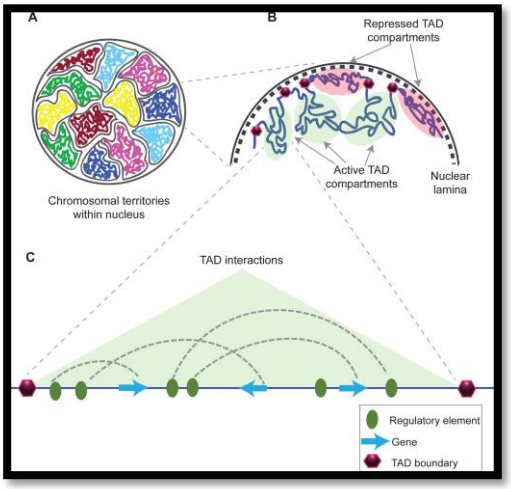
Podział genomu na domeny topologiczne – Raport

Autorzy: Sebastian Pergała, Aleksandra Samsel

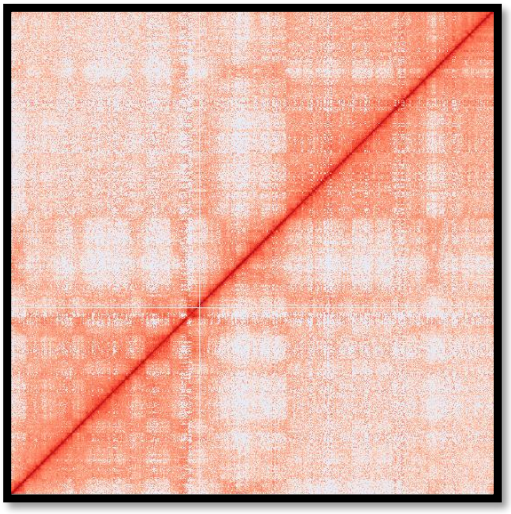
Spis treści:

Wstęp teoretyczny

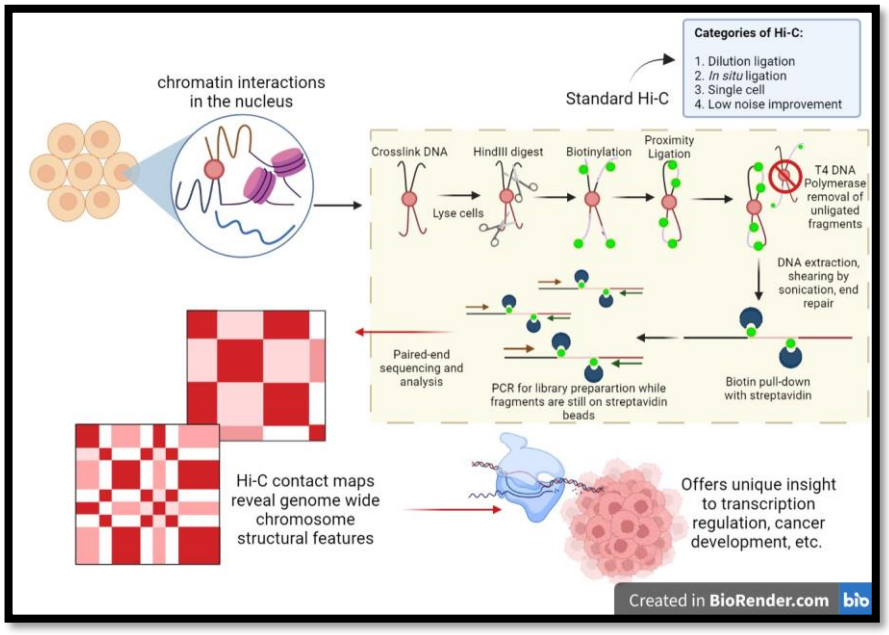
Domeny topologiczne (TAD – Topologically Associationg Domains) - regiony genomu, wchodzące w częste interakcje same ze sobą. Oznacza to, że sekwencje DNA znajdujące się wewnątrz domeny mają ze sobą więcej fizycznych interakcji, niż z sekwencjami poza nią.



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.



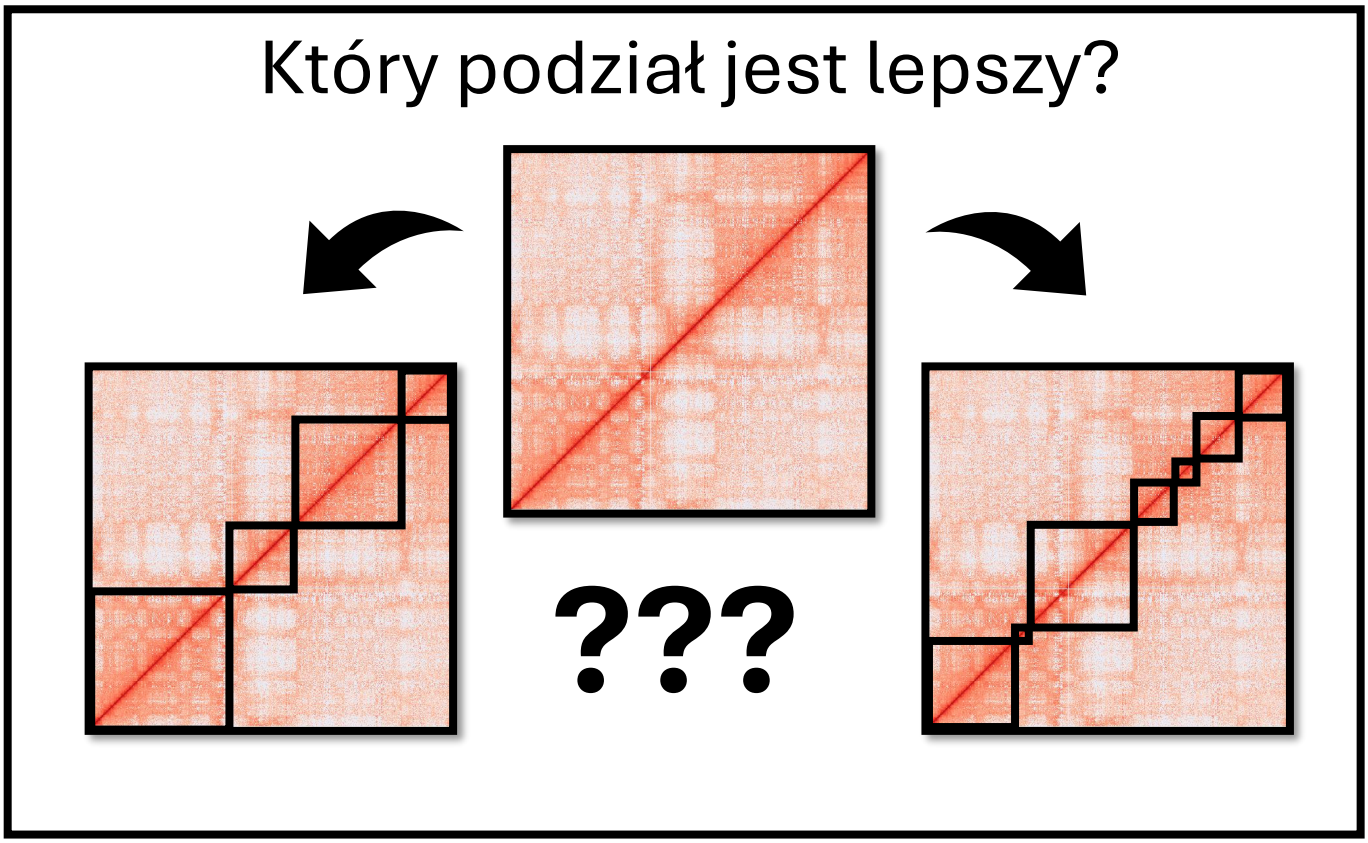
Hi–C – technika zbierania informacji na temat konformacji chromatyny. Zlicza częstotliwość (jako średnia na całą populację komórki) fizycznych kontaktów segmentów DNA, łącząc ze sobą trójwymiarową strukturę chromosomów i sekwencje DNA.



Cel biznesowy

Cel biznesowy projektu był następujący:

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

Pracę nad projektem można podzielić na pięć głównych etapów:

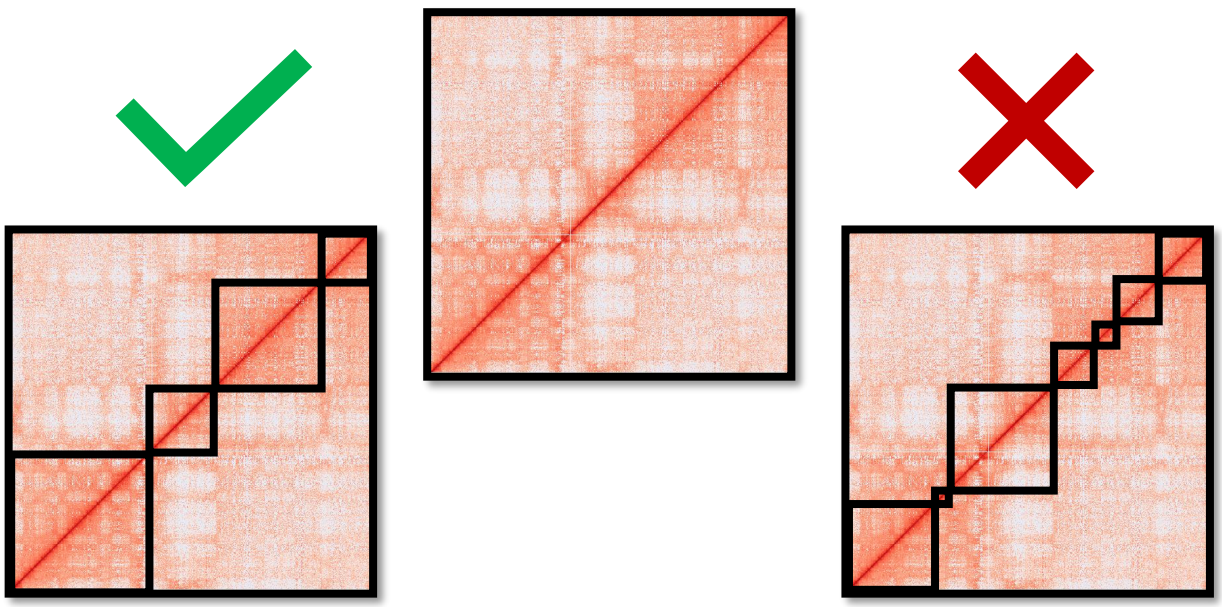
1. Eksploracyjna analiza danych.
2. Tworzenie różnych metod.
3. Wybór najlepszej metody i konstrukcja modelu.
4. Dobranie najlepszych parametrów modelu.
5. Testowanie modelu i porównanie z innymi metodami.

Jako sposób oceny jakości wykrytych domen można było wyróżnić:

1. Metodę „na oko” – człowiek ma bardzo dobre zdolności wykrywania wzorców geometrycznych w danych wizualnych, zatem użycie własnej oceny do wstępnej oceny jakości wyników nie jest niewskazane, a usprawnia tok pracy.
2. Za pomocą metryk – walidując model na większej próbie danych nie sposób jest ocenić wszystkie wyniki w sposób jednakowy i stosunkowo niewielkim czasie. Dlatego Posłużyliśmy się różnymi metrykami (które zostały omówione kolejnych sekcjach) w celu automatyzacji oceny i umożliwienia działania na większą skalę.

Dodatkowo, celem usprawnienia procesu twórczego, ocena jakości i budowa modeli odbywała się początkowo na danych o mniejszym rozmiarze, a dopiero po osiągnięciu zadowalającego efektu praca przechodziła na dane o większym detalu. Celem biznesowym było stworzenie algorytmu stabilnego (na różną wielkość danych), tak więc satysfakcjonujące wyniki na mniejszych danych są jednym z warunków koniecznych do osiągnięcia postawionego celu.

Wizualna ocena jakości przykładowo wyznaczonych domen



Eksploracyjna analiza danych

Dane, na których operowaliśmy są opisane w większym szczególe na stronie podanej w bibliografii jako źródło danych. W skrócie, dane dotyczyły interakcji fragmentów DNA w diploidalnej komórce macierzystej kobiety.

Podczas ładowania danych .hic za pomocą biblioteki straw w pythonie, można było ustawić następujące 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 (zwykle, 2500000, 1000000, 500000, 100000, 50000, 25000, 10000, 5000, itp.). Warto zauważyć, że rozdzielczość to parametr rozdrobnienia, czyli ile elementów przypada na jednostkę– im mniejsza wartość, tym większa ilość danych.

Tablica powstała z nieobrobionych danych miała następującą postać; 3 listy:

- Fragment A (sekwencja DNA),
- Fragment B,
- Ilość interakcji fragmentu A z fragmentem B.

Przykładowy wygląd danych:

[[0, 0, 0, 5000, 4000, 0, 420, ...],

[0, 1, 420, 69, 50000, 1100, ...],

[27759.044009373243, 5561.095348743029, 20778.80718787852, ...]].

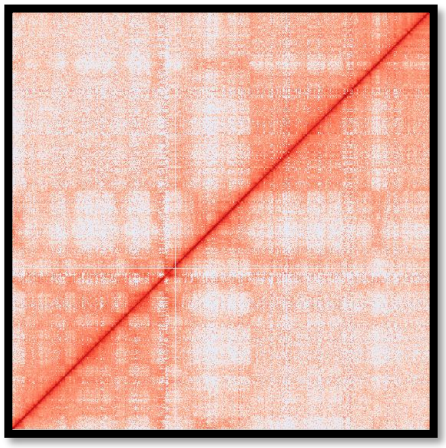
Takie dane zostały przekształcone do następującej postaci:

- 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

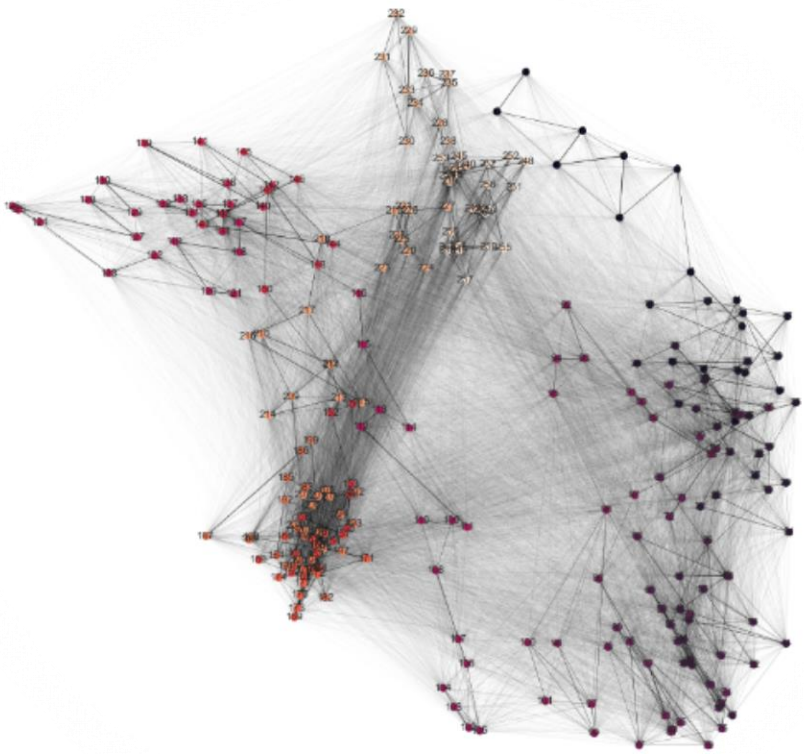
Następnie ramka danych była przekształcona do postaci macierzy. Z uwagi na to, że tylko zaobserwowane interakcji znajdowały się w oryginalnych danych, występowały braki danych. Te miejsca w macierzy zostały zastąpione wartością 0. W przypadku, gdy brakowało całego wiersza (fragment DNA o danym indeksie w ogóle nie wystąpił w danych), ten wiersz był dodawany do macierzy, a wszystkie wartości w nim były ustawiane jako zerowe.

Przykład wizualizacji macierzy interakcji za pomocą mapy ciepła.

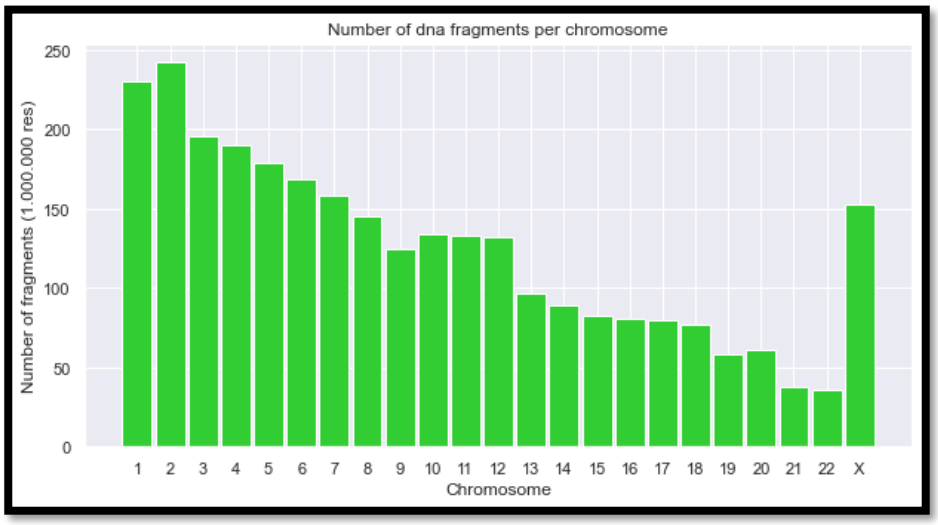


Innym testowanym sposobem wizualizacji danych były grafy. Za pomocą układu „spring laoyut" w bibliotece networkx, można było stworzyć graf, gdzie położenia wierzchołków zależały od wartości wag na krawędziach (wierzchołek – indeks wiersza w macierzy, waga – wartość w macierzy). Takie podejście spotkało się jednak z problemem; dla większych danych biblioteka osiągała limit obliczeniowy. Stworzone grafy nie przekazywały istotnych informacji na temat przynależności fragmentów do domen, choć kilka wyraźniejszych klastrów można było wyróżnić.

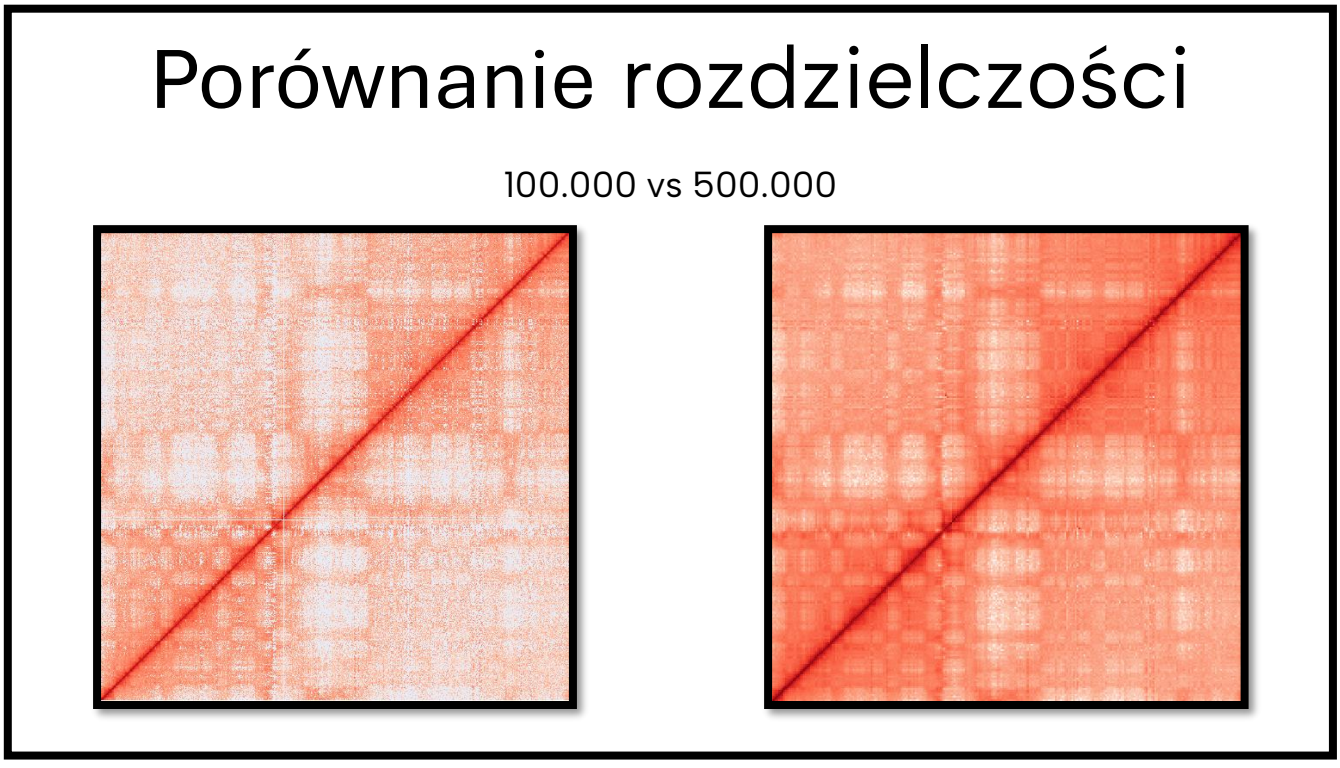
Przykład grafu dla powyższej macierzy.



Jako ciekawostkę sprawdziliśmy wielkość danych dotyczących każdego z chromosomów. Jak można się było spodziewać, ich wielkość maleje wraz ze wzrostem numeru.



Następnie porównaliśmy, jak parametr rozdzielczości wpływa na otrzymywane macierze. Wraz ze wzrostem detalu ilość wartości zerowych rosła. Można więc zaobserwować kompromis między pełniejszą informacją na temat struktury genomu, a danymi trudniejszymi do wykorzystania przez większe rozproszenie istotnych wartości.



Wnioski

Z etapu eksploracyjnej analizy danych wyciągnęliśmy następujące wnioski:

- Dane z większym detalem mają więcej braków danych (są rzadsze) – na miejscu jest znalezienie metody wygładzającej macierz.
- Domeny topologicznie nie są wyznaczone jednoznacznie – należy walidować metody na zasadzie odseparowania znalezionych domen od reszty danych.
- Kolejność fragmentów ma znaczenie. Wiele algorytmów klasteryzujące grafy ważone nie dadzą poprawnych wyników.

Kreacja i wybór metody

Następnym etapem było stworzenie kilku metod do wykrywania domen topologicznych, a następnie wybór najlepszej do doskonalenia.

Porównanie metod wygładzania macierzy

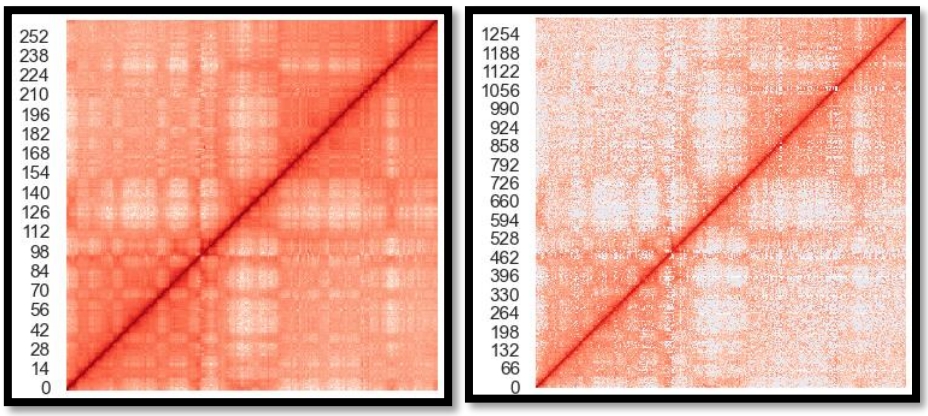
Na początku przyjrzeliśmy się metodom wygładzania macierzy. Szukaliśmy takiego algorytmu, który nie zaburzy wyglądu macierzy, domeny topologicznie nie będą znacznie gorzej widoczne i macierz będzie mniej rzadka.

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

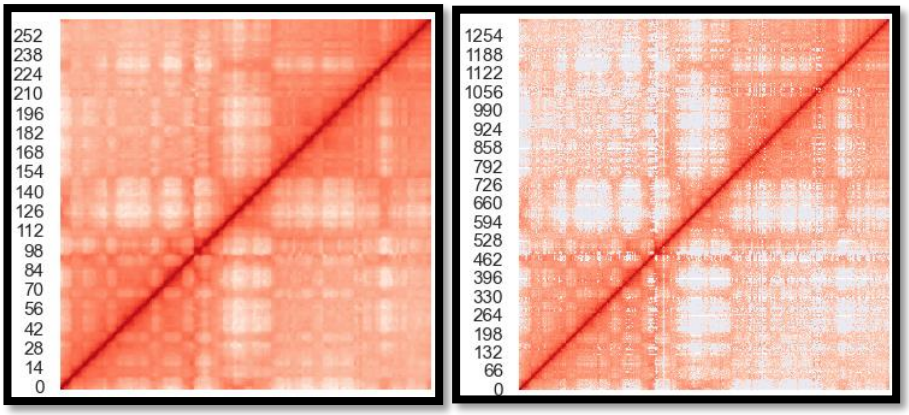
Porównywane były:
- Brak wygładzenia
- Filtr gaussowski, sigma=1
- Filtr średniej, rozmiar jądra=10
- Filtr mediany, rozmiar jądra=3

Parametry były dobrane tak, aby wynik na mniejszej macierzy był jak najlepszy. Ten sam parametr jest dla większej macierzy, aby zaobserwować, jak wartość parametru wpływa na wynik wygładzenia, gdy rozmiar macierzy się zmienia, a parametr zostaje niezmieniony (czyli jak bardzo istotne jest dostosowanie parametru).

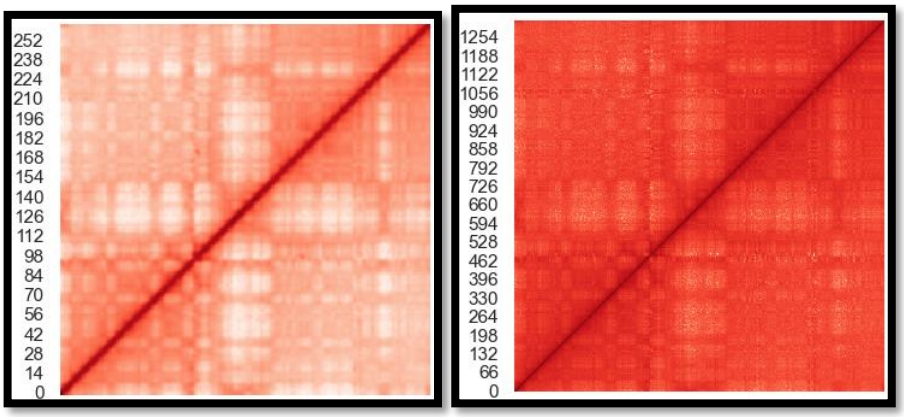
Brak wygładzenia



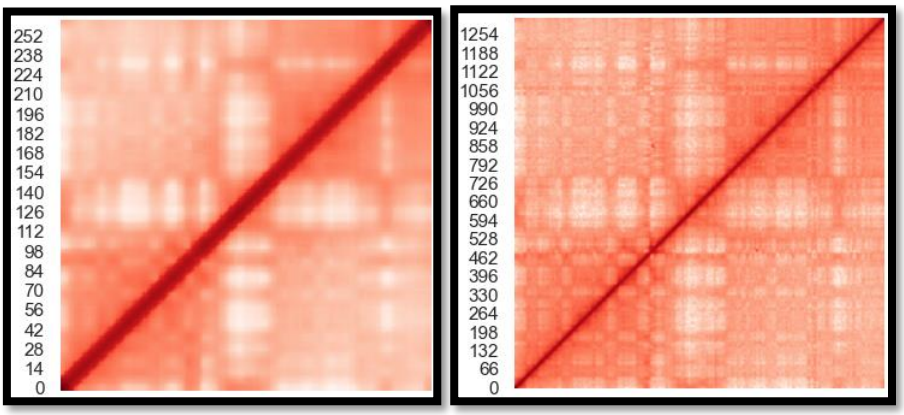
Filtr mediany



Filtr gaussowski



Filtr średniej

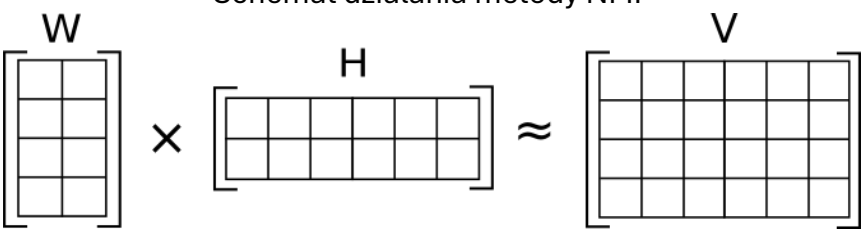


Regulowana grafami dekompozycja macierzy nieujemnej (graph regularized NMF)

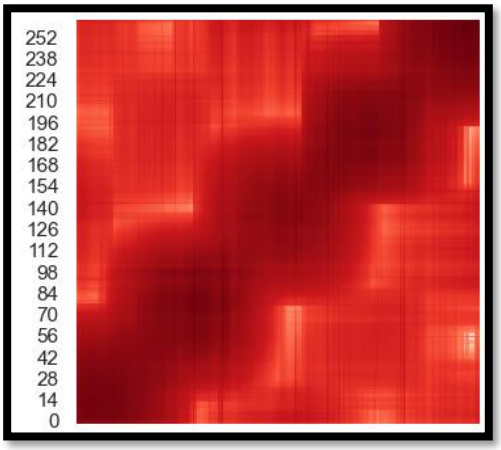
Braliśmy również pod uwagę algorytm dekompozycji macierzy NMF w celu wygładzania. Korzystając z modyfikacji tej metody – regulowania grafami sąsiedztwa, pomysł był taki, że domeny znajdujące się na diagonali będą uwidocznione.

Jednak podstawowa implementacja tego algorytmu nie dała najlepszych rezultatów. Poniżej jest pokazany wynik tego przedsięwzięcia. Parametry były ręcznie dobrane tak (siła regularyzacji, promień sąsiedztwa i ilość składowych), aby efekt był jak najlepszy. Dodatkowym minusem tej metody jest to, że jest bardzo czuła na wartości parametrów, których są aż 3 (łatwo jest przeoczyć ten optymalny).

Schemat działania metody NMF



Najlepszy wynik zastosowania metody



Najbardziej obiecująca była metoda z filtrem średniej, jednak dalsze dobieranie metody i odpowiedniego parametru zostało opisane w późniejszej sekcji.

Po wygładzeniu macierz może przestać być symetryczna (dolna trójkątna połowa będzie się nieznacznie różnić od górnej). W takim wypadku dolna trójkątna połowa była odbijana lustrzanie, przywracając symetrię.

Tworzenie modelu

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

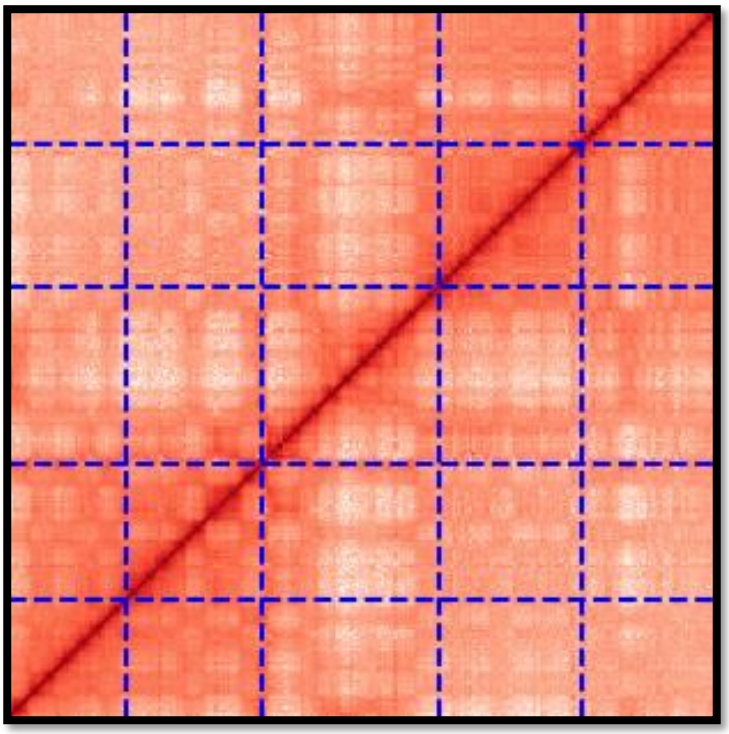
Ocena rezultatów odbywała się w sposób wizualny, następnie najlepsza metody była dalej testowana za pomocą metryk.

Dodatkowo ocenie podlegała łatwość obsługi algorytmu (to znaczy czułość na zmiany parametrów i liczba parametrów – im bardziej różne wyniki po zmianie i im więcej elementów do regulacji, tym trudniej jest dostosować optymalny zestaw wartości).

Metoda 1

Najpierw obliczany jest "insulation score", czyli wskaźnik izolacji dla każdego wiersza macierzy kontaktów. Wskaźnik izolacji dla danego wiersza jest obliczany jako średnia wartość w oknie (fragmentu macierzy) o określonej wielkości, skoncentrowanym wokół tego wiersza. Okno to jest określone przez parametr window_size, który definiuje, jak duży fragment macierzy jest brany pod uwagę do obliczeń. Wynik dla każdego wiersza jest zapisywany w tablicy insulation_scores.Następnie, przy użyciu obliczonych wskaźników izolacji, identyfikowane są lokalne minima, które wskazują na potencjalne granice domen topologicznych. Lokalizacja tych minimów jest realizowana za pomocą funkcji find_peaks z odwróconymi wartościami wskaźników izolacji, co pozwala na znalezienie dolin w wykresie wskaźników izolacji. Znalezione pozycje tych minimów są traktowane jako granice TADu i są zapisywane w tablicy tad_boundaries.

Wynik dawał TADy bardzo podobnej wielkości, co nie było zgodne ze stanem rzeczywistym.



Metoda 2

Na podstawie Directionality Index:

Na początek, dla każdego binu (komórki macierzy) w macierzy kontaktów, obliczana jest miara zwana indeksem kierunkowości (DI). DI określa, czy dany bin bardziej kontaktuje się z upstream (górnymi) czy downstream (dolnymi) binami.

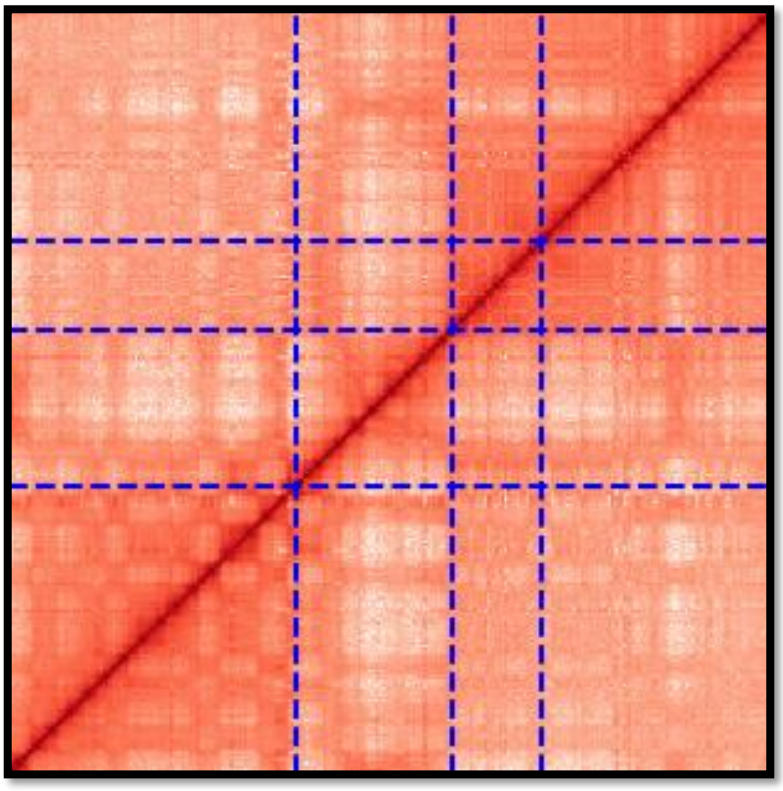
Obliczenia te są wykonywane w następujący sposób:

- Dla każdego binu, zaczynając od window_size do num_bins - window_size, sumowane są kontakty z upstream (od i - window_size do i) i downstream (od i + 1 do i + window_size + 1).
- Całkowita liczba kontaktów jest sumą kontaktów upstream i downstream.
- Jeśli suma kontaktów nie jest zerowa, indeks kierunkowości jest obliczany jako różnica między kontaktami downstream i upstream podzielona przez całkowitą liczbę kontaktów.

Wygładzanie sygnału - indeks kierunkowości jest następnie wygładzany przy użyciu filtra uśredniającego, aby uzyskać bardziej płynny sygnał. W tym celu stosowany jest jednolity filtr 1D z określonym rozmiarem okna wygładzania.

Wykrywanie granic TADów - Wygładzony sygnał DI jest analizowany w celu wykrycia granic TAD. Granice te są identyfikowane jako szczyty w wartościach bezwzględnych wygładzonego sygnału DI, z określoną minimalną odległością między szczytami (peak_distance).

Ta metoda była bardzo czuła na parametry wejściowe.



Metoda 3

Działanie metody można podzielić na następujące kroki:

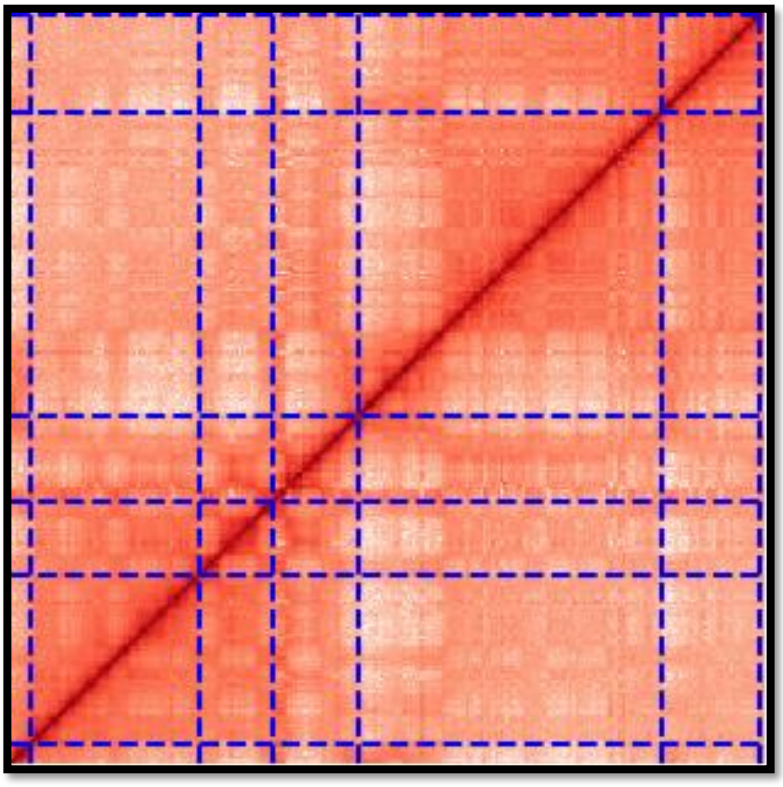
Normalizacja macierzy interakcji:
Macierz interakcji jest przekształcana w taki sposób, aby jej wartości mieściły się w przedziale od 0 do 1. Najpierw od wartości macierzy odejmowana jest minimalna wartość, a następnie wszystkie wartości są dzielone przez maksymalną wartość z macierzy.

Konstrukcja Laplasjanu grafu:
Na podstawie znormalizowanej macierzy kontaktowej tworzony jest Laplasjan grafu. Najpierw tworzona jest macierz wag W, w której sąsiednie węzły (definiowane przez zadany rozmiar sąsiedztwa) są połączone krawędziami o wadze 1. Następnie tworzona jest macierz diagonalna D zawierająca sumy wierszy macierzy W. Laplasjan L jest obliczany jako różnica macierzy D i W.

Regulowana grafami dekompozycja na nieujemne macierze:
Zastosowana jest metoda dekompozycji na nieujemne macierze (NMF) z regularyzacją grafową. Model NMF dekomponuje macierz kontaktową na dwie macierze H i W. Proces iteracyjny dostosowuje H i W, minimalizując błąd aproksymacji i uwzględniając regularyzację przez Laplasjan grafu, co wpływa na spójność wykrytych struktur.

Identyfikacja TADów:
Na podstawie macierzy H identyfikowane są granice domen topologicznych. Różnice między wartościami średnimi są analizowane w celu zlokalizowania miejsc największych zmian, które są interpretowane jako granice TADów.

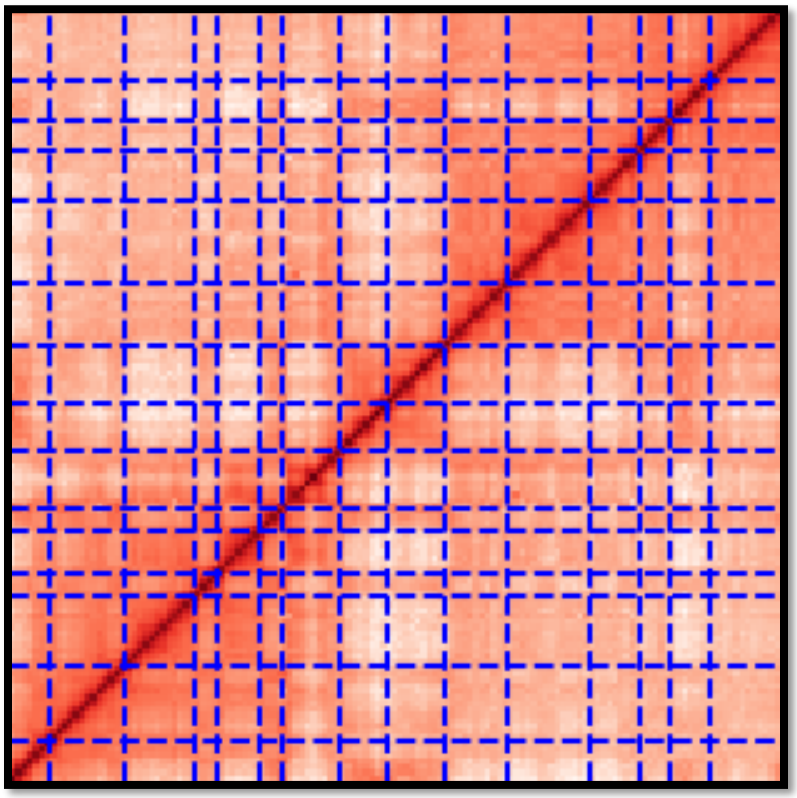
Podobnie jak poprzednie podejście, niewielka zmiana parametrów dawała o wiele gorsze rezultaty.



Metoda 4

Metoda wykorzystująca funkcję „louvian_communities” w bibliotece networkx.

Brak regulowania parametrami i duże rozdrobnienie znalezionych domen powoduje wyniki, które jakościowo nie są odpowiednie do naszego celu.



Metody 5-8

W tych algorytmach została wykorzystana metoda do obliczania średnich wartości sum elementów wewnątrz kwadratowych podmacierzy leżących wzdłuż głównej przekątnej danej macierzy. Proces ten odbywa się w kilku krokach.

Na początku, metoda przyjmuje dwie wartości wejściowe: macierz (matrix) oraz rozmiar kwadratowej podmacierzy (r). Następnie, sprawdza wymiar macierzy (n) oraz tworzy pustą listę, która będzie przechowywać wyniki obliczeń.

Kolejnym krokiem jest iteracja przez macierz od pierwszego elementu do elementu, gdzie można jeszcze wyodrębnić pełną kwadratową podmacierz o rozmiarze r. W każdej iteracji, metoda wyodrębnia podmacierz (jądro) o rozmiarze r x r, zaczynając od elementu leżącego na przekątnej głównej. Następnie, oblicza sumę wszystkich elementów w tej podmacierzy i dzieli ją przez liczbę elementów w podmacierzy (r x r), co daje średnią wartość elementów w podmacierzy. Wynik ten jest dodawany do listy. Na koniec, metoda zwraca listę, która zawiera średnie wartości sum elementów dla każdej wyodrębnionej podmacierzy wzdłuż przekątnej głównej macierzy wejściowej.

Metoda 5

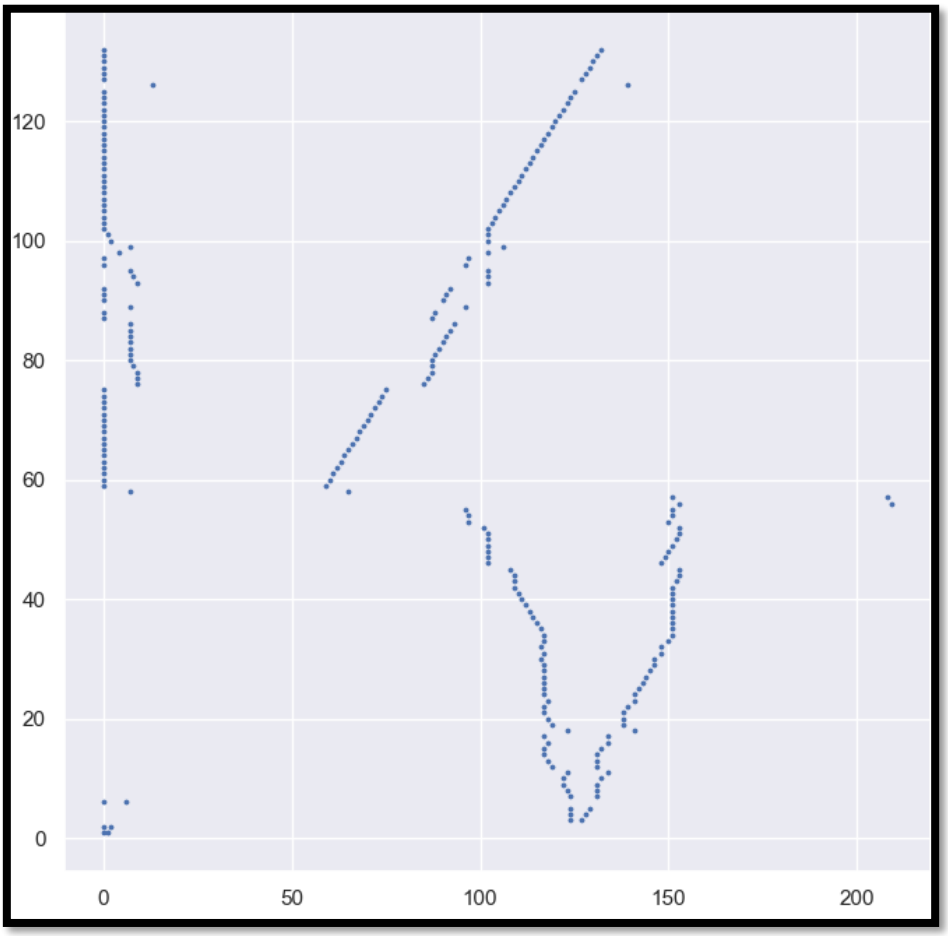
Metoda najpierw definiuje pustą listę result, która będzie przechowywać wyniki. Następnie, dla każdej podlisty w liście głównej, metoda wykonuje następujące kroki:

- Znajduje indeks maksymalnej wartości w podliście.
- Inicjalizuje zmienne second_max oraz second_max_index jako None.
- Przeszukuje podlistę zaczynając od pozycji tuż za maksymalną wartością, dodając do indeksu wartość r, gdzie r jest numerem bieżącej podlisty (indeks podlisty w liście głównej), aż do końca podlisty. Celem jest znalezienie drugiej co do wielkości wartości, która jest różna od maksymalnej wartości. Gdy taka wartość zostanie znaleziona, zapamiętuje jej indeks.
- Jeśli nie udało się znaleźć drugiej co do wielkości wartości w kroku powyżej, przeszukuje podlistę wstecz, od pozycji tuż przed maksymalną wartością, odejmując od indeksu wartość r, aż do początku podlisty, w poszukiwaniu wartości różnej od maksymalnej. Gdy taka wartość zostanie znaleziona, zapamiętuje jej indeks.
- Dodaje parę indeksów (maksymalnej i drugiej maksymalnej wartości) do listy result.

Na koniec metoda zwraca listę result, zawierającą dla każdej podlisty parę indeksów maksymalnej oraz drugiej co do wielkości wartości.

Ta metoda nie daje jasno określonego indeksu w macierzy, gdzie jest granica domen, oraz wybór liczby wartości maksymalnych (w tym przypadku były 2) nie jest oczywisty.

Wykres przestawia na osi y – rozmiary jądra (czyli też numer iteracji), a na osi x – indeks macierzy (czyli numer fragmentu rozpatrywanej sekwencji DNA).



Metoda 6

Metoda ta ma za zadanie znajdywanie domen topologicznych poprzez analizę wartości w listach zawartych w liście głównej.

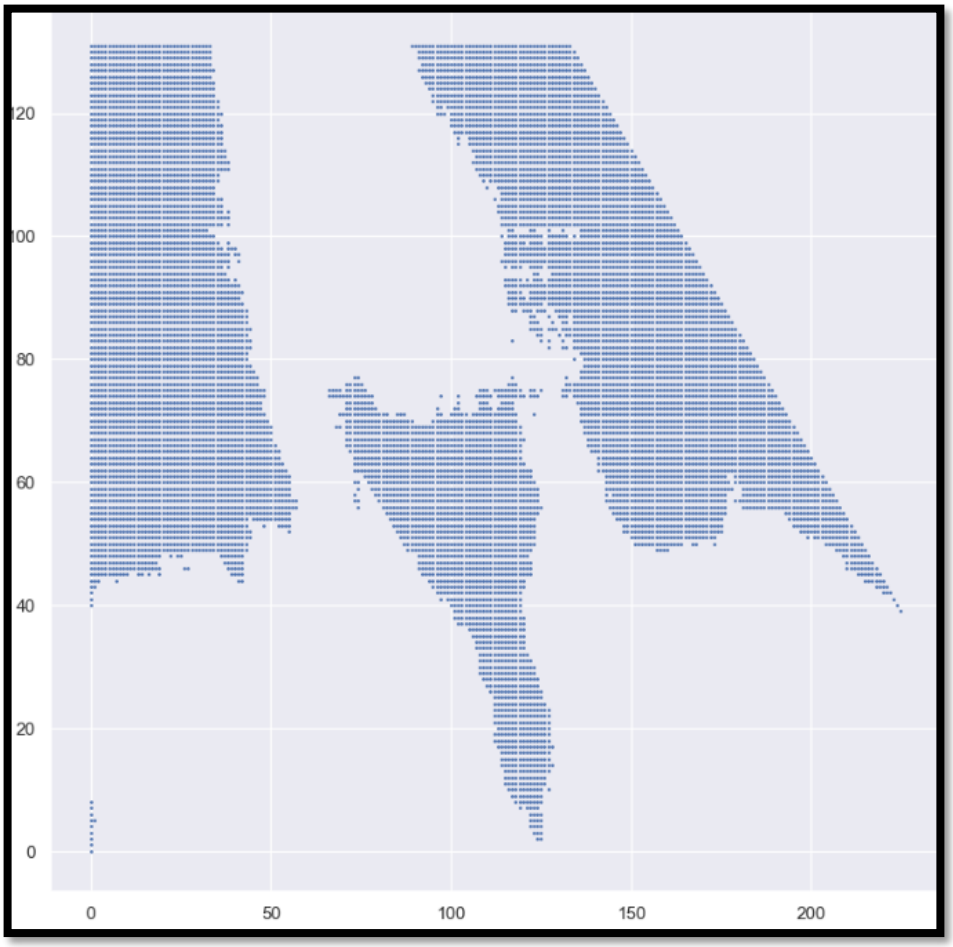
Proces działa w następujący sposób:

- Funkcja get_max_and_less_than_10_percent przyjmuje listę list, gdzie każda podlista zawiera liczby.
- Dla każdej podlisty określa wartość maksymalną (max_value) oraz próg wartości, który wynosi 95% wartości maksymalnej (threshold).
- Następnie identyfikuje indeksy wszystkich elementów w podliście, które są większe lub równe wartości progowej.
- Indeksy te są zapisywane w nowej liście.
- Dodatkowo tworzona jest lista rozmiarów, która zawiera numery wierszy (indeksów podlist) odpowiadające ilości wybranych indeksów z każdej podlisty.
- Funkcja zwraca dwie listy: zawierającą indeksy wartości spełniających warunki informującą o ilości wybranych indeksów w każdej podliście.

W skrócie, metoda ta znajduje indeksy wartości w każdej podliście, które są bliskie wartości maksymalnej.

Wynikowe dane są trudne do analizy – nie jest jasno określona wartość indeksu z granicą domeny.

Wykres przestawia na osi y – rozmiary jądra (czyli też numer iteracji), a na osi x – indeks macierzy (czyli numer fragmentu rozpatrywanej sekwencji DNA).



Metoda 7

Metoda ta ma na celu podział tablicy na k podtablic, tak aby suma największej z tych podtablic była jak najmniejsza.

Działanie metody:

- Najpierw oblicza się tablicę prefix_sum, która przechowuje sumy prefiksowe tablicy wejściowej. Suma prefiksowa to suma wszystkich elementów od początku tablicy do danego indeksu włącznie. Ułatwia to późniejsze obliczenia sum fragmentów tablicy.
- Następnie tworzy się dwie tablice pomocnicze: dp i partition. Tablica dp przechowuje minimalną możliwą wartość największej sumy w jj-tej podtablicy dla pierwszych i elementów tablicy wejściowej. Tablica partition przechowuje indeksy, które określają, gdzie należy podzielić tablicę, aby uzyskać optymalny podział.
- Metoda wypełnia tablicę dp, iterując przez wszystkie możliwe końcowe indeksy i i liczby podtablic jj. Dla każdego z tych przypadków rozważa wszystkie możliwe podziały, porównując obecnie uzyskaną wartość największej sumy z wcześniej obliczonymi wartościami. Jeśli nowa wartość jest mniejsza, aktualizuje tablicę dp i zapisuje punkt podziału w tablicy partition.
- Po wypełnieniu tablicy dp i partition, metoda rekonstruuje podtablice. Zaczyna od końca tablicy wejściowej i używając informacji z tablicy partition, odtwarza podtablice, przesuwając się do początkowego indeksu podtablicy określonego w tablicy partition.Ostatecznie metoda zwraca listę kk podtablic, w których suma największej z tych podtablic jest minimalna.

Wraz ze wzrostem rozmiaru jądra wartości wyznaczonych granic maleją. Ten spadek wartości ma różne tempo dla różnych granic, przez co nie jest jasne, która wartość jest tą najlepszą.

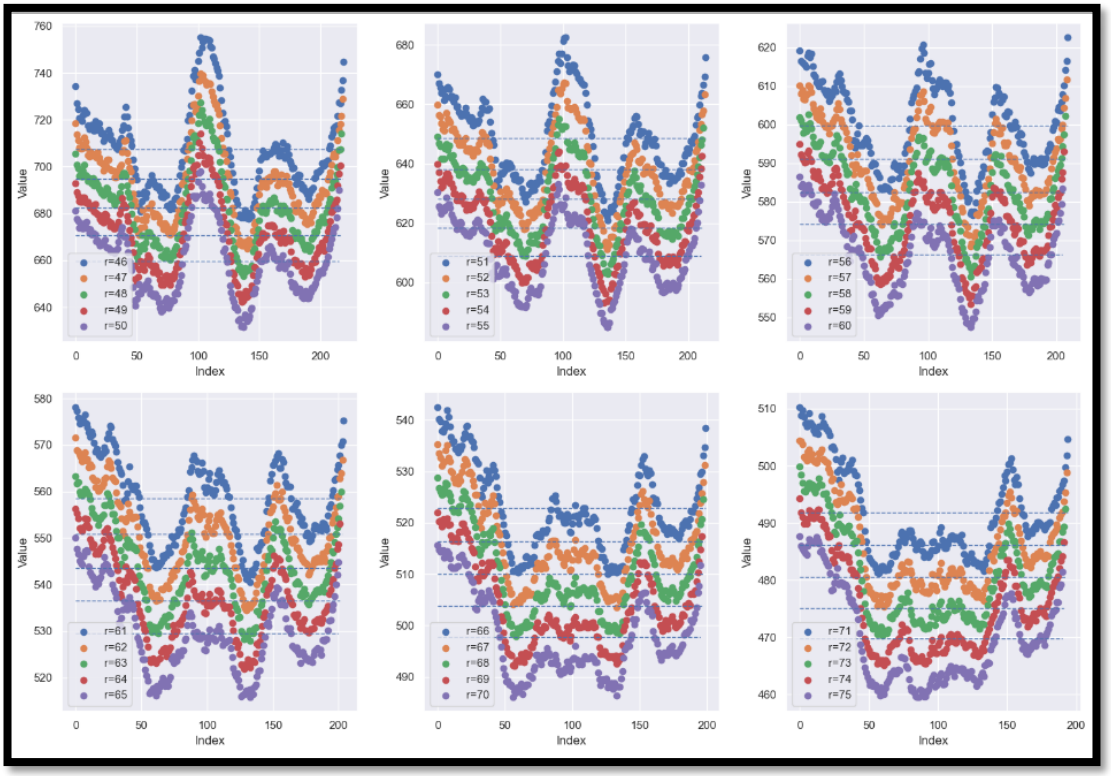
```
r: 20: 49 100 145 197
r: 21: 49 100 145 196
r: 22: 48 99 144 195
r: 23: 48 99 144 195
r: 24: 48 99 144 194
r: 25: 48 98 143 193
r: 26: 48 98 142 192
r: 27: 47 97 141 191
r: 28: 47 97 141 190
r: 29: 47 96 140 189
r: 30: 47 96 140 189
r: 31: 47 96 140 188
r: 32: 47 95 139 187
r: 33: 46 94 138 186
r: 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

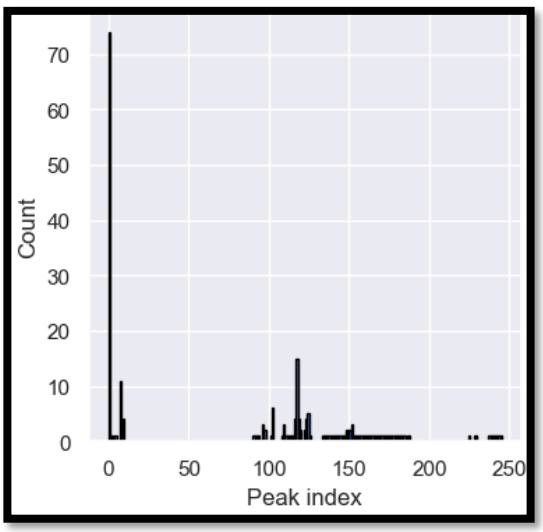
Działanie metody:

- Najpierw dla każdej listy obliczana jest średnia wartość.
- Następnie separowane są wartości większe od średniej, a pozostałe ustawiane są na zero.
- Szukane są ciągłe sekwencje, w których więcej niż 10% wartości jest niezerowych.
- Dla tych sekwencji znajdywana jest maksymalna wartość.
- Indeksy odpowiadające maksymalnym wartościom są umieszczane w liście.
- Na końcu metoda zwraca listę, która zawiera listy indeksów odpowiadających maksymalnym wartościom dla każdej z początkowych list.

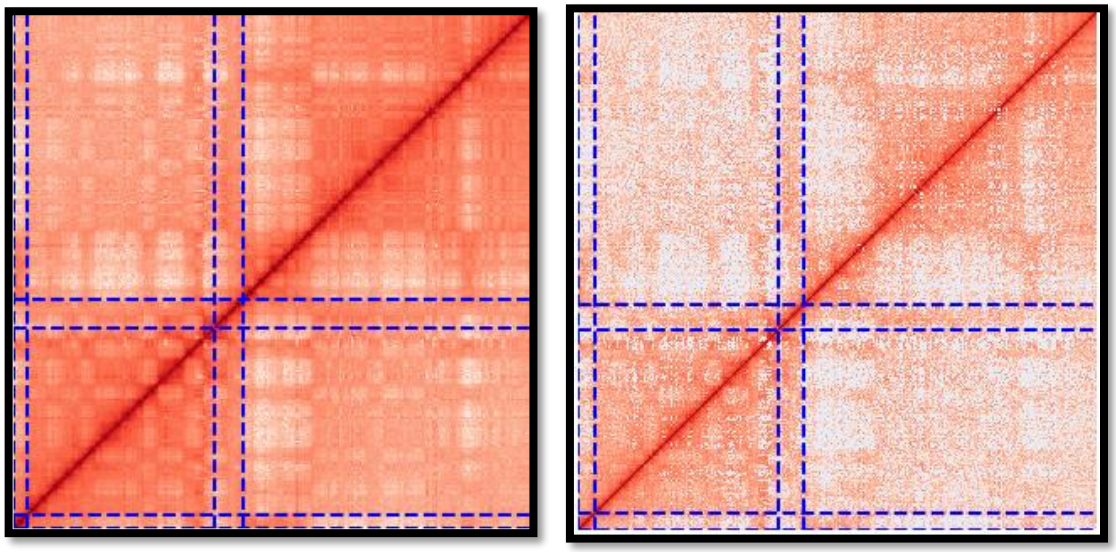
Poniżej jest pokazany fragment zestawu wykresów, gdzie x – indeks wiersza (lub kolumny) macierzy (czyli punkt „zaczepienia" jądra sumującego, idącego po diagonali), y – wartość zliczona opisaną metodą z jądrem, kolor – rozmiar jądra (i numer iteracji), pozioma linia – średnia wartość dla danego koloru.



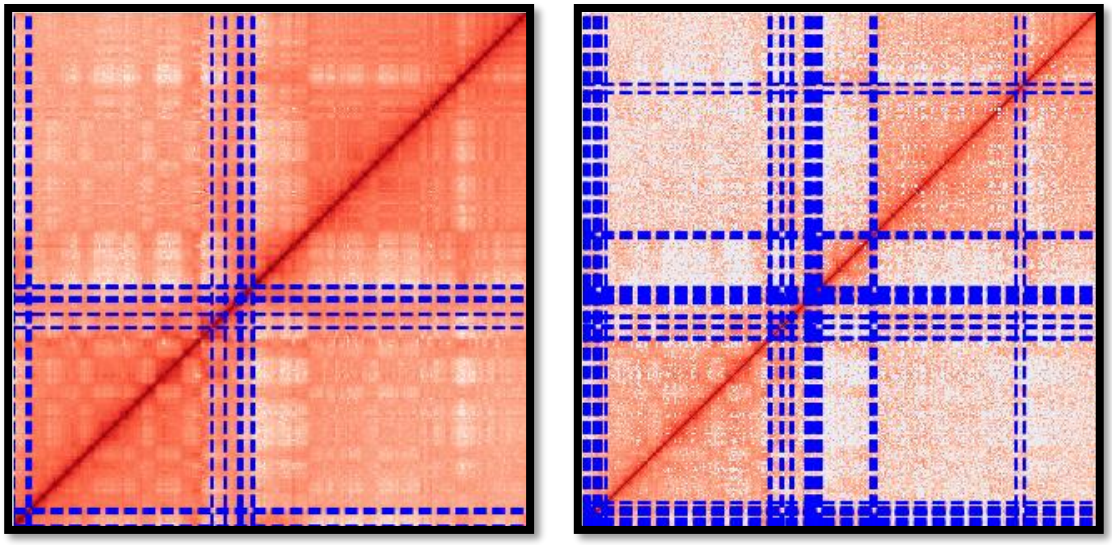
Następnie, te wyniki są sortowane w celu znalezienia najczęstszych maksymalnych wartości – poniżej histogram tych maksymalnych wartości.



Poniżej przykładowe wyniki tej metody.



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



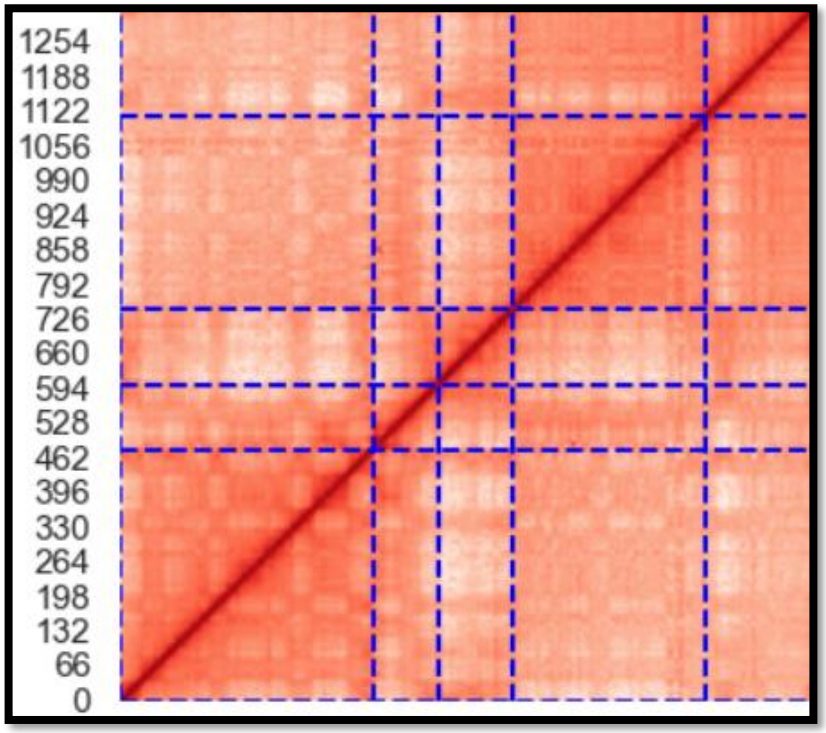
Na szczęście istnieje na to rozwiązanie – można scalić ze sobą domeny, jeśli ich granice są zbyt blisko siebie. Implementacja takiego rozwiązania w poprawny sposób sprawi, że granice będą wyznaczone w sposób jednoznaczny i nie będzie wykrywanych domen o bardzo małym rozmiarze.

W celu poprawienia wyników została zaimplementowana następująca funkcja:

- Inicjalizacja tablicy wynikowej: Na początku tworzona jest pusta lista RESULT, która będzie przechowywać elementy spełniające warunki.
- Pętla główna: Funkcja przechodzi przez każdy element tablicy wejściowej IN za pomocą pętli.
- Dodawanie elementów: Każdy element IN[i] jest dodawany do tablicy RESULT.
- Sprawdzanie warunków odległości: Gdy długość RESULT osiągnie co najmniej 2, następuje sortowanie tablicy RESULT. Następnie iteruje się po posortowanej tablicy i sprawdza się warunek odległości pomiędzy sąsiednimi elementami. Jeśli warunek jest spełniony, mniejsza z wartości jest zachowywana, a większa jest usuwana z tablicy. Proces ten zapobiega przekroczeniu odległości dist pomiędzy elementami w RESULT.
- Warunek liczby elementów: Jeśli długość RESULT osiągnie wartość tad_num, funkcja zwraca RESULT.
- Zwracanie wyniku: Jeśli nie zostanie znaleziona odpowiednia liczba elementów TADs do momentu zakończenia pętli, funkcja zwraca aktualny stan RESULT.

Podsumowując, funkcja ma na celu znalezienie i zwrócenie tablicy elementów z tablicy wejściowej IN, które spełniają warunki odległości pomiędzy sobą oraz liczby elementów.

Tak zaimplementowana poprawka skutkuje metodą, której wyniki są bardzo zadowalające. W celu dalszego doskonalenia algorytmu metoda została sparametryzowana i stworzony został model posiadający funkcjonalność
Odczytu i obróbki danych,
Rysowania macierzy interakcji (z i bez granic domen),
Wykrywania domen za pomocą metody 8.



W następnych etapach odbyło się testowanie modelu w celu doboru jak najlepszych parametrów.

Testowanie i strojenie modelu

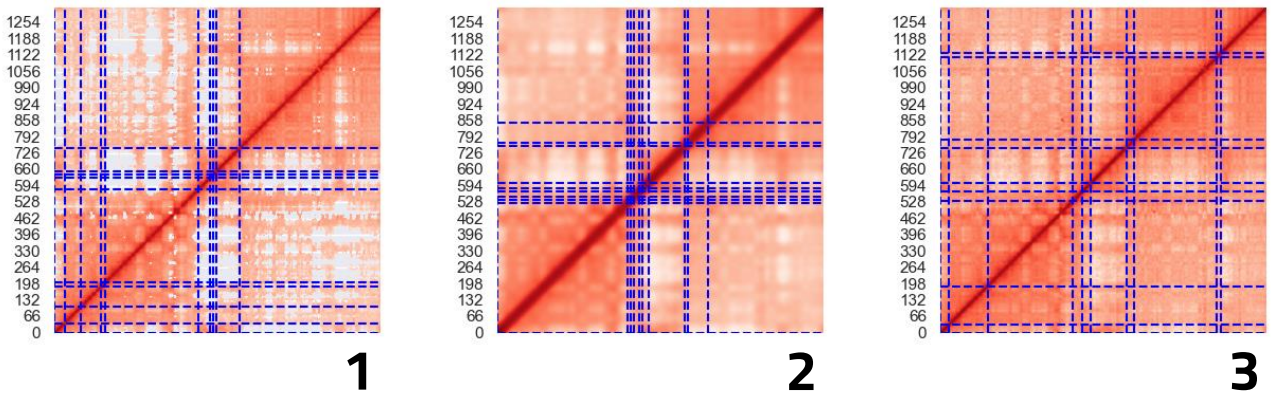
Po stworzeniu modelu przyszedł czas na dobranie odpowiednich parametrów. Nasz oparty jest na 4 parametrach:

- rodzaj filtra
- wskaźnik wygładzenia
- maksymalna liczba TADów
- minimalny rozmiar TADu

Parametry zostały potraktowane jako niezależne od siebie. Przeprowadziliśmy testy odnoszące się do kolejnych paramentów. Wyniki i wnioski zostały przedstawione na następnych stronach.

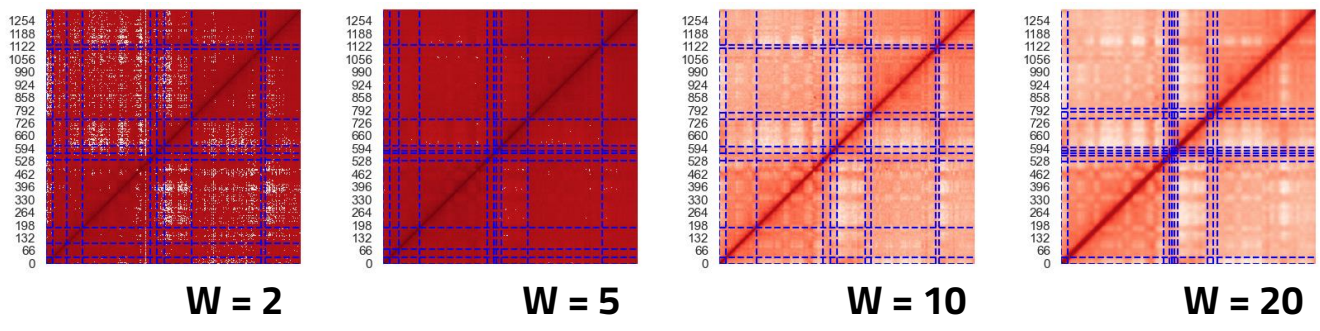
Wybór zastosowanego filtra

Poniżej zostały przedstawione wyniki 3 testów dotyczące 3 filtrów dostępnych w naszym modelu (1 – filtr mediany, 2 – filtr gaussowski, 3 – filtr średniej). Przy tych samych pozostałych parametrach widzimy, że najlepsze wyniki dostajemy dla filtra 3, czyli używającego średnią. Jednak według wskaźnika TAD-adjR^2 (który zostanie opisany w dalszej części), to filtr 2 - gaussowski jest najlepszy.



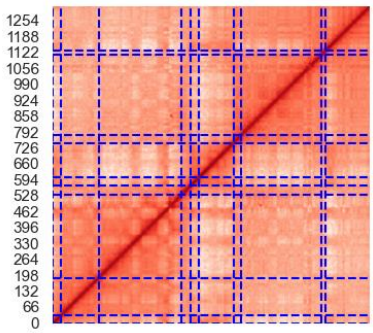
Wybór wskaźnika wygładzenia

Teraz testy dotyczyły wskaźnika wygładzenia. Doświadczenia zostały przeprowadzone dla następujących wartości: 2, 5, 10 i 20. Wyniki zostały pokazane na obrazkach. Przy tych samych pozostałych parametrach widzimy, że im większy wskaźnik tym lepsze wyniki, ale do pewnego momentu. Najlepsza wartość to 10.

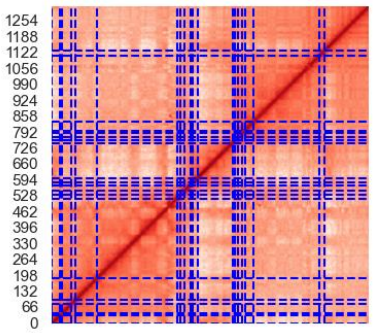


Porównanie maksymalnej liczby TADów

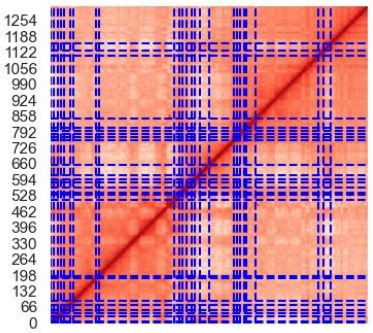
Na stan naszej wiedzy to wyraźne dla ludzkiego oka TADy są najlepszym wyborem, jednak po uwagach prowadzącego doszliśmy do wniosku, że niekoniecznie tak jest. Według nas najlepszy wynik widzimy na obrazku 1, w którym maksymalna liczba TADów była wyznaczona przez 1% wymiaru macierzy. Jednak dokładniejszym i bardziej poprawnym założeniem powinno być wybranie ok. 4% jako maksymalnej liczby TADów.



ok. 1%



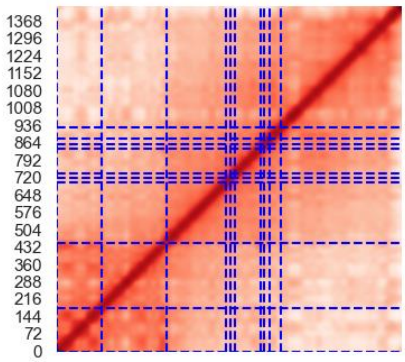
ok. 2%



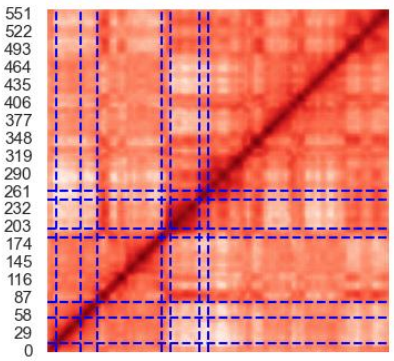
ok. 4%

Wyniki modelu

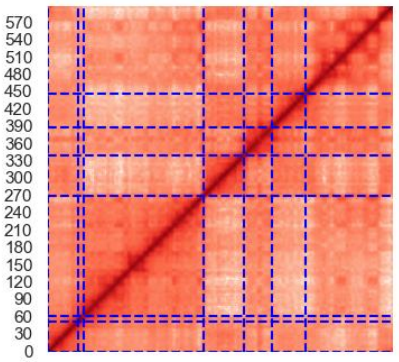
W następnej sekcji pokażemy przykłady wywołania naszego modelu i reprezentację wyników na macierzach kontaktów. Oczywiście znalezione TADy są wyznaczone poprzez niebieskie przerywane linie tworzące prostokąty na przekątnej. Wykonaliśmy 3 próby dla losowo wybranych chromosomów – 12, 19 i 20. Są to macierze przedstawiające całe chromosomy.



Chromosom 12
Liczba znalezionych TADów - 10



Chromosom 19
Liczba znalezionych TADów - 8



Chromosom 20
Liczba znalezionych TADów - 7

Porównanie z innymi metodami

W ostatniej części prezentacji skupiliśmy się na porównani u naszej metody z innymi dostępnymi metodami do znajdowania TADów. Porównywaliśmy się z metodami Arrowhead (od juicer tools – aplikacja javowa) oraz SpectralTAD (paczka dostępna dla języka R). Metody porównaliśmy dla 5 losowych chromosomów. Wykonaliśmy porównanie za pomocą wskaźników i miar oraz wizualnej reprezentacji znalezionych TADów na macierzy kontaktów.

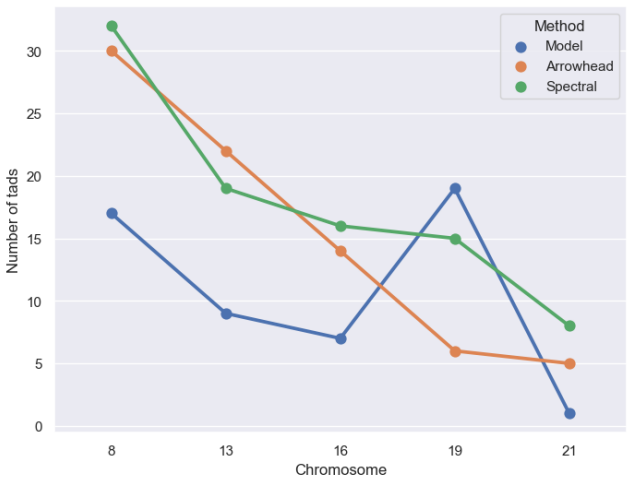
Arrowhead

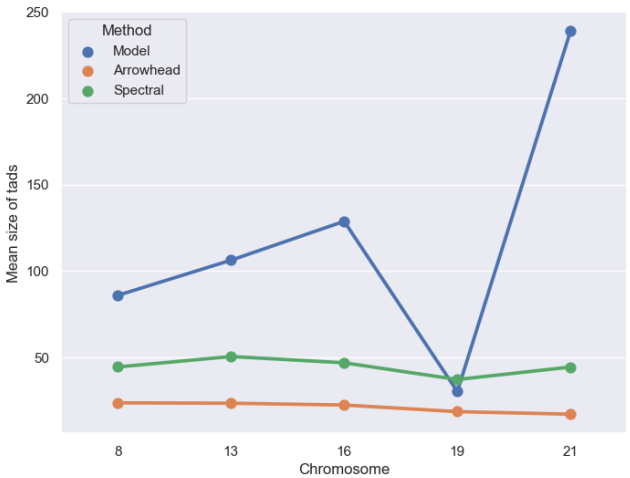
1. **Analiza Macierzy Kontaktów** - Na początku metoda analizuje macierz kontaktów Hi-C, która reprezentuje częstotliwość kontaktów między różnymi regionami genomu. Każda komórka w macierzy reprezentuje liczbę kontaktów między dwoma regionami genomu.
2. **Identyfikacja Zmian Kontaktów** - Metoda identyfikuje punkty, w których występują znaczne zmiany w liczbie kontaktów. Te zmiany mogą wskazywać na granice między różnymi TADami.
3. **Tworzenie Profilu** - Następnie metoda tworzy profil kontaktów, który reprezentuje zmiany w liczbie kontaktów wzdłuż chromosomu. Na podstawie tego profilu można wykryć potencjalne granice TADów.
4. **Detekcja Granic TADów** - Granice TADów są identyfikowane w miejscach, gdzie profil kontaktów wykazuje największe zmiany. Te punkty są interpretowane jako granice oddzielające różne TADy.

SpectralTAD

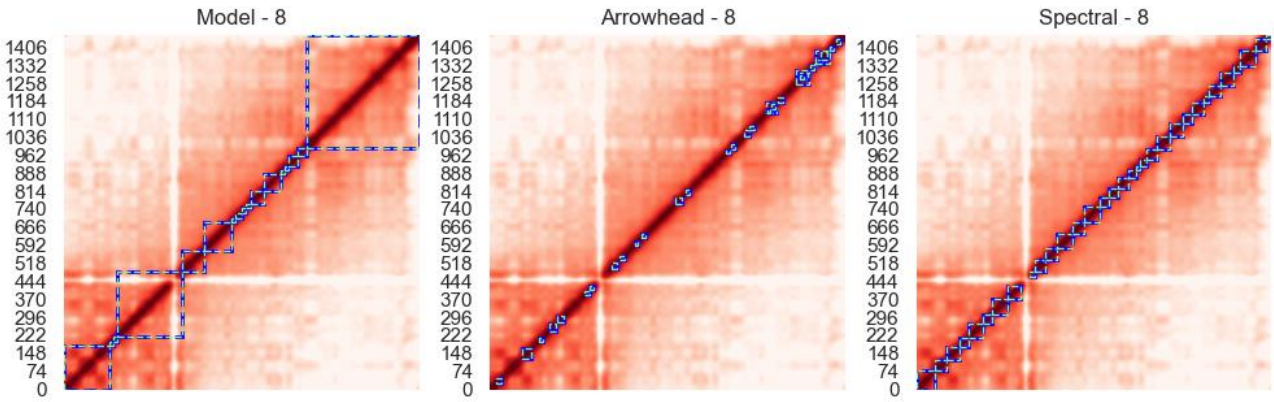
1. **Analiza Spektralna** - Metoda ta rozpoczyna się od analizy spektralnej macierzy kontaktów Hi-C. Analiza spektralna polega na dekompozycji macierzy na zestaw wektorów własnych i wartości własnych, które reprezentują różne wzorce kontaktów w danych.
2. **Wybór Wektorów Własnych** - Spośród wszystkich wektorów własnych, metoda wybiera te, które najlepiej reprezentują strukturalne cechy genomu, takie jak TADy. Wektory te są używane do stworzenia nowej reprezentacji danych, która podkreśla granice TADów.
3. **Klasteryzacja** - Na podstawie wybranych wektorów własnych, metoda przeprowadza klasteryzację regionów genomu. Regiony, które mają podobne wzorce kontaktów, są grupowane razem, co pozwala na identyfikację TADów.
4. **Identyfikacja Granic TADów** - Granice TADów są identyfikowane jako punkty, w których kończą się jedne klastry i zaczynają inne. Dzięki temu można określić, które regiony genomu tworzą poszczególne TADy.

Wyniki przy resolution = 100000 dla losowych chromosomów





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

TAD-adjR^2

	Model	Arrowhead	SpectralTAD
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

Dla TAD-adjR^2 wyższe wyniki wskazują na lepszą jakość wykrytych TADów. Nasza metoda jest porównywalna z metoda Arrowhead, jednak nie otrzymały one zbyt dobrych rezultatów. SpectralTAD radzi sobie dużo lepiej.

Delta Contact Count

	Model	Arrowhead	SpectralTAD
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

Wyższe wartości wskaźnika DCC wskazują na lepszą jakość wykrytych TADów. Tym razem najlepszą okazała się metoda Arrowhead, natomiast nasza metoda trochę odstaje od pozostałych.

Davies-Bouldin index

	Model	Arrowhead	SpectralTAD
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

Dla DBI wartości dodatnie <1 wskazują na dobrą jakość klastrowania, wartości <2 dają zadowalające klastrowanie, natomiast wartości wyższe to złe klastrowanie. Nasza metoda mieści się poniżej 2, więc jesteśmy zadowoleni z wyników.

Measure of Concordance

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

MoC pokazuje jak zbliżone są osiągnięte wybory TADów, widzimy że nasza metoda jest bardziej podoba do SpectralTAD.

Bibliografia

Grafika struktury chromatyny:
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)

Metoda HiC - obraz:
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

Grafika NMF:
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