Streszczenie

Przetwarzanie informacji biologicznej wewnątrz jądra komórkowego *metazoa* jest niezwykle złożonym procesem. Integruje ono wiele poziomów jej przechowywania i regulowania, takich jak sekwencja DNA, znaczniki epigenetyczne, elementy cis-regulacyjne oraz trójwymiarową strukturę chromatyny. Rosnące zapotrzebowanie na zaawansowane narzędzia obliczeniowe, umożliwiające poznanie złożonej organizacji genomu, zrozumienie różnic między populacjami oraz między komórkami osób zdrowych i chorych, jest bardziej aktualne niż kiedykolwiek. Pomimo iż obecnie dysponujemy zaawansowanymi metodami eksperymentalnymi pozyskiwania informacji o przestrzennych kontaktach chromatynowych, takimi jak ChIA-PET i Hi-C, ich zastosowanie wciąż wiąże się z wysokimi kosztami i jest czasochłonne, co ogranicza ich wykorzystanie w badaniach w skali populacji ludzkiej. W związku z tym, aby zmniejszyć koszty i zwiększyć dostępność badań nad przestrzenną organizacją genomu, niezbędny jest rozwój odpowiednich metod obliczeniowych.

W odpowiedzi na te potrzeby niniejsza praca prezentuje zaawansowane narzędzia informatyczne umożliwiające modyfikację wzorów kontaktów chromatynowych wynikających ze zmian sekwencji DNA, co umożliwia generowanie i porównywanie różnych modeli 3D odzwierciedlających zróżnicowanie populacyjne wariantów strukturalnych. To innowacyjne narzędzie zostało włączone do serwisu internetowego 3D-GNOME w wersji 2.0, umożliwiając unikalne badania trójwymiarowych struktur chromatyny dla tysięcy genomów ludzkich.

Dodatkowo, w celu zwiększenia wydajności obliczeń, opracowano narzędzie *cudaMMC*, które powstało na bazie algorytmu modelowania 3D-GNOME. Jest to metoda oparta na metodzie symulowanego wyżarzania Monte Carlo, rozbudowana o możliwość masowego zrównoleglania obliczeń na kartach graficznych (GPU). Pozwoliło to na znacznie szybsze generowanie trójwymiarowych struktur chromatyny (do 25 razy szybciej), przy jednoczesnym zachowaniu wysokiej jakości modeli.

W pracy przedstawiono również metodę obliczeniową służącą do tworzenia zespołów modeli 3D, zarówno dla struktur referencyjnych, jak i zmodyfikowanych przez warianty strukturalne. Ta nowatorska technika została zaimplementowana w wersji 3.0 serwisu internetowego 3D-GNOME. Umożliwia ona mapowanie enhancerów oraz promotorów genów na modele 3D, a

także obliczanie zmian w rozkładach odległości między tymi elementami regulatorowymi i genami w strukturach referencyjnych i zmodyfikowanych przez warianty. W celu obsługi generowania zespołów statystycznych modeli 3D oraz przetwarzania dużych zestawów danych, w serwisie 3D-GNOME zaimplementowano metodę *cudaMMC*. Obliczenia wykonano na klastrze Eden^{N}, będącym wewnętrznym heterogenicznym wysoko-wydajnym klastrem obliczeniowym HPC wyposażonym w węzły Nvidia DGX A100 i zarządzanym przez oprogramowanie kolejkowe Slurm.

Dzięki tym innowacjom, niniejsza praca dostarcza kompleksową platformę komputerową do badania wpływu wariantów strukturalnych na przestrzenną organizację genomu. Opisane narzędzia stanowią unikatowe źródło wiedzy pozwalającej na zrozumienie wpływu przestrzennej organizacji chromatyny na ekspresję genów, a także na badanie mechanizmów regulacji transkrypcji i chorób.

Abstract

Processing biological information within a metazoan cell nucleus is highly complex, as it must integrate multiple information storage and regulation layers such as DNA sequence, epigenetic marks, cis-regulatory elements, and the 3D structure of chromatin. The demand for advanced computational tools to unravel the intricate organisation of the genome, understand population differences, and discern between healthy and diseased cells is continually growing. While we have advanced experimental methods to obtain chromatin contacts, like ChIA-PET and Hi-C, their application is still costly and time-consuming, limiting their use in population-scale studies. This necessitates the adoption of computational approaches to reduce costs and increase accessibility.

Addressing the need for a sophisticated computational tool to apply changes to the chromatin contact pattern due to modifications of the underlying DNA sequence, this thesis introduces a comprehensive solution that facilitates generating and comparing distinct 3D models underpinned by Structural Variants (SV) driven changes. This innovative tool was incorporated into the 3D-GNOME 2.0 web service, enabling a unique exploration of chromatin 3D structures.

Moreover, to enhance the efficiency of these calculations and the manipulation of large chromatin models, this thesis presents the *cudaMMC* method. This method employs GPU-accelerated computing and the Simulated Annealing Monte Carlo approach, allowing for faster generation of chromatin 3D structures while maintaining model quality.

Furthermore, the study unveils a computational method designed to create ensembles of models for both reference and SV-altered structures. This novel technique, encapsulated within the 3D-GNOME 3.0 web server update, empowers researchers to map enhancers and gene promoters onto the 3D models. As a result, it's possible to calculate changes in the distribution of distances between these genomic features in reference and SV-affected structures. To handle the generation of 3D model ensembles alongside new large datasets, we implemented *cudaMMC* and established calculations on Eden^{N} high performance computing (HPC) cluster, an in-house heterogeneous computing resources equipped with Nvidia DGX A100 nodes and managed by Slurm. Through these advancements, this PhD thesis provides a comprehensive computational platform for studying the influence of structural variants on the genome's spatial organisation. These tools serve as a unique resource for understanding the effect of chromatin spatial organisation on genetic expression and investigating transcriptional regulation and disease mechanisms.