Zawęźlenie DNA i jego zastosowania

Opiekunowie.

- dr hab. Maciej Borodzik, MIM, matematyk, specjalista z teorii węzłów.
- dr hab. Dariusz Plewczyński, CeNT, biolog, biolog, specjalista z genomiki obliczeniowej, struktury trójwymiarowej chromatyny, biofizyki

Cel badań. Podstawowym celem jest badanie struktur topologicznych w łańcuchach chromatyny (DNA z białkami strukturalnymi) i ich wpływu na biologię komórki, w szczególności na ekspresję genów w zależności od zawęźlenia. Jako dane wejściowe bierzemy dwumiarową macierz interakcji, określającą które części łańcucha chromatyny są ze sobą połączone (fizyczne kontakty). Standardowe techniki poszukiwania struktur topologicznych opierają się o trójwymiarowe modelowanie na podstawie danych wejściowych i postulatów minimalizowania energii. Jest to proces bardzo czasochłonny, większość modeli bierze pod uwagę tylko część oddziaływań ze względu na złożoność obliczeń.

Pierwszym krokiem w badaniach jest zastosowanie zaawansowanych metod topologicznych do zbadania, czy sama macierz interakcji nie wymusza pojawienia się węzła nie z przyczyn energetycznych, ale z przyczyn czysto topologicznych. Wiedza o tym, że konkretne fragmenty łańcucha są blisko siebie, może być wystarczająca do stwierdzenia, że gdzieś w łańcuchu DNA pojawi się węzeł, niezależnie od algorytmu i sposobu wizualizacji. Co więcej, jeśli łańcuch DNA zostanie przecięty za pomocą topoizomerazy, a następnie sklejony, ale macierz interakcji się nie zmieni, być może węzeł zniknie, ale wtedy będzie musiał pojawić się w innym miejscu.

Topologiczne badanie zawęźleń DNA nie doprowadzi do stwierdzenia, że węzeł pojawia się w konkretnym miejscu. Spodziewany wynik polega raczej na wskazaniu np. trzech par fragmentów łańcucha chromatynowego, z których co najmniej jedna jest zaczepiona. Oznacza to, przynajmniej w teorii, że spośród genów tych kodowanych w tych sześciu fragmentach łańcucha co najwyżej cztery będą aktywne jednocześnie (zakładając, że zaczepione pary są nieaktywne).

Potencjalne zastosowania tej teorii są niezwykle szerokie. Można badań np. defekty łańcucha DNA polegające na wyłączeniu pewnego oddziaływania, bądź pojawieniu się dodatkowego. To oznacza zmianę topologii łańcucha przy braku mutacji, a więc zmianę ekspresji genów. Z drugiej strony, mutacje genetyczne również mogą oddziaływać na topologię łańcucha, co daje przynajmniej teoretyczną możliwość wyjaśniania mechanizmu niektórych chorób genetycznych.

Plan badań. W pierwszej kolejności zamierzamy pracować nad macierzami oddziaływań DNA ludzkiego pochodzącymi z doświadczeń populacyjnych genomiki trójwymiarowej (HiC, ChIA-PET, HiChIP), głównie na oddziaływaniach mediowanych przez białko CTCF o skierowaniu motywów typu convergent, włączając oddziaływania rzadsze, a więc typu: tandem left, tandem right i divergent. Następnie uwzględnimy kontakty mediowane przez inne białka strukturalne i czynniki transkrypcyjne. Włączenie tych typów nie stanowi żadnego algorytmicznego problemu, a obliczenia wciąż mieszczą się w możliwościach dobrego komputera przenośnego. Znaleziono już potencjalnie zawęźlone struktury w niektórych łańcuchach chromatynowych (chr1, chr10). Należy teraz, będzie to pierwszym zadaniem doktoranta, sprawdzić, czy te potencjalnie zawęźlone struktury przekładają się na statystyczne własności ekspresji genów. Z początku należy pracować na prostych zawęźlonych

strukturach, które można zrozumieć bez użycia komputera, następnie przejść do badania bardziej zaawansowanych.

Poszukiwanie węzłów można rozszerzyć na łańcuchy DNA innych organizmów żywych. Szczególnie bakterie mogą obfitować w struktury zawęźlone, te struktury ponadto mogą zmieniać się w sposób dynamiczny.

Najważniejszym celem, ale też i najtrudniejszym, jest badanie zawęźleń ludzkiego DNA opierając się na danych typu Single Cell, a więc konkretnej komórki. Główna trudność polega na zdobyciu danych o odpowiednio dużej rozdzielczości. Struktury topologicznie zawęźlone nie mogą bowiem zostać znalezione gdy liczba oddziaływań jest stosunkowo mała, nawet jeśli przy wizualizacji pojawiają się węzły.

Metodyka badań. Program badawczy łączy trzy dyscypliny naukowe: biologię, matematykę i informatykę. Dane dotyczące macierzy interakcji są przetłumaczone na język teorii grafów a następnie przetwarzane komputerowo używając oprogramowania opracowanego przez dr. hab. Marcina Pilipczuka i dr. Michała Pilipczuka z Instytutu Informatyki UW. Wyniki tych opracowań są przetłumaczone na język topologii przez dr. hab. Macieja Borodzika, następnie zespół dr. hab. Dariusza Plewczyńskiego podaje interpretację biologiczną oraz zbuduje modele trójwymiarowe. Praca doktoranta będzie więc polegała na analizie danych topologicznych w kontekście zastosowań biologii, tak więc będzie miała charakter interdyscyplinarny niezależnie od projektu. Należy też dodać, że pełne przetłumaczenie języka grafów na język topologiczny oraz jego implementacje trójwymiarową jeszcze nie zostało dokonane i również jest wyzwaniem.

Ekspertyza dr hab. Dariusza Plewczyńskiego, prof. UW skupia się na symulacjach biofizycznych genomu ludzkiego wykorzystujących informację jednowymiarową (dane epigenomiczne, warianty strukturalne, sekwencyjne motywy wiążące białka architektoniczne oraz czynniki transkrypcyjne) oraz dwuwymiarową (mapy kontaktów binarnych między regionami chromatyny otrzymane z masowych danych trójwymiarowej genomiki).

Znaczenie projektu. Końcowym celem jest funkcjonalna interpretacja otrzymanych wyników toplogicznych (np. stopnia zawęźlenia domen genomicznych). Spodziewamy się pokazania zależności między miarami topologicznymi a poziomem wyciszenia danego regionu DNA. Wstępne wyniki pokazują złożoność modelu wiążącego poziom eksprtesji mRNA oraz cechy strukturalne (takie jak odległości fizyczne między regionami regulatorowymi i genami). Analiza strukturalna porównująca konformacje trójwymiarowe wydaje się nie być właściwa. Niezbędne jest opracowanie nowej metody opisu zróżnicowania strukturalnego, bazującego właśnie na topologii i teorii węzłów, co spodziewamy się że pozwoli zinterpretować różnice w poziomie ekspresji genów zlokalizowanych w wybranych domenach genomicznych.

Dodatkowo w ostatnim etapie badań przeprowadzimy studium mapujące dane populacyjne (zarówno w populacji zdrowej - badanej w projekcie 1000 Genomes, jak i populacji osób chorych - badanych w ramach kohort GWAS: genome wide association studies) na strukturalnie istotne regiony sekwencji genomicznej, ze szczególnym uwzględnieniem mocno zapętlonych domen. Zwrócimy szczególną uwagę na choroby autoimmunologiczne (takie jak cukrzyca typu 1), oraz nowotwory (wykorzystując masowe dane z publicznie dostępnej bazy TCGA) - np. potrójnie negatywny nowotwór piersi, czy białaczki przewlekłe lub ostre dziecięce.