Algorytmiczne aspekty analizy podobieństwa struktur RNA

Tomasz Żok

STRESZCZENIE ROZPRAWY DOKTORSKIEJ

Promotor: **dr hab. inż. Marta Szachniuk**

Promotor pomocniczy: **dr Mariusz Popenda**

Politechnika Poznańska

Instytut Informatyki

Poznań, 2018

Wstęp

Bioinformatyka to interdyscyplinarna dziedzina nauki, w której problemy biologii molekularnej rozwiązywane są metodami zaczerpniętymi z informatyki i badań operacyjnych. Szczególnie istotne znaczenie należy przypisać w tej mierze bioinformatyce strukturalnej, której metody pozwalają zrozumieć wpływ struktur cząsteczek na ich funkcje *in vivo*. Zagadnienie to ma coraz większe znaczenie dla badań nad cząsteczkami RNA, pełniącymi zasadnicze funkcje w komórkach, a jednocześnie zwijającymi się w struktury o dużej złożoności.

W ramach badań opisanych w rozprawie doktorskiej opracowano nowe modele i metody oceny podobieństwa struktury RNA. Kwestia ta stanowi istotne zagadnienie naukowe, gdyż ocena podobieństwa strukturalnego na podstawie metod porównawczych jest wykorzystywana na różnych etapach przetwarzania struktur RNA, tj. podczas przewidywania, modelowania, klastrowania i klasyfikacji. Jednocześnie jest to trudne zadanie ze względu na złożoność struktur RNA i ilość danych do przetwarzania.

Nowe metody analizy podobieństwa struktur 2D RNA

Struktura 2D RNA zawiera informację o interakcjach przestrzennych między nukleotydami wchodzącymi w skład cząsteczki. Główne interakcje poddawane analizie to kanoniczne parowania między nukleotydami G-C, A-U lub G-U. Analiza takich par w obrębie całej struktury 2D pozwala wydzielić motywy, czyli powtarzające się elementy strukturalne o określonych cechach. Przykładowo, motyw *stem* to ciąg kolejnych parowań kanonicznych o określonej długości. Informacje o motywach strukturalnych są gromadzone w bazach danych i dostępne z poziomu wyspecjalizowanych wyszukiwarek, ponieważ ich zrozumienie pozwala badać cząsteczki RNA w szerszym kontekście jako struktury składające się z powtarzalnych bloków.

Powyższe ujęcie jest stosowane w bazie danych i wyszukiwarce RNA FRABASE. Narzędzie do predykcji struktury 3D RNA o nazwie RNAComposer rozwija tę koncepcję, ponieważ tworzy model z fragmentów opisanych motywami 2D. Oba narzędzia są dobrze znane w społeczności naukowej. Cechę wspólną RNA FRABASE i RNAComposera stanowi wykorzystywanie tekstowego formatu *dot-bracket* do opisu struktur 2D, jak i poszukiwanych motywów. Format ten jest kompaktowy i czytelny, a motywy strukturalne można wyszukać jako podciągi tekstowe. Należy jednak uwzględnić fakt, że złożone motywy składają się z kilku łańcuchów, które w metodzie podciągów tekstowych trzeba szukać w zmieniającej się cyklicznie kolejności.

Dodatkową komplikacją są tzw. pseudowęzły występujące w części struktur RNA.

W pracy badawczej stworzono nowy model mieszanego grafu RNA, którym można reprezentować dowolną strukturę 2D RNA albo motyw strukturalny. Na tej podstawie zaproponowano: (1) algorytm PMMG do dokładnego znajdowania fragmentów struktury RNA oraz (2) algorytm IMMG do dopasowania niedokładnego z oceną podobieństwa pozwalającą na ranking wyników. W eksperymencie obliczeniowym pokazano, że część motywów generowanych losowo nie ma dokładnego dopasowania w żadnej znanej strukturze RNA. W takim przypadku RNAComposer generuje fragment, który jest poprawny, ale nie zawsze optymalny. Algorytm IMMG odnajduje natomiast fragmenty, które zachowują wymaganą topologię strukturalną, mimo niedokładnego dopasowania. Kolejny eksperyment wykazał, że odnalezione przez algorytm IMMG fragmenty pozytywnie wpływają na wynik przewidywania struktury przestrzennej przez RNAComposer. Wyniki te dostępne są w tabeli poniżej.

Porównanie modeli RNAComposera z fragmentem generowanym (G5 lub G4), znalezionym (F5 lub F4) albo znalezionym i ręcznie modyfikowanym (F5m). RMSD, DI, MCQ, MedCQ i p-value to miary odległości (im mniejsze, tym bardziej podobne). Miara INF to miara podobieństwa (im większa, tym bardziej podobne). Czerwonym tłem zaznaczono najgorszy wynik na danym kryterium, natomiat na zielonym tle umieszczono wynik najlepszy.

	RMSD [Å]	INF	DI	MCQ [°]	MedCQ [°]	p-value
G5/G4	49.40	0.817	60.49	44.37	32.25	$9.85E-01$
G5/F4	48.11	0.827	58.20	41.98	30.00	$9.28E-01$
F5/G4	36.57	0.830	44.06	46.00	34.71	$3.80E-07$
F5/F4	35.95	0.887	39.71	39.71	26.41	$6.03E-08$
FM5/G4	37.58	0.853	44.08	41.45	28.11	$5.77E-06$
FM5/F4	34.91	0.836	43.83	43.83	31.70	$2.15E-09$

Nowe metody analizy podobieństwa struktur 3D RNA

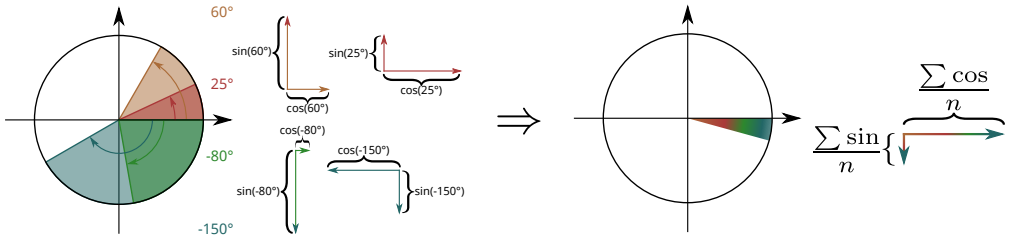
Struktura 3D RNA opisuje przestrzenny kształt cząsteczki. Najczęściej jest on reprezentowany przez współrzędne 3D wszystkich atomów tworzących daną strukturę. W pracy doktorskiej skupiono się natomiast na alternatywnej reprezentacji trygonometrycznej, w której kształt cząsteczki opisują wartości kątów torsyjnych, a zatem obroty wokół wiązań między kolejnymi atomami.

Na podstawie tej koncepcji zaproponowano metrykę odległości określającą poprawnie różnicę między wartościami kątów torsyjnych. Następnie zaadaptowano wzór na średnią wartości okresowych (ang. *mean of circular quantities, MCQ*) do wyznaczenia średniej różnicy kątowej jako globalnej miary odległości dwóch struktur RNA. W pracy doktorskiej wykazano, że podejście to jest skalowalne i umożliwia porównywanie zarówno fragmentów, jak i całych struktur, a nadto umożliwia dobór typów kątów torsyjnych, także niestandardowych. Uwzględniając różne zapotrzebowania metod do analizy struktur RNA, miara MCQ pozwala na porównywanie w trybie: jeden do jednego, jeden do wielu, wiele do wielu. Wynik porównania może służyć do klastrowania w celu automatycznego odkrycia zależności między danymi. Rozprawa doktorska zawiera również liczne wizualizacje pomocne dla określenia źródeł podobieństwa lub jego braku.

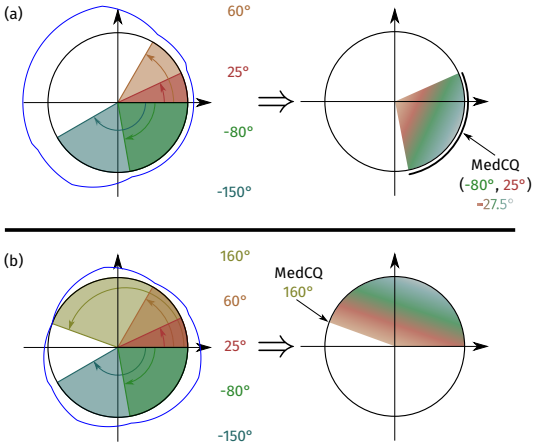
MCQ, jak każda średnia, jest miarą centralności, w tym przypadku wartości różnic kątowych. Dlatego jest podatna na wpływ wartości ekstremalnych

w próbie. W pracy badawczej sprawdzono zatem cechy mediany jako alternatywy dla średniej. W przypadku danych kątowych określenie wartości mediany wymaga jednak optymalizacji funkcji celu, by znaleźć globalne minimum. W przypadku zbioru różnic kątów torsyjnych między parą struktur RNA wartość funkcji celu została nazwana po angielsku *median of circular quantities (MedCQ)*.

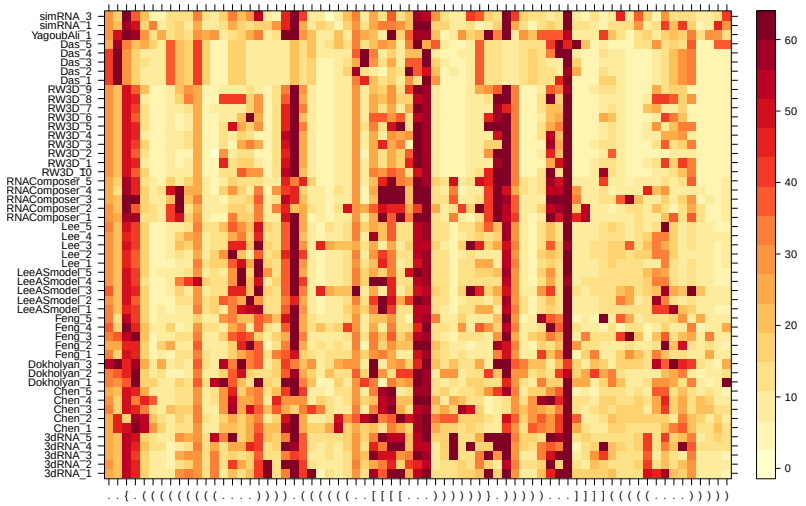
W pracy doktorskiej zweryfikowano działanie miar kątowych na kilkudziesięciu modelach 3D RNA zgłoszonych w ramach Puzzle 18, edycji konkursu RNA-Puzzles. Uczestnicy konkursu nadsyłają co najwyżej kilka modeli struktury 3D RNA, powstałych *in silico*, wyłącznie w oparciu o podstawowe dane udostępnione przez organizatora. Z dniem opublikowania struktury natywnej, modele są porównywane, a ich ranking umieszczany na stronie internetowej poświęconej konkursowi. Jedną z metryk podobieństwa, używaną przez organizatorów RNA-Puzzles od 2014 r., jest właśnie MCQ w wersji podstawowej. W pracy doktorskiej stwierdza się, że inne sposoby użycia MCQ ujawniają znacząco więcej szczegółów odnośnie do podobieństwa modeli struktur 3D RNA. Miara MCQ klarownie pokazała, że różnice w obrębie modeli jednego zespołu są znacząco mniejsze niż względem reszty zgłoszeń. Uwzględniając strukturę natywną, miara ta pozwoliła także wskazać fragmenty najtrudniejsze do przewidzenia. Tym samym MCQ daje informację zwrotną o możliwościach potencjalnego rozwoju metod do modelowania struktury 3D RNA.



Wizualna prezentacja miary MCQ. Każdy kąt dekomponowany jest na składową poziomą i pionową. Składowe uśredniane są niezależnie. Wartość MCQ równa jest kątowi między wynikowymi wektorami.



Wizualna prezentacja miary MedCQ dla (a) parzystej i (b) nieparzystej liczby danych wejściowych. Niebieska krzywa to funkcja, której minimum wyznacza medianę. W przypadku (a) wynikiem jest zakres liczb, natomiast w (b) wynik równy jest jednej z danych wejściowych.



Mapa ciepła na podstawie MCQ. Jasny kolor oznacza podobieństwo, ciemny natomiast jego brak. Wiersz opisuje pojedynczy model, a kolumna nukleotyd. Czerwone kolumny oznaczają błąd we wszystkich modelach.

Podsumowanie

Tematem pracy doktorskiej jest analiza podobieństwa struktur RNA, stanowiąca kluczowy element wielu metod bioinformatycznych. Na poprawę ich działania może zatem znacząco wpłynąć rozwój algorytmów do porównywania wspomnianych struktur. Opracowany model grafowy oraz algorytmy PMMG i IMMG pozwalają wyszukiwać fragmenty struktur 2D RNA, spełniające cechy zadanego motywu strukturalnego. W pracy badawczej wykazano, że przy braku dokładnego dopasowania algorytm IMMG znajduje wynik najbardziej podobny. W eksperymencie obliczeniowym zweryfikowano pozytywny wpływ wynikowych fragmentów na jakość modeli struktury RNA złożonej przy ich użyciu.

Zaproponowana metryka kątowa i wynikające z niej średnia MCQ oraz mediana MedCQ umożliwiają precyzyjną ocenę podobieństwa struktur RNA. W rozprawie doktorskiej stwierdza się, że obie metody można wykorzystywać w różnych scenariuszach, a wyniki są miarodajne i pozwalają zrozumieć źródła podobieństwa lub jego braku. Miara MCQ jest używana do oceny zgłoszeń w konkursie RNA-Puzzles, w którym zespoły badawcze konkurują, by zaproponować najlepszy model nieznanej struktury.

W przyszłości planowana jest większa integracja zaproponowanych rozwiązań z istniejącymi narzędziami oraz ich rozwój na podstawie wymagań użytkowników. W ramach trygonometrycznej reprezentacji struktury 3D dodana zostanie możliwość swobodniejszego definiowania niestandardowych kątów torsyjnych i pseudotorsyjnych.

Najważniejsze publikacje

1. Maciej Antczak, Mariusz Popenda, **Tomasz Zok**, Michal Zurkowski, Ryszard W. Adamiak, and Marta Szachniuk. New algorithms to represent complex pseudoknotted RNA structures in dot-bracket notation. *Bioinformatics*, page btx783, 2017.

2. Jakub Wiedemann, **Tomasz Zok**, Maciej Milostan, and Marta Szachniuk. LCS-TA to identify similar fragments in RNA 3D structures. *BMC Bioinformatics*, 18(1):456, 2017.

3. Zhichao Miao, Ryszard W. Adamiak, Maciej Antczak, Robert T. Batey, Alexander J. Becka, Marcin Biesiada, Michał J. Boniecki, Janusz M. Bujnicki, Shi-Jie Chen, Clarence Yu Cheng, Fang-Chieh Chou, Adrian R. Ferré-D'Amaré, Rhiju Das, Wayne K. Dawson, Feng Ding, Nikolay V. Dokholyan, Stanisław Dunin-Horkawicz, Caleb Geniesse, Kalli Kappel, Wipapat Kladwang, Andrey Krokhotin, Grzegorz E. Łach, François Major, Thomas H. Mann, Marcin Magnus, Katarzyna Pachulska-Wieczorek, Dinshaw J. Patel, Joseph A. Piccirilli, Mariusz Popenda, Katarzyna J. Purzycka, Aiming Ren, Greggory M. Rice, John Santalucia, Joanna Sarzynska, Marta Szachniuk, Arpit Tandon, Jeremiah J. Trausch, Siqi Tian, Jian Wang, Kevin M. Weeks, Benfeard Williams, Yi Xiao, Xiaojun Xu, Dong Zhang, **Tomasz Zok**, and Eric Westhof. RNA-Puzzles Round III: 3D RNA structure prediction of five riboswitches and one ribozyme. *RNA*, 23(5):655–672, 2017.

4. Maciej Antczak, Mariusz Popenda, **Tomasz Zok**, Joanna Sarzynska, Tomasz Ratajczak, Katarzyna Tomczyk, Ryszard W Adamiak, and Marta Szachniuk.

New functionality of RNAComposer: an application to shape the axis of miR160 precursor structure. *Acta Biochimica Polonica*, 63(4):737–744, 2016.

5. Zhichao Miao, Ryszard W Adamiak, Marc-frédérick Blanchet, Michal Boniecki, Janusz M. Bujnicki, Shi-jie Chen, Clarence Cheng, Grzegorz Chojnowski, Fang-chieh Chou, Pablo Cordero, José Almeida Cruz, Adrian R. Ferré-D'Amaré, Rhiju Das, Feng Ding, Nikolay V. Dokholyan, Stanislaw Dunin-Horkawicz, Wipapat Kladwang, Andrey Krokhotin, Grzegorz Lach, Marcin Magnus, François Major, Thomas H Mann, Benoît Masquida, Dorota Matelska, Mélanie Meyer, Alla Peselis, Mariusz Popenda, Katarzyna J Purzycka, Alexander Serganov, Juliusz Stasiewicz, Marta Szachniuk, Arpit Tandon, Siqi Tian, Jian Wang, Yi Xiao, Xiaojun Xu, Jinwei Zhang, Peinan Zhao, **Tomasz Zok**, and Eric Westhof. RNA-Puzzles Round II: assessment of RNA structure prediction programs applied to three large RNA structures. *RNA*, 21(6):1066–1084, 2015.

6. **Tomasz Zok**, Maciej Antczak, Martin Riedel, David Nebel, Thomas Villmann, Piotr Lukasiak, Jacek Blazewicz, and Marta Szachniuk. Building the library of RNA 3D nucleotide conformations using clustering approach. *International Journal of Applied Mathematics and Computer Science*, 25(3):689–700, 2015.

7. Agnieszka Rybarczyk, Natalia Szostak, Maciej Antczak, **Tomasz Zok**, Mariusz Popenda, Ryszard W. Adamiak, Jacek Blazewicz, and Marta Szachniuk. New in silico approach to assess RNA secondary structures with non-canonical base pairs. *BMC Bioinformatics*, 16(1):276, 2015.

8. Maciej Antczak, **Tomasz Zok**, Mariusz Popenda, Piotr Lukasiak, Ryszard W. Adamiak, Jacek Blazewicz, and Marta Szachniuk. RNApdbee–a webserver to derive secondary structures from pdb files of knotted and unknotted RNAs. *Nucleic Acids Research*, 42(Web Server issue):W368–72, 2014.

9. **Tomasz Zok**, Mariusz Popenda, and Marta Szachniuk. MCQ4Structures to compute similarity of molecule structures. *Central European Journal of Operations Research*, 22(3):457–473, 2014.