Joanna Miśkiewicz

Bioinformatics methods of motif analysis in RNA structures

Streszczenie rozprawy doktorskiej

Promotor: Prof. dr hab. inż. Marta Szachniuk

Poznań 2022

W literaturze, sztuce, czy muzyce spotykamy się z pewnymi powtarzalnymi schematami, po których rozpoznajemy ich twórców lub epokę, z której dzieła pochodzą. Te powtarzające się wzory nazywane są motywami i występują również w naukach o życiu – w sieciach metabolicznych, procesach regulacyjnych komórki, czy też w strukturach kwasów nukleinowych. Każdy biologiczny motyw ma nie tylko określoną formę, ale też specyficzną rolę do odegrania w organizmie. Odnalezienie wzorców (czyli powtarzalnych fragmentów określonych jako kombinację elementów strukturalnych) nie jest tożsame z odkryciem motywu. Do pełnej definicji potrzebujemy połączenia odkrytych wzorców do ich funkcjonalności. Odnajdując dany motyw w cząsteczce naukowcy są w stanie powiązać z nim funkcję jaką pełni w systemie. Trudniejszym zadaniem jest jednak poszukiwanie motywu, który odpowiada za konkretne działania cząsteczki.

Motywy w kwasach nukleinowych (DNA i RNA) możemy poszukiwać na każdym z trzech poziomów strukturalnych – sekwencji, strukturze drugorzędowej, oraz strukturze trzeciorzędowej. Cząsteczka zdefiniowana na każdym z tych poziomów może być przedstawiona w dwóch formach – tekstowej bądź graficznej. Na poziomie sekwencji poszukiwane są powtarzalne słowa, najczęściej nad czteroliterowym alfabetem – {A,U,G,C} dla RNA, oraz ze zmianą U na T dla DNA. Inną metodą zapisu sekwencji jest jej kodowanie według konwencji IUPAC [Johnson 2010, Hendrix et al., 2005]. Alfabet IUPAC pozwala na reprezentację dwóch lub więcej zasad jako pojedynczego symbolu – przykładem może być symbol R oznaczający purynę (guanina lub adenina) oraz Y symbolizujący pirymidynę (cytozyna, tymina, uracyl). Odkrywanie nowych motywów sekwencyjnych, nawet przy użyciu kodowania IUPAC lub wyrażenia regularnego, jest problemem nietrywialnym, uważanym za problem NP-trudny, a co za tym idzie stanowiącym wyzwanie dla bioinformatyków i biologów obliczeniowych [Li et al., 2010, Rajasekaran et al., 2011, Rampasek et al., 2016, Ashraf et al., 2020].

Kolejny poziom strukturalny, struktura drugorzędowa, odzwierciedla układ fragmentów sparowanych i niesparowanych. W RNA zazwyczaj jest to kombinacja podstawowych elementów strukturalnych, takich jak dupleks, pętla symetryczna, multipętla, czy pojedyncze niedopasowanie [Batey et al., 1999, Chheda et al., 2014]. Poprawnie zdeterminowana struktura 2D cząsteczek pozwala na trafniejsze wyszukiwanie motywów w ich architekturze.

Struktura 3D cząsteczki RNA stanowi jej biologicznie aktywną formę [Batey et al., 1999, Hoehndorf et al., 2011]. Zwinięta struktura przedstawia pozycje wszystkich atomów molekularnych oraz trzeciorzędowe oddziaływania wiążące motywy strukturalne [Halder et al., 2013, Picardi 2015, Miao 2017]. Zdeterminowanie struktury 3D cząsteczki kwasu nukleinowego zazwyczaj odbywa się przy użyciu metod spektroskopii NMR lub krystalografii rentgenowskiej, jednak w przypadku ograniczeń czasowych i finansowych, naukowcy mogą w tej kwestii zwrócić się w kierunku bioinformatyki. Algorytmy do przewidywania struktur 3D stają się coraz dokładniejsze, w szczególności dzięki zastosowaniu metod głębokiego nauczania maszynowego [Huang et al., 2020, A. Kryshtafovych et al., 2021, Jumper et al., 2021, Townshend et al., 2021]. Wykorzystanie metod bioinformatycznych pozwala w takim przypadku na poznanie niemal dokładnej struktury trzeciorzędowej przy niskich kosztach czasowych i finansowych.

Wyszukiwanie motywów w cząsteczkach kwasów nukleinowych jest zadaniem wymagającym wiedzy z różnych dziedzin, statystyki, biologii, informatyki. Różnorodność wzorców oraz format zapisu danych wejściowych daje szerokie pole do badań, w szczególności dla bioinformatyków.

Niniejsza praca doktorska poświęcona jest badaniom motywów strukturalnych w cząsteczkach RNA pochodzących z różnych organizmów. Prace wykonane podczas doktoratu polegały na wyszukiwaniu i analizie motywów w sekwencjach oraz strukturach drugo- i trzeciorzędowych. Pierwsze badania skupione były na poszukiwaniu motywów strukturalnych w zbiorze roślinnych mikroRNA na przykładzie organizmu modelowego – *Arabidopsis thaliana*. Zaobserwowano schemat powtarzania się małych pętli wewnętrznych w okolicach dupleksu miRNA:miRNA*, co może wskazywać na obecność motywu rozpoznawalnego przez enzym wycinający dupleks z cząsteczki. Uzyskane wyniki były inspiracją do rozszerzenia badań na pre-miRNA z całego królestwa roślin zielonych – *Viridiplantae*. W analizowanych strukturach wykryto podobny motyw jak przy analizach pre-miRNA w *Arabidopsis thaliana*. Kolejne badania dotyczyły struktury pierwotnego transkryptu miR-125a w dwóch wariantach sekwencyjnych (zmiana pojedynczego nukleotydu, SNP). Bioinformatyczna analiza wskazywała na zależność rodzaju wiązanych białek do transkryptu od wybranego typu wariantu sekwencyjnego. Ponadto, predykcja struktury drugorzędowej wskazywała na różnice strukturalne wynikające ze zmiany pojedynczego nukleotydu w transkrypcie. Najnowsze badania koncentrowały się na motywach kwadrupleksów, ich topologii oraz analizie parametrycznej z użyciem narzędzi bioinformatycznych. Zaowocowały one opracowaniem nowej klasyfikacji kwadrupleksów w oparciu o ich strukturę drugorzędową oraz stworzeniem nowych reprezentacji umożliwiających zapisywanie informacji o strukturze drugorzędowej w dwuliniowej notacji kropkowo-nawiasowej i w postaci dwuczęściowego diagramu łukowego. Przebadaliśmy wszystkie dostępne zasoby bioinformatyczne pod kątem ich wykorzystania do badań kwadrupleksów RNA oraz utworzyliśmy bazę danych ONQUADRO gromadzącą i przetwarzającą dane o strukturach kwadrupleksów otrzymanych drogą eksperymentalną. Przeanalizowaliśmy ludzkie sekwencje mikroRNA pod kątem ich potencjału do formowania motywów kwadrupleksów. W tym celu wykorzystaliśmy algorytm bazujący na dopasowaniu wyrażeń regularnych. Sekwencje zostały również zbadane pod kątem nasycenia guaninami, w celu sprawdzenia wielkości zbioru, który spełnia minimalny wymóg do posiadania motywu kwadrupleksu (8G i 12G kolejno dla dwu- i trójtetradowych kwadrupleksów).

W badaniach do pracy doktorskiej wykorzystywane były dostępne narzędzia bioinformatyczne, jak również nowo stworzone metody do analizy zbiorów danych strukturalnych. Wszystkie analizowane dane pochodzą z publicznie dostępnych repozytoriów.

Bibliografia

Johnson, A. D. (2010). An extended IUPAC nomenclature code for polymorphic nucleic acids. *Bioinformatics*, **26**(10), 1386–1389.

Hendrix, D. K., Brenner, S. E., & Holbrook, S. R. (2005). RNA structural motifs: building blocks of a modular biomolecule. *Quarterly Reviews of Biophysics*, **38**(3), 221–243.

Li, G., Chan, T.-M., Leung, K.-S., & Lee, K.-H. (2010). A cluster refinement algorithm for motif discovery. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **7**(4), 654–668.

Rajasekaran, S. & Dinh, H. (2011). A speedup technique for (l, d)-motif finding algorithms. *BMC Research Notes*, **4**(1), 54.

Rampášek, L., Jimenez, R. M., Lupták, A., Vinar, T., & Brejová, B. (2016). RNA motif search with data-driven element ordering. *BMC Bioinformatics*, **17**(1), 216.

Ashraf, F. B. & Shafi, M. S. R. (2020). MFEA: An evolutionary approach for motif finding in DNA sequences. *Informatics in Medicine Unlocked*, **21**, 100466.

Batey, R. T., Rambo, R. P., & Doudna, J. A. (1999). Tertiary motifs in RNA structure and folding. *Angewandte Chemie International Edition*, **38**(16), 2326–2343.

Chheda, N. & Gupta, M. K. (2014). RNA as a permutation. *arXiv: Biomolecules*.

Hoehndorf, R., Batchelor, C., Bittner, T., Dumontier, M., Eilbeck, K., Knight, R., Mungall, C. J., Richardson, J. S., Stombaugh, J., Westhof, E., & et al. (2011). The RNA Ontology (RNAO): An ontology for integrating RNA sequence and structure data. *Applied Ontology*, **6**(1), 53–89.

Halder, S. & Bhattacharyya, D. (2013). RNA structure and dynamics: A base pairing perspective. *Progress in Biophysics and Molecular Biology*, **113**(2), 264–283.

Picardi, E. (2015). RNA Bioinformatics. *Springer New York*.

Miao, Z. & Westhof, E. (2017). RNA Structure: Advances and Assessment of 3D Structure Prediction. *Annual Review of Biophysics*, **46**(1), 483–503.

Huang, B., Du, Y., Zhang, S., Li, W., Wang, J. & Zhang, J. (2020). Computational prediction of RNA tertiary structures using machine learning methods. *Chinese Physics B* **29**(10), 108704.

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. (2021). Critical Assessment of methods of protein structure prediction (CASP)Round XIV. *Proteins: Structure, Function, and Bioinformatics* **89**(12), 1607–1617.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zdek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* **596**(7873), 583–589.

Townshend, R. J. L., Eismann, S., Watkins, A. M., Rangan, R., Karelina, M., Das, R. & Dror, R. O. (2021). Geometric deep learning of RNA structure. *Science* **373**(6558), 1047–1051.