streszczenie

Trójwymiarowa organizacja ludzkiego genomu oraz jej wpływ na ekspresję genów i funkcje komórkowe jest wciąż otwartym zagadnieniem. Struktura genomu jest hierarchiczna, od terytoriów chromosomalnych do gęsto upakowanych regionów tzw. domen topologicznych (TAD, ang. Topologically associating domain), domen genomicznych związanych ze statystycznie istotnymi przestrzennymi kontaktami chromatyny (CCD, ang. Chromatin contact domains), oraz pętli chromatynowych powstających w oparciu o białko CTCF i kohezynę. Zaproponowałem nowatorski model in-silico oparty na teorii wybuchowej perkolacji do modelowania dynamicznego procesu zwijania się genomu ludzkiego. Model ten reprezentuje pętle chromatynowe jako krawędzie w grafach reprezentujących poszczególne chromosomy przy użyciu różnych strategii dodawania połączeń, opartych na właściwościach topologicznych sieci, sile pętli, kompartmentalizacji chromatyny i cechach epigenetycznych. Zaproponowałem również biofizyczny model procesu zwijania się chromatyny kontrolowanego przez skalarny parametr porządku. Parametr ten obliczany jest przy użyciu Liniowej Analizy Dyskryminacyjnej zastosowanej do epigenomicznych cech chromatyny. Do symulacji dynamiki tworzenia pętli i trajektorii zwijania się użyłem modelu LEM (Loop Extrusion Model). Mój model komputerowy wykazuje separację faz chromatyny i jej kondensację w topologiczne domeny w miarę przekraczania krytycznej liczby dodanych kontaktów. Wykazałem, że do kondensacji włókien chromatyny w przestrzeni 3D wymagana jest obecność około 80% pętli, co sprawdziłem dla wielu linii komórkowych.

W trakcie moich studiów doktoranckich badałem również sieci interakcji białko-białko (PPIN) oraz ich rolę w przetwarzaniu informacji w komórkach. Zrozumienie tych procesów jest ograniczone przez brak informacji na temat kierunkowości interakcji w sieciach PPIN. Aby rozwiązać ten problem, zaproponowałem MultiNet, metodę opartą na dyfuzji, która przypisuje kierunkowość do grafu PPIN utworzonych z ośmiu sieci obejmujących większość ludzkiego genomu. Użyta przeze mnie metoda osiągnęła najwyższy wynik AUC wynoszący 0,94 na zbiorze testowym "Protein DNA Interaction" i przewyższa obecne algorytmy wykorzystujące sieci pochodzące z pojedynczych źródeł.

Ponadto zaproponowałem nowe podejście integrujące dane genomiczne i proteomiczne dane o interakcjach. Użyłem danych ChIA-PET z ludzkiego genomu do skonstruowania sieci interakcji chromatynowych (CIN), które wykorzystywały mapowanie genów. Następnie wykonałem mapowanie ludzkich genów na MultiNet, ukierunkowaną sieć interakcji proteomic-znych, tworząc ludzką sieć interakcji biomolekularnych (hBIN). Oceniłem funkcjonalny wpływ wariantów strukturalnych na ekspresję genów i strukturę ludzkiego

genomu w wielu skalach poprzez statystyczne testowanie wpływu tych wariantów na geny bliskie sobie w przestrzeni 3D. Badałem również ich wpływ na hBIN, odzwierciedlający komórkowe interakcje między biomolekułami. Mój model hBIN obejmuje wiązanie DNA z białkami i interakcje w chromatynie, co umożliwia połączenie genomiki z kolejnymi procesami biomolekularnymi odbywa-jącymi się w komórce. Używając mojego modelu hBIN, zbadałem interakcje chromatynowe mediowane przez białko CTCF w ludzkiej limfoblastoidalnej linii komórkowej GM12878. Biorąc pod uwagę naturalny podział chromatyny na domeny genomiczne CCD potwierdziłem statystycznie, że zidentyfikowane przez moją metodę sieci mają wyższą modularność niż sieci generowane losowo. Zmapowałem również polimorfizmy pojedynczych nukleotydów (SNP) z bazy GWAS, identyfikując miejsca częstych mutacji cechujące się znacznym wzbogaceniem w mutacje SNP związane z chorobami autoimmunologicznymi. Wykonałem również analizę wariantów strukturalnych (SVs) pochodzących od zdrowych osób z bazy danych projektu 1000 Genomes Project, identyfikując brakujący kompleks białkowy związany z białkiem Q6GYQ0 z powodu delecji na chromosomie 14. Moja analiza pokazuje przydatność meta-sieciowego modelu BIN do oceny wpływu zmienności genetycznej na przestrzenną organizację genomu oraz jej funkcjonalny efekt w komórce. Moje podejście oferuje skuteczne narzędzie do modelowania systemów biologicznych w skali komórkowej i zapewnia ich bardziej kompleksowe zrozumienie.

Moje badania dotyczyły również fundamentalnego zagadnienia w bioinformatyce, jakim jest przewidywanie funkcji białek. Opracowałem nową metodę do porządkowania ontologii genów (GO terms) i przewidywania funkcji białek. Potwierdziłem wyniki uzyskane za pomocą sieci hBIN i mapowania białek i genów identyfikując dziesięć krytycznych genów, oraz określając ich trójwymiarowe położenie w genomie ludzkim. Wyniki podkreślają znaczenie metod obliczeni-owych zaprojektowanych przeze mnie we wcześniejszych etapach pracy nad doktoratem.

Podsumowując, niniejsza rozprawa doktorska dostarcza nowych algorytmów, narzędzi i modeli obliczeniowych w celu znalezienia odpowiedzi na konkretne pytania badawcze oraz ułatwia zrozumienie systemów biologicznych. Wykorzystując teorię grafów oraz inne narzędzia matematyczne oraz obliczeniowe, lepiej opisując złożone zjawiska komórkowe.

Słowa kluczowe: Genomika 3D, Pętle chromatynowe CTCF, domeny topologicznie związane (TAD, ang. Topologically associating domain), domeny kontaktowe (CCD, ang. Chromatin contact domains), Sieci, Perkolacja, Separacja faz, Zwijanie chromatyny, Parametr porządku, Wyciąganie pętli, Kompartmentalizacja, Warianty strukturalne, Ekspresja genów, Meta-sieci, Analiza sieci, Centralność, Interakcje białko-białko, Sekwencje białek, Domeny białkowe, Asocjacja genów 3D, GO terms, Przewidywanie funkcji białek.

ABSTRACT

The three-dimensional organization of the human genome and its impact on gene expression and cellular function remain a topic of ongoing research. The genome is structured into hierarchical levels, from chromosomal territories to densely-packed genomic regions known as topologically associating domains (TADs) or chromatin contact domains (CCDs) and loops. This thesis proposes novel graph-based algorithms to explore the folding of the genome in 3D and its impact on cellular function on a whole-cell scale.

To investigate how the genome folds, a novel *in-silico* model based on explosive percolation theory over graphs was developed to simulate the dynamic folding of the genome into hierarchical structures. The applied CTCF-Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) experimental data identifies CTCF-mediated interactions in the GM12878 cell line, to construct Chromatin Interaction Networks (CIN). This model is able to simulate the addition of CTCF chromatin loops as edges in CIN, following various edge addition strategies based on network properties, chromatin loop frequencies, compartmentalization, and epigenomic features. Additionally, a biophysical pseudo-time process guided by a single scalar order parameter is proposed, calculated using Linear Discriminant Analysis (LDA) over chromatin features. The incorporate Loop Extrusion Model (LEM) allows for a simulation of the dynamics of loop formation and folding. The model observes chromatin phase separation and condensation into topological domains and compartments triggered by the critical number of contacts. It is demonstrated that at least 80% loops are required for chromatin fiber to condense in 3D space, as a constant across various cell lines (H1ESC and HFFc6) and experiment types (ChiA-PET and Hi-C).

The protein-protein interaction networks (PPIN) and their role in information processing and decision-making in cells were also investigated. The directionality of PPIN is currently unknown, which limits our understanding of the flow of information in a cell, for which MultiNet, a diffusion-based method that assigns directionality to PPIN created from eight networks covering most of the human genome is proposed. The method used achieved the highest AUC score of 0.94 over the Protein DNA Interaction test set and outperformed current state-of-the-art algorithms that use networks from single databases. Using the CIN and MultiNet, a novel approach integrating genomic and proteomic interaction data was proposed. The Chromatin Interaction Networks (CIN) was mapped to genes, and the genes were mapped onto MultiNet, creating a human biomolecular interaction network (hBIN). The hBIN model used in this study includes DNA-protein binding and chromatin interactions, enabling the connection of genomics with downstream

biomolecular processes present in a cell. Using this hBIN model, the chromatin interactions mediated by the CTCF protein in the human lymphoblastoid cell line GM12878 were analysed. Considering the natural partitioning of chromatin into CCDs, it was statistically confirmed that the networks this approach identified have higher modularity than randomly generated networks. The functional impact of genomic reorganization was evaluated for gene expression and the multi-scale structures of the human genome through statistical testing of the influence of those reorganizations on spatially close genes and their impact using hBIN, reflecting the cellular interactome. The single nucleotide polymorphisms (SNPs) from GWAS studies were mapped, identifying chromatin mutational hot spots associated with significant enrichment of SNPs related to autoimmune diseases. Structural variants (SVs) from healthy individuals of the 1000 Genomes Project were also mapped, identifying the missing protein complex associated with protein Q6GYQ0 due to a deletion on chromosome 14. This analysis demonstrates the usefulness of the meta-network BIN model in evaluating the influence of genetic variation on the spatial organization of the genome and its functional effect in a cell. It offers a powerful tool for modelling biological systems and provides a comprehensive understanding of genotype-to-phenotype connections.

In addition, the study addressed the fundamental problem of predicting protein function, developing a multi-source approach to rank gene ontology (GO) terms and predict protein functions. The results were validated using hBIN networks and back-mapping from proteins to genes, identifying ten essential target genes and inferring their 3D organization in the genome.

Overall, this thesis provides innovative computational approaches to address specific research gaps in understanding biological systems, using graph theory and mathematical tools to gain insights into complex biological phenomena. These approaches offer a powerful tool for modeling biological systems and provide a comprehensive understanding of genotype-to-phenotype connections.

Keywords: 3D genomics, CTCF Chromatin loops, Topologically associating domains (TADs), Chromatin contact domains (CCDs), Networks, Percolation, Phase separation, Chromatin folding, Networks, Scalar parameter, Loop extrusion, Compartmentalisation, Structural Variants, Gene expression, Meta-networks, Network analysis, Centrality, Protein-protein interaction, Protein Sequence, Protein Domain, 3D Gene-Gene Association, Ranked GO, Protein Function Prediction.