Przewidywanie struktur 3D RNA z wykorzystaniem generatywnych sieci przestawnych

Marta Szachniuk

Instytut Informatyki, Politechnika Poznańska

Tematyka projektu osadzona jest w bioinformatyce, dziedzinie będącej fuzją informatyki i nauk biologicznych. Celem badań jest stworzenie systemu przewidującego na podstawie sekwencji, wiarygodne trójwymiarowe struktury cząsteczek RNA o wysokiej precyzji i rozdzielczości. Cel ten zamierzamy osiągnąć łącząc elementy danetyki (ang. *data science*) i sztucznej inteligencji z koncepcją budowania cząsteczki *in silico* poprzez rekombinację elementów strukturalnych.

Komputerowe przewidywanie struktur 3D, do jakich zwijają się cząsteczki o określonej sekwencji staje się coraz częściej stosowanym podejściem uzyskiwania ogólnej wiedzy o budowie białek czy RNA. Jest to wynikiem szybkiego wzrostu rozmiaru danych przechowywanych w bazach sekwencji aminokwasowych i nukleotydowych, którym najczęściej nie towarzyszy informacja o przestrzennej strukturze molekularnej. Dysponując sekwencją możemy znaleźć cząsteczki podobne strukturalnie i na podstawie homologii wysnuć wnioski o funkcjach badanej cząsteczki oraz zarysować jej uogólniony kształt (np. przewidując go algorytmem komputerowym). Nie wystarcza to jednak, aby odpowiedzieć na pytania wymagające precyzyjnej wiedzy o strukturze 3D. Na przykład podczas identyfikacji miejsc wiązania cząsteczek z innymi – kluczowego zagadnienia zgłębianego przy projektowaniu leków ukierunkowanych molekularnie. Dlatego w wielu laboratoriach na świecie wykonuje się eksperymenty – m.in. spektroskopii rentgenowskiej, magnetycznego rezonansu jądrowego, mikroskopii krioelektronowej – pozwalające na obrazowanie molekuł i określanie ich struktur z dokładnością atomową. Dane eksperymentalne pozwalają ulepszać metody predykcji oraz udokładniać generowane przez nie modele.

Pierwsze algorytmy do przewidywania struktur 3D pojawiły się w latach 1970-tych i były dedykowane białkom. Ich intensywny rozwój zawdzięczamy m.in. uruchomieniu w 1994 roku inicjatywy CASP (Critical Assessment of protein Structure Prediction) – ogólnoświatowych mistrzostw w predykcji struktur białkowych. W niedawno zakończonym CASP14, spektakularne zwycięstwo we wszystkich kategoriach turnieju odniósł algorytm sztucznej inteligencji AlphaFold 2, a organizatorzy CASP obwieścili początek nowej ery w predykcji białek (Callaway 2020). Podobne osiągniecie w przypadku RNA jest kwestią czasu, wymaga jednak dostępu do wiarygodnych danych strukturalnych, na których można wytrenować algorytmy AI. Danych tych nie ma jeszcze wiele: Protein Data Bank zgromadził struktury 1 516 cząsteczek RNA określonych eksperymentalnie w porównaniu z 150 694 strukturami białek (dane z początku grudnia 2020). Jest to jeden z powodów, dla których postęp w przewidywaniu struktur 3D RNA nie dotrzymuje kroku postępowi w predykcji białek.

W niniejszym projekcie planujemy opracować pierwszy w świecie algorytm przewidywania struktur 3D RNA bazujący na sztucznej inteligencji. Algorytm będzie budował cząsteczkę poprzez rekombinację trójwymiarowych elementów strukturalnych, które zostaną wymodelowane przez moduł wykorzystujący generatywne sieci przestawne. Sieci zostaną wytrenowane na zbiorze niewielkich motywów 3D pochodzących zarówno ze struktur eksperymentalnych jak i z symulatorów *de novo*. Metody predykcji *de novo*, choć - z powodu złożoności obliczeń - nie nadają się do modelowania dużych struktur, potrafią generować dokładne i wiarygodne modele niewielkich cząsteczek lub ich fragmentów. Dobór elementów strukturalnych będzie oparty o algorytm wyszukiwania konsensusu (RNAtive). Ogólny zwój będzie przewidywany z użyciem istniejących metod predykcji, m.in. opracowanego w naszym laboratorium systemu RNAComposer (Popenda *et al*. 2012) oraz z wykorzystaniem wiedzy o geometrii motywów strukturalnych, które decydują o przebiegu łańcucha RNA, m.in. multipętli. Wiedzę tę pozyskamy ze struktur zgromadzonych w bazach danych, w tym opracowanych przez nas RNA FRABASE (Popenda *et al.* 2008) i RNAloops (Wiedemann *et al*. 2020), stosując metody analizy i inżynierii danych.

Stworzona przez nas metoda przyczyni się do generowania struktur 3D RNA, których precyzja i rozdzielczość będą odpowiadały tym, jakie uzyskujemy laboratoryjnymi technikami eksperymentalnymi. Umożliwi ona stosunkowo szybkie otrzymywanie informacji, których pozyskanie wymaga obecnie czasochłonnych eksperymentów. Przyczyni się to do intensyfikacji badań nad relacją między strukturą a funkcją biomolekuł oraz przyspieszy postęp w biomedycynie i bioinżynierii.

Callaway E (2020) ‘It will change everything’: DeepMind’s AI makes gigantic leap in solving protein structures, *Nature*.

Popenda M, Blazewicz M, Szachniuk M, Adamiak RW (2008) RNA FRABASE version 1.0: an engine with a database to search for the three-dimensional fragments within RNA structures, *Nucleic Acids Res* 36(1):D386-D391.

Popenda M, Szachniuk M, Antczak M, Purzycka KJ, Lukasiak P, Bartol N, Blazewicz J, Adamiak RW (2012) Automated 3D structure composition for large RNAs, *Nucleic Acids Res* 40:e112.

Wiedemann J, Kaczor J, Antczak M, Milostan M, Zok T, Szachniuk M (2020) RNAloops - a database of RNA multiloops, *zgłoszone do czasopisma*.