POZNAN UNIVERSITY OF TECHNOLOGY

FACULTY OF COMPUTING AND TELECOMMUNICATIONS

Institute of Computing Science

Algorithms for feature exploration and modeling of quadruplex structures

Michał Paweł Żurkowski, M.Sc.

Supervisor: **Marta Szachniuk, Prof. Dr Eng.**

Poznań, 2024

Abstract

The subject of this dissertation is derived from bioinformatics, in which biology and computing science converge to address complex challenges of life sciences. As one of the fastest-growing branches of science, bioinformatics plays a crucial role in generating, processing, and analyzing bioscience-related data. This type of data includes quadruplex (G4) structures, the discovery and analysis of which is the focus of this work. Through the development of new computational methods and bioinformatic tools, it contributes to a better understanding of these motifs with implications for the development of bioinformatics, molecular medicine, and biotechnology. The first result of this work is *ONQUADRO*, a self-updating repository dedicated to quadruplexes. This resource retrieves data from the Protein Data Bank (PDB) and, through further analysis and processing, enables exploration of quadruplex structures, including sequences, and the secondary and tertiary structures of tetrads, G4s, and G4 helices. Complementing this database is *WebTetrado*, a web server to analyze structures that were obtained *in silico* or experimentally and have not yet been submitted to the PDB, or structures with simulated modifications. Both tools are consistent in their parametric descriptions of structures, forming a duo that provides a comprehensive analysis of quadruplexes. Another result is *DrawTetrado* that automates the creation of 2.5D layer diagrams, offering optimized visualization to facilitate understanding of G4s and the discovery of their

structural relationships. From my previous work related to improvements with visualizations of 2D structures of RNAs which led to the discovery of a new ONZ classification for quadruplexes, it was of utmost importance to provide that improvement to the most common way of visualizing quadruplexes. *DrawTetrado* is integrated into *ONQUADRO* and *WebTetrado*.

The analysis of structural features and similarities between biomolecules often requires the alignment of 3D structures. Algorithms designed for this task are also used to evaluate structure modeling. However, aligning the motifs with the unusual architecture adopted by quadruplexes presents a significant bioinformatic challenge. Therefore, as part of my doctoral thesis, I developed two proprietary algorithms for flexible alignment of nucleic acid structures, *GEOS* and *GENS*. I then created a system called *RNAhugs*, which enables these algorithms to operate independently and dependently on the sequence, with adjustable RMSD cutoff values. I validated the algorithms through extensive benchmarking against all currently available methods dedicated to 3D structure alignment. These tests confirmed that the new methods can produce longer alignments while maintaining the same or better structural similarity.

The *LinkTetrado* algorithm is the latest achievement in this dissertation. It is the world's first automatic method for identifying multimeric nucleotide assemblies based on G4 structures. It allows the discovery of nucleotides in DNA and RNA molecules that interact with tetrads so that pentads, hexads, heptads, and beyond are formed. The application of *LinkTetrado* expands the catalog of known motifs and allows the analysis of their properties. The algorithm was validated against experimental data from Nuclear Magnetic Resonance (NMR). In analyzing the NMR data, I collaborated with researchers from the Department of Biomolecular NMR, IBCH PAS, and Dr. Maja Marušič from the Slovenian NMR Center.

Streszczenie

Tematyka rozprawy wywodzi się z bioinformatyki, w której biologia i nauki komputerowe łączą się, aby stawić czoła wyzwaniom nauk przyrodniczych. Bioinformatyka, jedna z najszybciej rozwijających się gałęzi nauki, odgrywa kluczową rolę w generowaniu, przetwarzaniu i analizowaniu danych z bionauk. Do tego rodzaju danych należą struktury kwadrupleksów (G4), na których odkrywaniu i analizie skupia się niniejsza praca doktorska. Poprzez opracowanie nowych metod obliczeniowych i narzędzi bioinformatycznych, przyczynia się ona do lepszego zrozumienia tych motywów mając wpływ na rozwój bioinformatyki, medycyny molekularnej i biotechnologii.

Pierwszym wynikiem niniejszej pracy jest *ONQUADRO*, samoaktualizujące się repozytorium dedykowane kwadrupleksom. Pozyskuje ono dane z Protein Data Bank (PDB), a dzięki dalszej analizie i przetwarzaniu umożliwia eksplorację ich struktur, w tym sekwencji oraz struktur drugo- i trzeciorzędowych, tetrad, G4 oraz helis G4. Uzupełnieniem tej bazy danych jest *WebTetrado*, aplikacja do analizy struktur, które zostały otrzymane *in silico* lub eksperymentalnie i nie przesłano ich jeszcze do PDB, lub struktur z symulowanymi modyfikacjami. Oba narzędzia są spójne jeśli chodzi o parametryczny opis struktur tworząc duet zapewniający kompleksową analizę kwadrupleksów. Kolejnym rezultatem jest narzędzie *DrawTetrado*, które automatyzuje tworzenie diagramów warstwowych w grafice 2.5D, oferując wizualizację ułatwiającą zrozumienie G4 i odkrywanie ich relacji

strukturalnych. Na podstawie moich wcześniejszych prac związanych z poprawą wizualizacji struktur 2D RNA, które doprowadziły do odkrycia nowej klasyfikacji kwadrupleksów ONZ, niezwykle istotne było wprowadzenie tej poprawy do najbardziej powszechnego sposobu przedstawiania kwadrupleksów. *DrawTetrado* jest zintegrowane z *ONQUADRO* i *WebTetrado*. Analiza cech strukturalnych i podobieństw między biocząsteczkami często wymaga dopasowania ich struktur 3D. Algorytmy zaprojektowane do tego zadania są też wykorzystywane do oceny modelowania struktur. Jednak dopasowanie motywów o nietypowej architekturze, jaką przyjmują kwadrupleksy, stanowi istotne wyzwanie bioinformatyczne. W związku z tym, w ramach mojej pracy doktorskiej, opracowałem dwa autorskie algorytmy do elastycznego dopasowania struktur kwasów nukleinowych, *GEOS* i *GENS*. Następnie stworzyłem system o nazwie *RNAhugs*, który umożliwia uruchomienie tych algorytmów zarówno zależnie, jak i niezależnie od sekwencji, z możliwością dostosowania wartości odcięcia RMSD. Algorytmy te zostały zweryfikowane przez szeroko zakrojone testy porównawcze z wszystkimi dostępnymi obecnie metodami dedykowanymi dopasowaniu struktur 3D. Testy potwierdziły, że *GEOS* i *GENS* mogą zapewnić dłuższe dopasowania, przy zachowaniu tego samego lub większego podobieństwa strukturalnego. Algorytm *LinkTetrado* to ostatnie z osiągnięć niniejszej pracy. Jest to pierwsza w świecie automatyczna metoda do detekcji motywów multimerycznych opartych na strukturach G4. Pozwala ona na odkrycie w cząsteczkach DNA i RNA nukleotydów wchodzących w interakcje z tetradami w taki sposób, że tworzą się pentady, heksady, heptady, itd. Zastosowanie *LinkTetrado* pozwala rozszerzyć katalog znanych motywów i umożliwia analizę ich właściwości. Algorytm został zweryfikowany w kontekście danych eksperymentalnych z Magnetycznego Rezonansu Jądrowego (MRJ). Przy analizie tych danych współpracowałem z badaczami z Zakładu Biomolekularnego NMR, ICHB PAN oraz z dr Maja Marušič ze Słoweńskiego Centrum NMR.

Contents