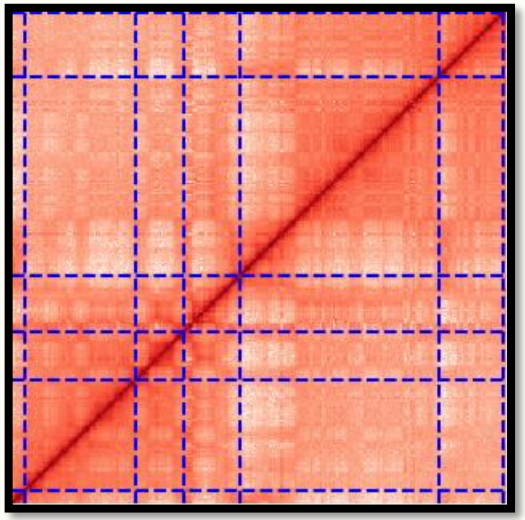
Metoda 2

Na podstawie directionality index.



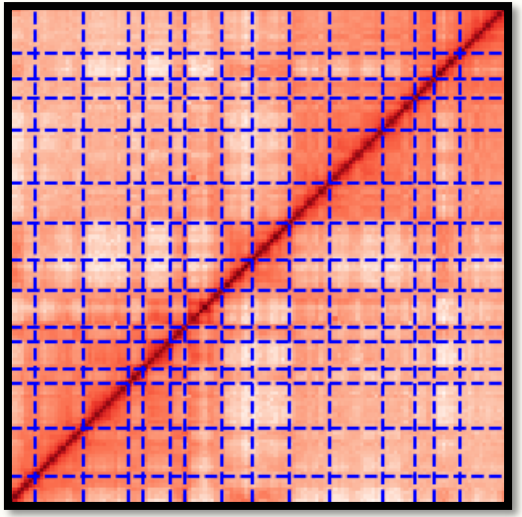
Metoda 3

Na podstawie laplacjanu z regulowanej grafami nieujemnej dekompozycji macierzy.



Metoda 4

Na podstawie algorytmu detekcji społeczności w grafie (louvian_communities w networkx).



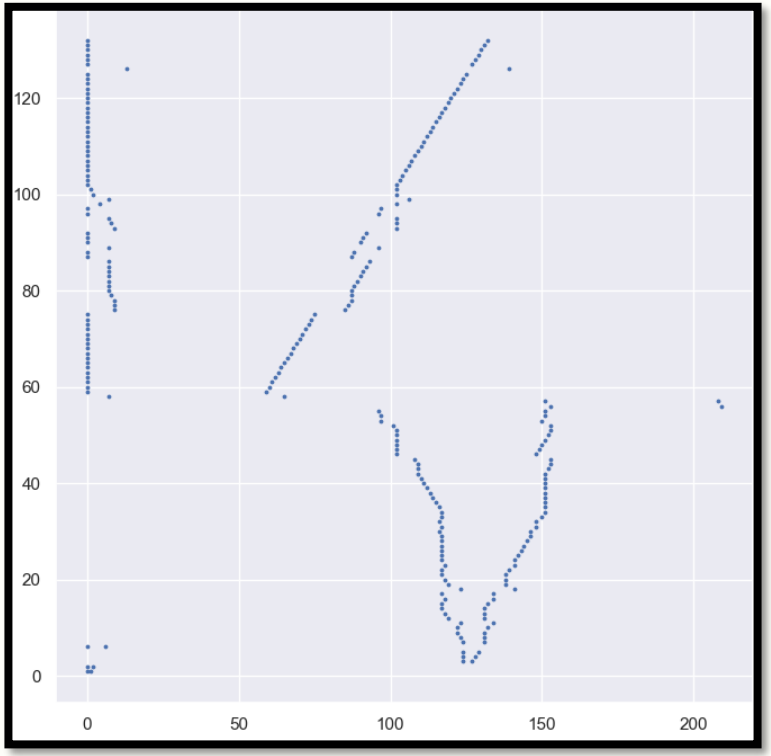
Metody 5–8

Korzystają z wyznaczania średniej wartości pola macierzy wewnątrz jądra o zmiennym promieniu, poruszającym się wzdłuż diagonali macierzy interakcji.

Dla każdego rozmiaru jądra r (liczba wierszy/kolumn) jest brana średnia wartość elementów wewnątrz jądra. Element jądra r:1 jest na elemencie i:i w i–tej iteracji dla danego r.

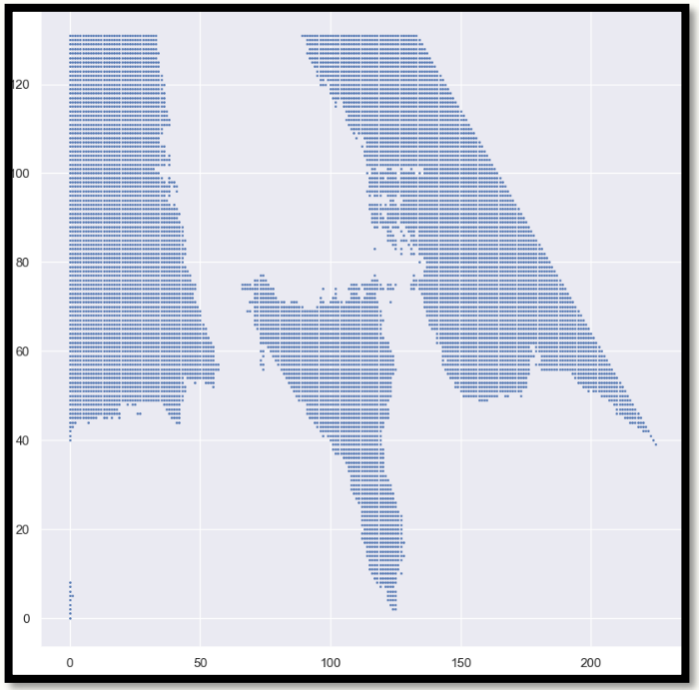
Metoda 5

Branie 2 maksymalnych wartości w odległości większej niż r od siebie.



Metoda 6

Branie wartości niemniejszych niż maksimum * 0.9 w każdej iteracji (dla każdego r).



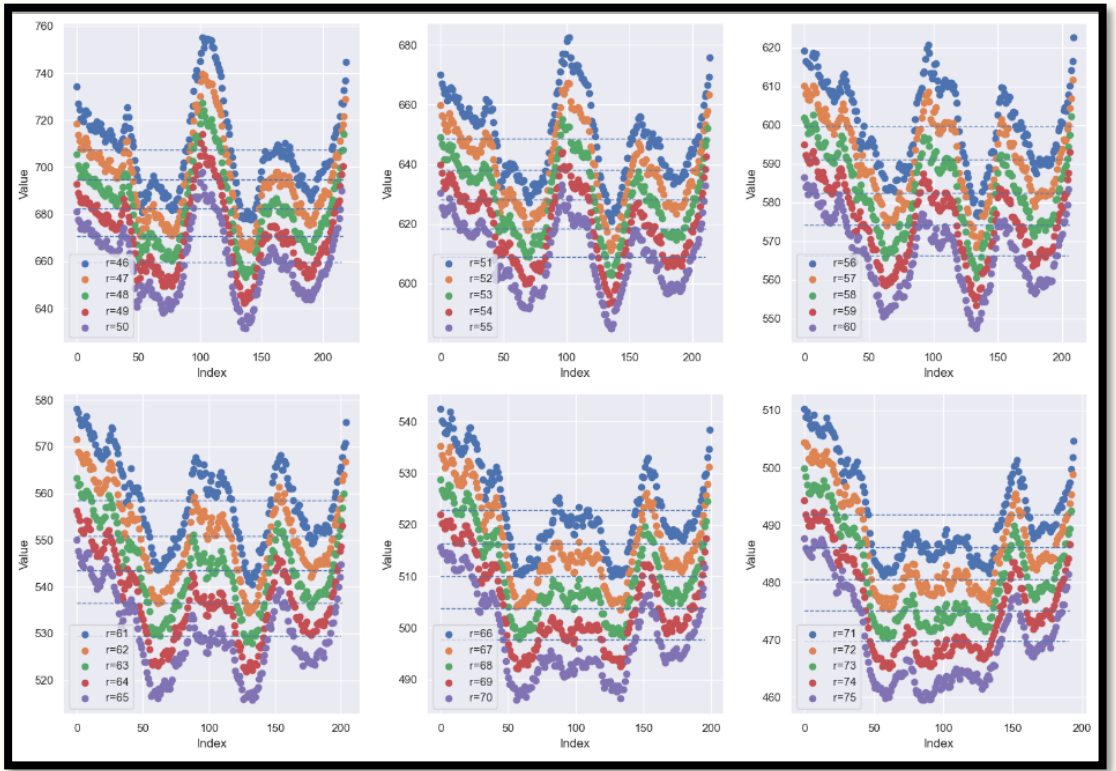
Metoda 7

Rozdzielanie tablicy wartości na sekwencje o podobnych sumach.

```
r: 20: 49 100 145 197
r: 21: 49 100 145 196
r: 22: 48 99 144 195
r: 23: 48 99 144 195
r: 24: 48 99 144 194
r: 25: 48 98 143 193
r: 26: 48 98 142 192
r: 27: 47 97 141 191
r: 28: 47 97 141 190
r: 29: 47 96 140 189
r: 30: 47 96 140 189
r: 31: 47 96 140 188
r: 32: 47 95 139 187
r: 33: 46 94 138 186
r: 34: 46 94 138 186
r: 35: 46 94 138 185
r: 36: 46 93 137 184
r: 37: 45 92 136 183
r: 38: 45 92 135 182
r: 39: 45 92 135 181
r: 40: 45 91 134 180
r: 41: 45 91 134 180
r: 42: 44 90 133 179
```

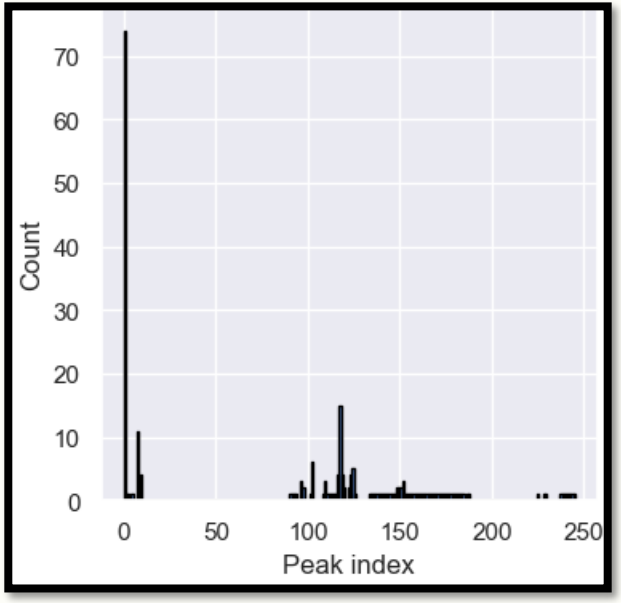
Metoda 8

Dla każdej iteracji (każdego rozmiaru r jądra) jest tworzona tablica wartości. Z tej tablicy zerowane są wartości mniejsze od średniej. W sekwencjach nieoddzielonych zerami są wyznaczane wartości maksymalne, reszta jest zerowana. Wynikiem jest zbiór indeksów wartości niezerowych.



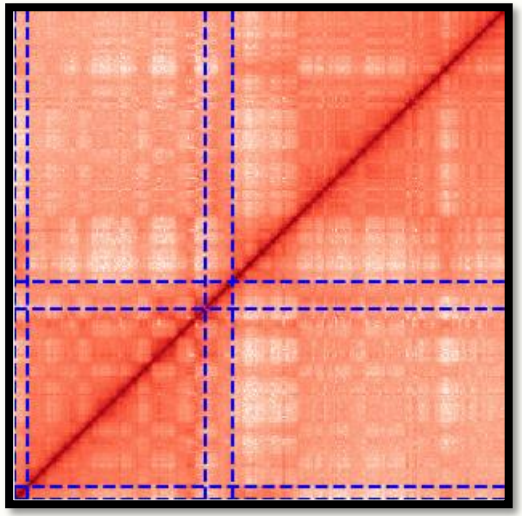
Metoda 8 cd.

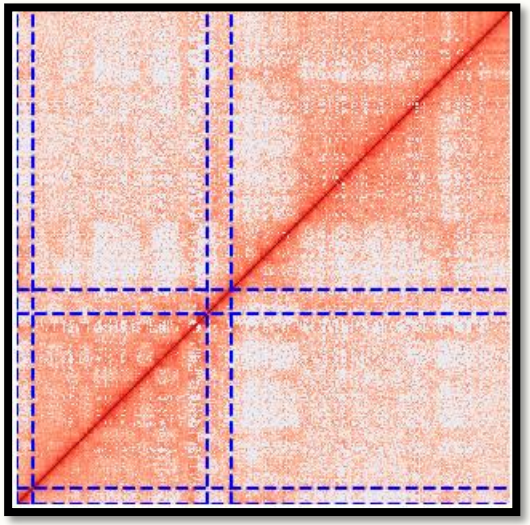
Histogram otrzymanych opisaną metodą wartości.



Wyniki

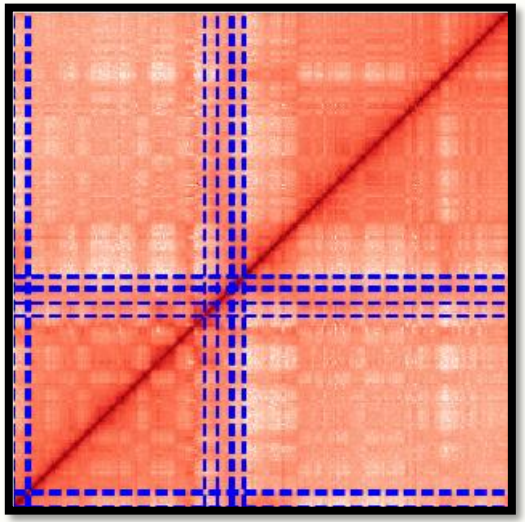
Najlepsze wyniki po dobraniu odpowiednich parametrów.

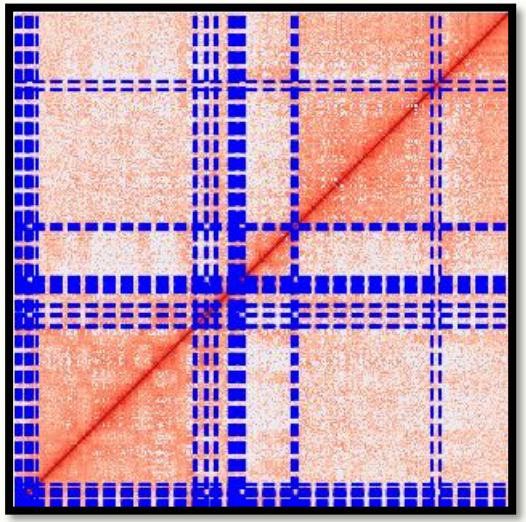




Problem

Dla większej ilości dopuszczanych granic domen widać, że granice są wyznaczane niejednoznacznie, jednak już znalezione domeny nie są dalej dzielone w nieodpowiednich miejscach.

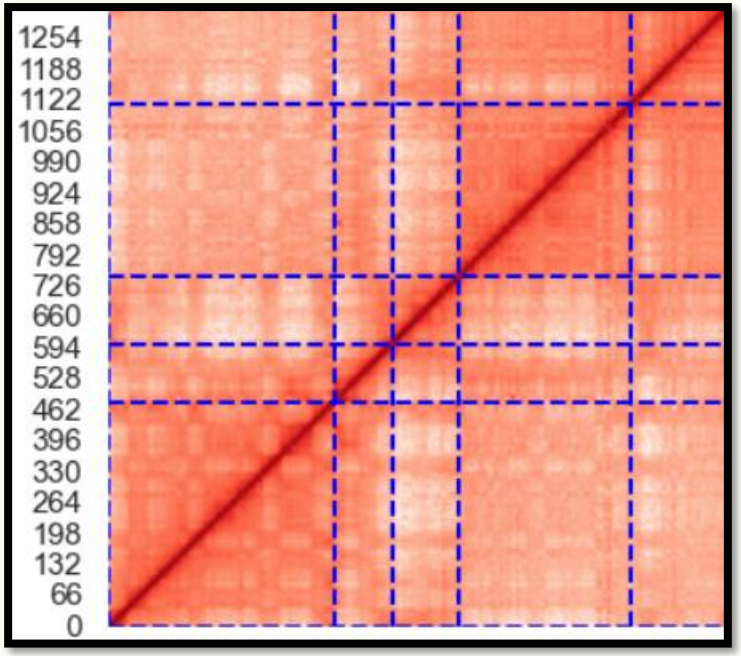




Rozwiązanie

Rozwiązanie – scalanie granic, jeśli są zbyt blisko siebie.

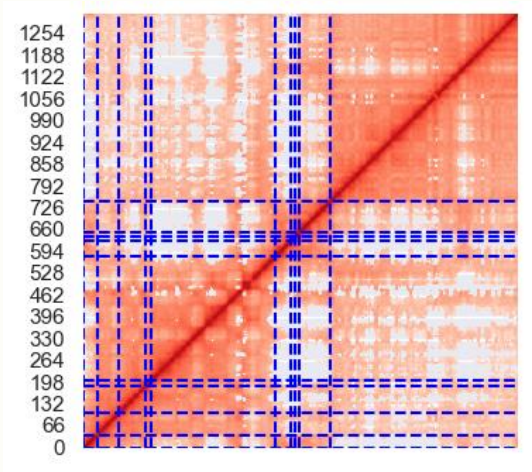
Z ręcznie dobranymi parametrami otrzymano następujący rezultat.



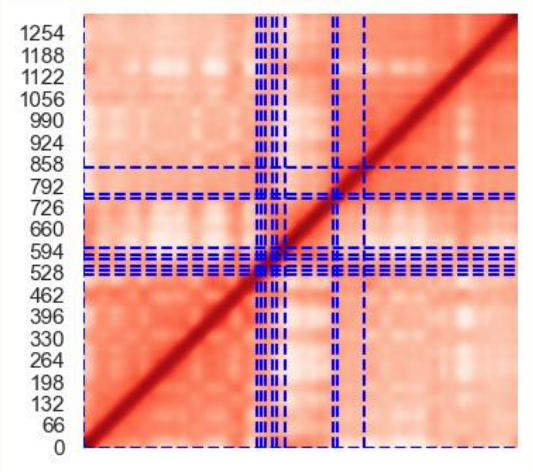
Testowanie i strojenie modelu

Wybór zastosowanego filtra

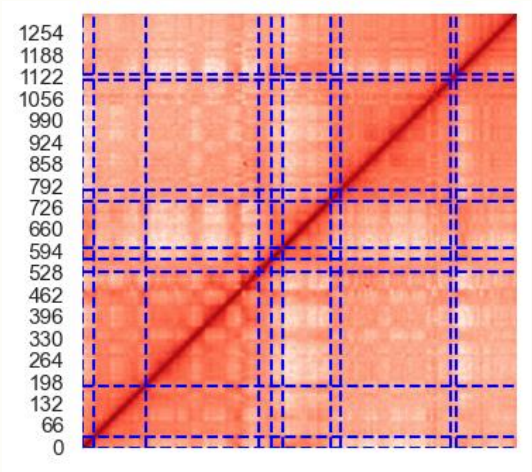
Przy tych samych pozostałych parametrach widzimy, że najlepsze wyniki dostajemy dla filtra 3, czyli używającego średnią. Jednak według wskaźnika TAD-adjR^2, to filtr gaussowski jest najlepszy.



1



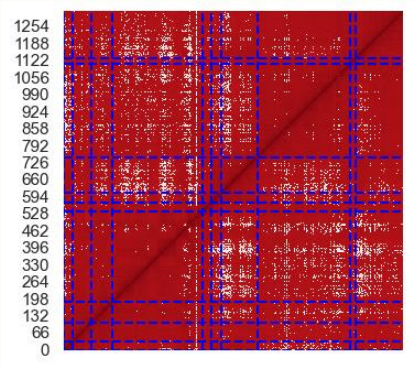
2



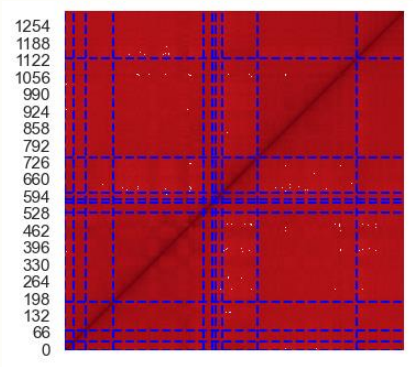
3

Wybór wskaźnika wygładzenia

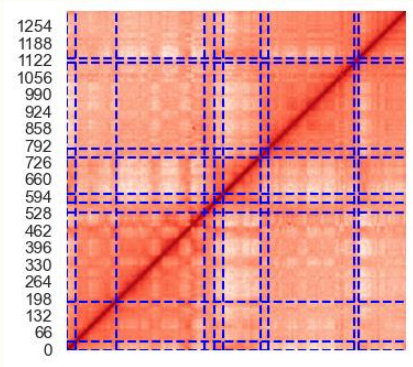
Przy tych samych pozostałych parametrach widzimy, że im większy wskaźnik tym lepsze wyniki, ale do pewnego momentu. Najlepsza wartość to 10.



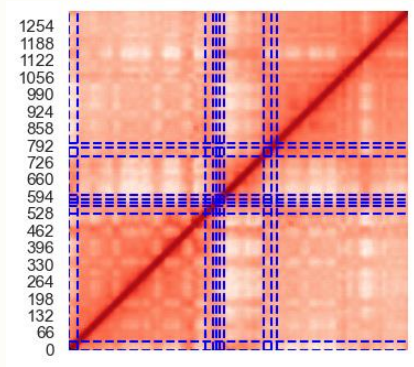
W = 2



W = 5



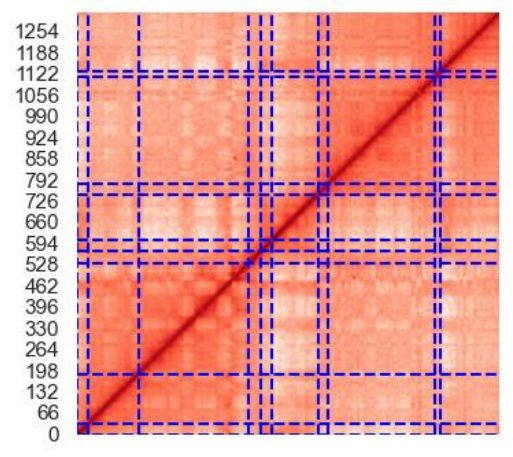
W = 10



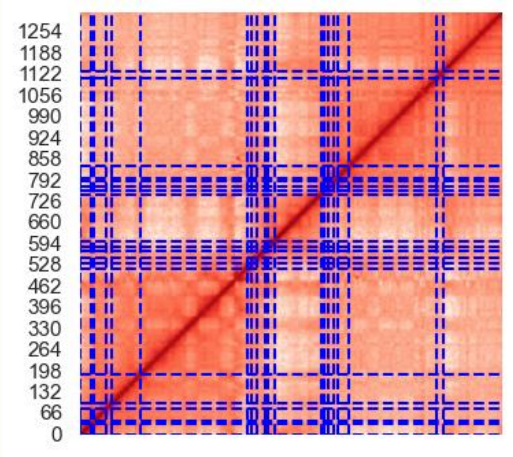
W = 20

Porównanie maksymalnej liczby TADów

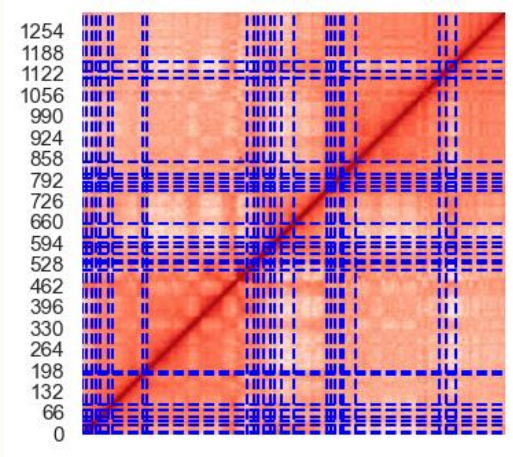
W tym przypadku widzimy, że więcej nie znaczy lepiej. Maksymalna liczba TADów wyznaczona jako 1% z wielkości macierzy jest zadowalającą opcją.



ok. 1%



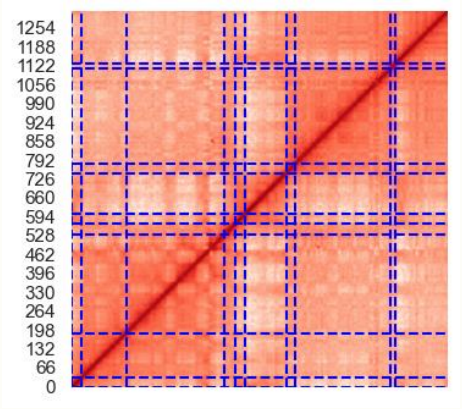
ok. 2%



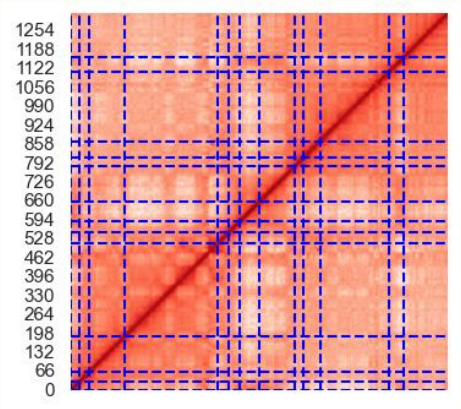
ok. 4%

Porównanie maksymalnej liczby TADów a minimalny rozmiar TADu

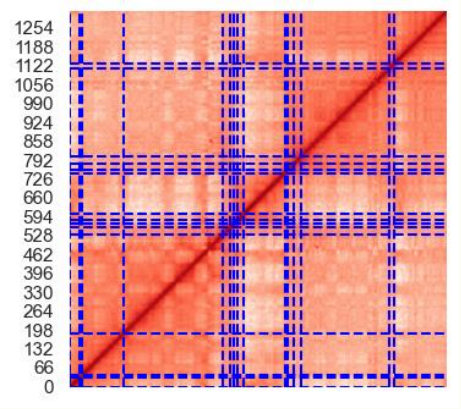
Tutaj zależało mi na sprawdzeniu czy z większą maksymalną liczbą TADów powinna iść mniejszy minimalny rozmiar TADu, czy większy.



max_liczba = 10
min_rozmiar = 10

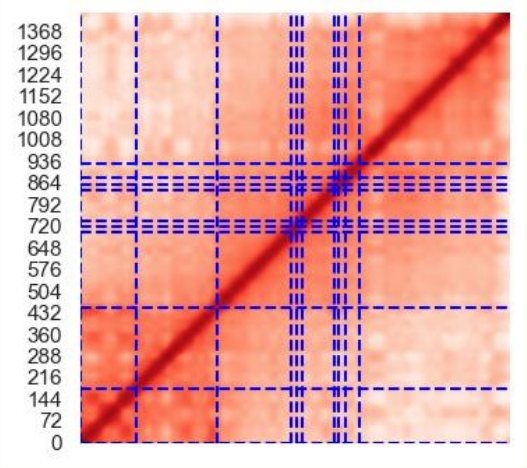


max_liczba = 15
min_rozmiar = 30

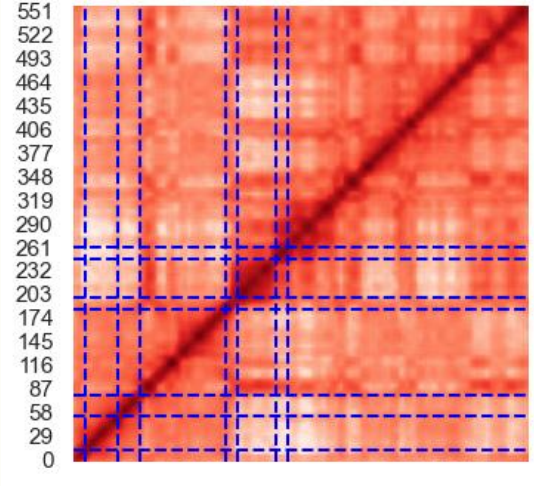


max_liczba = 15
min_rozmiar = 8

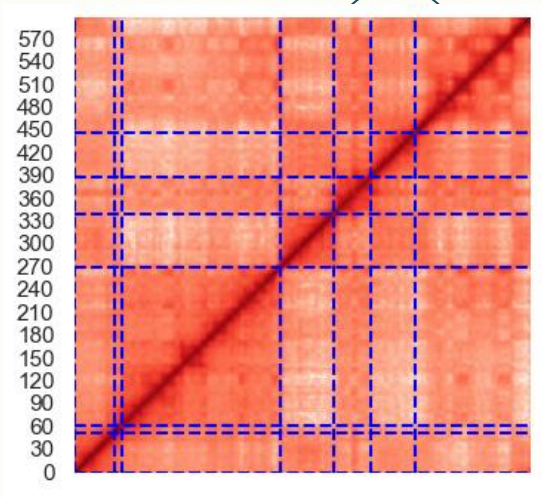
Wyniki modelu



Chromosom 12



Chromosom 19



Chromosom 20

Porównanie z innymi metodami

Porównywane metody

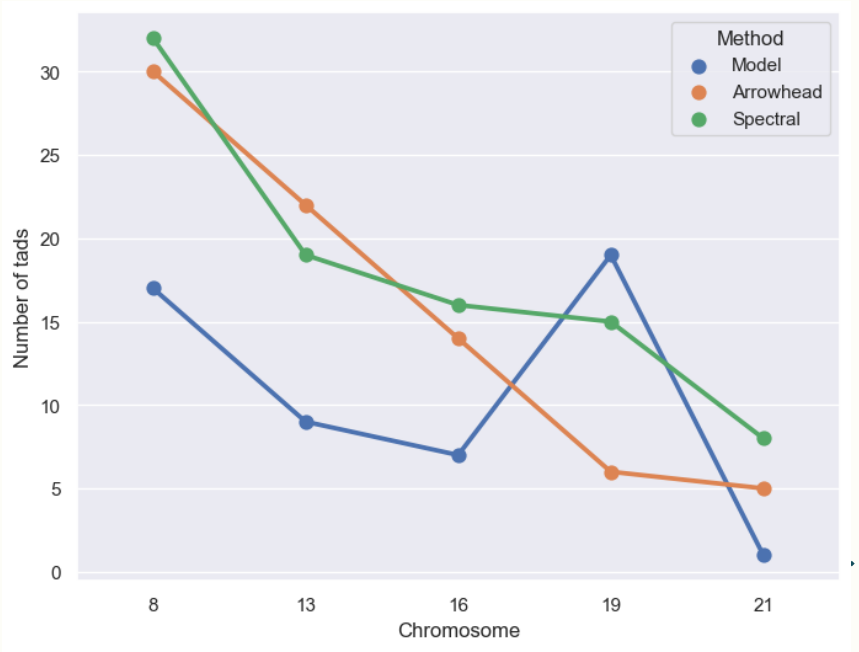
Arrowhead

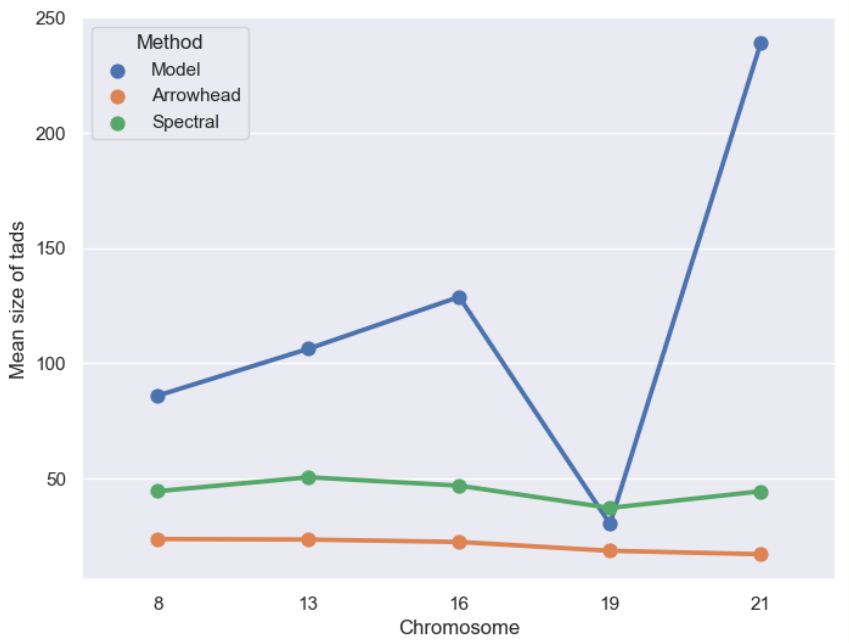
1. Analiza macierzy kontaktów
2. Identyfikacja zmian kontaktów
3. Tworzenie profilu
4. Detekcja granic TADów

SpectralTAD

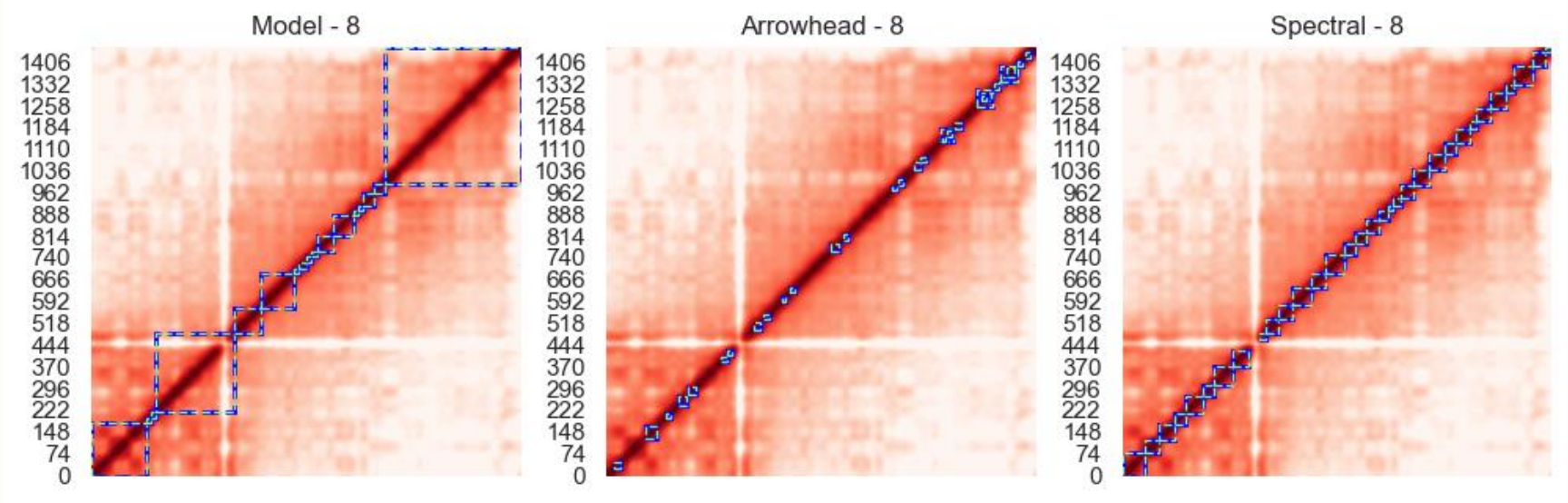
1. Analiza spektralna
2. Wybór wektorów własnych
3. Klasteryzacja
4. Identyfikacja granic TADów

Resolution: 100 000, losowe chromosomy





Przykładowy wynik dla chromosomu 8



Metryki jakich użyliśmy do porównania

Davies-Bouldin index

mierzy średnią podobieństwa między każdą domeną TAD a najbardziej podobną do niej domeną, gdzie niższe wartości wskazują na lepsze rozdzielenie i kompaktowość TAD-ów

Measure of Concordance

miara oceniająca zgodność między dwoma podziałami TAD-ów

TAD-adjR^2

miara oceniająca, w jakim stopniu klasyfikacja TAD-ów wyjaśnia zmienność danych Hi-C, uwzględniając liczbę i rozmiary wyznaczonych TAD-ów

Delta Contact Count

miara oceniająca zmiany w liczbie kontaktów między regionami DNA przed i po zastosowaniu metody wykrywania TAD-ów

Wyniki dla TAD-adjR^2

	Model	Arrowhead	Spectral
Chromosom 8	0.2701	0.1358	0.6034
Chromosom 13	0.315	0.1966	0.5849
Chromosom 16	0.1672	0.135	-
Chromosom 19	0.3142	0.193	0.5488
Chromosom 21	0.0883	0.2586	0.5199

Wyniki dla Delta Contact Count

	Model	Arrowhead	Spectral
Chromosom 8	29.7848	139.96	109.1745
Chromosom 13	59.4726	183.0016	133.3686
Chromosom 16	34.2688	244.6567	171.9981
Chromosom 19	55.0101	155.518	106.8379
Chromosom 21	50.8365	223.2853	147.6177

Wyniki dla

Davies-Bouldin index **oraz** Delta Contact Count

	Model	Arrowhead	Spectral
Chromosom 8	1.4175	0.9891	0.8169
Chromosom 13	1.5932	1.2437	0.9324
Chromosom 16	1.7423	0.6348	-
Chromosom 19	0.9023	0.3144	0.7393
Chromosom 21	-	0.3538	0.8876

Arrowhead	Spectral
0.1478	0.4281
0.2418	0.4404
0.1158	0.3594
0.1015	0.3837
-0.6025	-0.0290

Podsumowanie

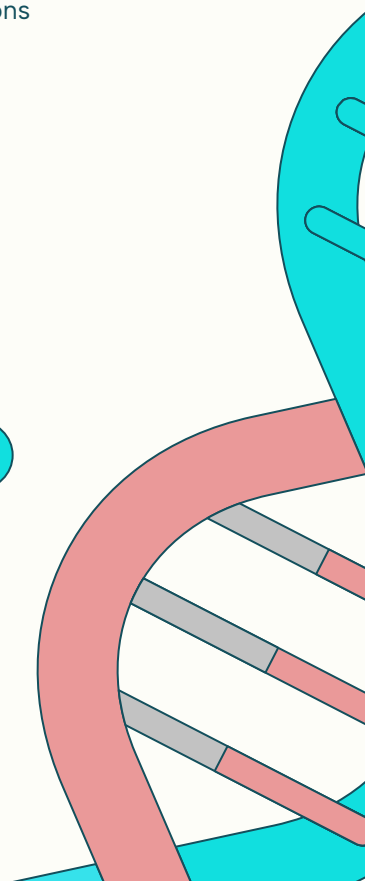
Dziękujemy za uwagę

Bibliografia

- Slajd 3: https://commons.wikimedia.org/wiki/File:Structural_organization_of_chromatin.png - Creative Commons Attribution-Share Alike 4.0 International license (CC BY-SA 4.0 DEED)
- Slajd 5: https://commons.wikimedia.org/wiki/File:HiCschematic.png - CC BY-SA 4.0 DEED
- Dane użyte w przykładach: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525 - GSE63525_GM12878_diploid_maternal.hic
- Slajd 31: https://en.wikipedia.org/wiki/Non-negative_matrix_factorization#/media/File:NMF.png

Użyte wskaźniki porównania pochodzą z następujących źródeł:

- https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04674-2
- https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1596-9
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8152090/#Sec11



Repozytorium projektu

https://github.com/Sebislaw/WB1-DomenyTopologiczne

Thanks!

Do you have any questions?

youremail@freepik.com
+91 620 421 838
yourwebsite.com