„Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”
projekt finansowany ze środków Programu Operacyjnego Polska Cyfrowa POPC.03.02.00-00-0001/20

Sztuczna inteligencja w informatyce biomedycznej

Projekt 06: Przewidywanie struktury drugorzędowej RNA

Maciej Antczak

9 i 16 maja 2024

Spis treści

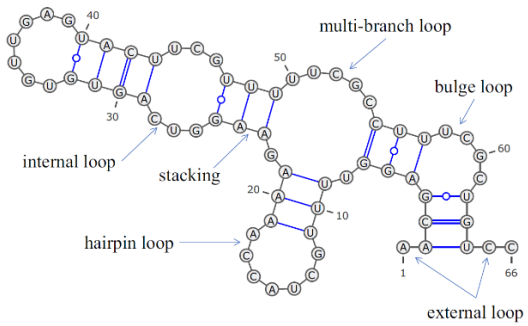
Cele szóstego cyklu zajęć laboratoryjnych są następujące:

1. Zapoznanie się z udostępnionymi zbiorami danych obejmującymi sekwencje (w formacie *FASTA*) i struktury drugorzędowe (w formacie *BPSEQ*) wyekstrahowane zarówno z (1) nieredundantnych, wysokorozdzielczych (rozdzielczość <= 3.0Å) eksperymentalnie określonych struktur 3D RNA [1] oraz (2) 10 rodzin cząsteczek RNA [2], których struktury drugorzędowe zostały manualnie udokładnione.
2. Opracowanie, wytrenowanie, walidacja i ewaluacja podejść, wykorzystujących szeroko stosowane techniki uczenia maszynowego np. *Residual Networks* [3], *2D-Bidirectional Long Short-Term Memory* [4,5], *U-net* [6], itd., pozwalających na przewidywanie struktury drugorzędowej [tzn. określenie dla każdego nukleotydu w wejściowej sekwencji czy jest on niesparowany czy też sparowany; jeśli jest sparowany to z którymi nukleotydami] na podstawie znanej sekwencji RNA. Dopuszczalne, a nawet wskazane jest czerpanie z rozwiązań opublikowanych w literaturze np. *SPOT-RNA* [7], czy też *UFold* [8], ale oczywiście można również zaproponować nowe rozwiązania o charakterze autorskim.

1 Wprowadzenie

1.1 Motywacja i biologiczna natura problemu.

Struktura drugorzędowa (2D) RNA **opisuje lokalnie uporządkowane motywy** powstałe w wyniku **wiązań wodorowych** zachodzących **między zasadami nukleotydów** wchodzącymi w interakcje. Przykładowa struktura drugorzędowa została zaprezentowana na Rysunku nr 1.



Rysunek 1. Przykładowe elementy topologiczne rozpatrywane w ramach struktury drugorzędowej RNA [9].

Znaczna część ludzkiego genomu ulega transkrypcji do niekodujących RNA, których struktury i funkcje wciąż nie są znane. Skuteczność prognozowania funkcji RNA wymaga dostępności

wiarygodnej struktury drugorzędowej, która znacząco ułatwia proces efektywnego modelowania struktury 3D RNA.

W cząsteczkach RNA wyróżniamy następujące **rodzaje par zasad**:

- **Kanoniczne** (Watson-Crick): A:U, G:C oraz para typu **Wobble**: G:U, gdzie A – Adenina, G – Guanina, C – Cytozyna, U – Uracyl.
- **Niekanonicze: klasy interakcji trzeciorzędowych** zaprezentowano na Rysunku nr 2.

Klasa	Symbol graficzny	Klasa	Symbol graficzny
cis Watson-Crick/ Watson-Crick		trans Watson-Crick/ Watson-Crick	
cis Watson-Crick / Hoogsteen		trans Watson-Crick / Hoogsteen	
cis Watson-Crick / Sugar		trans Watson-Crick / Sugar	
cis Hoogsteen / Hoogsteen		trans Hoogsteen / Hoogsteen	
cis Hoogsteen / Sugar		trans Hoogsteen / Sugar	
cis Sugar / Sugar		trans Sugar / Sugar	

Rysunek 2. Klasyfikacja niekanonicznych par zasad według Leontisa-Westhofa [10].

Analizując parowania w strukturach RNA należy mieć świadomość występowania pewnych szczególnych przypadków interakcji zachodzących pomiędzy nukleotydami, które zostały krótko zaprezentowane poniżej:

- **Izolowana para zasad** to taka para nukleotydów, dla których bezpośrednio sąsiadujące nukleotydy w sekwencji nie tworzą par z pozostałymi rozpatrywanymi nukleotydami.
- **Multiplet** zachodzi wtedy, jeśli pojedynczy nukleotyd tworzy wiązania wodorowe z więcej niż jednym spośród rozpatrywanych nukleotydów.
- **Pseudowęzły** zachodzą wtedy, gdy dla danej pary nukleotydów *(i, i')* istnieje inna para nukleotydów *(j, j')*, a numery porządkowe tworzących je nukleotydów (wskazujące na

lokalizację poszczególnych nukleotydów w rozważanej sekwencji) spełniają następującą zależność $i < j < i' \; AND \; j' < i \; OR \; j' > i'$.

Od przeszło dekady pojawiają się nowe metody do przewidywania struktur drugorzędowych RNA [11]. Aczkolwiek ich skuteczność od wielu lat utrzymuje się na stałym poziomie (nieprzekraczającym 80%), ponieważ istniejące rozwiązania, ze względu na ograniczenia technologiczne, często rozpatrują jedynie kanoniczne parowania zasad.

Podsumowując, wciąż nie jest znane jedno, standardowe podejście pozwalające wiarygodnie przewidywać strukturę drugorzędową RNA, a w szczególności parowania niekanoniczne, pseudowęzły oraz inne interakcje dalekiego zasięgu.

1.2 Jakie warianty problemu przewidywania struktury 2D RNA są rozpatrywane?

Rozważane są następujące warianty problemu przewidywania struktury drugorzędowej RNA:

- Dla każdego nukleotydu w wejściowej sekwencji RNA określa się jego przynależność do jednej z dwóch klas (liczba całkowita ze zbioru {0;1}): nukleotydów sparowanych (1), czy też niesparowanych (0). Potem stosowane są metody np. programowania dynamicznego w celu odczytania struktury drugorzędowej charakteryzujące się minimalnej wartością współczynnika energetycznego.
- Dla każdej pary różnych nukleotydów w wejściowej sekwencji RNA określa się jej przynależność do jednej z dwóch klas (liczba całkowita ze zbioru {0;1}): nukleotydów pomiędzy którymi zachodzą interakcje przestrzenne (tworzą się wiązania wodorowe) (1), czy też nie (0).
- Dla każdej pary różnych nukleotydów w wejściowej sekwencji RNA określa się wartość prawdopodobieństwa (liczba rzeczywista z przedziału <0;1>), że pomiędzy nimi tworzy się parowanie określonego typu (prawdopodobieństwo bliskie 1.0), czy też nie (prawdopodobieństwo bliskie 0.0).

Naszym celem jest opracowanie metody, która będzie potrafiła skutecznie i efektywnie rozwiązywać problem klasyfikacji dwuklasowej dla każdej pary różnych nukleotydów rozpatrywanych w wejściowej sekwencji RNA.

1.3 Użyteczne biblioteki.

Scikit-learn: https://github.com/scikit-learn/scikit-learn

Biopython: https://github.com/biopython/biopython

Keras: https://github.com/keras-team/keras

Tensorflow: https://github.com/tensorflow/tensorflow

PyTorch: https://github.com/pytorch

2 Dane

Udostępnione zostały dwa zbiory danych, a mianowicie: (1) 3970 sekwencji oraz manualnie udokładnionych struktur drugorzędowych RNA, pochodzących z 10 reprezentatywnych rodzin cząsteczek RNA [2] (katalog *ArchiveII*), oraz 594 sekwencje i struktury drugorzędowe wyekstrahowane z nieredundantnych, wysokorozdzielczych (rozdzielczość <= 3.0Å), eksperymentalnie określonych struktur 3D RNA pobranych z repozytorium *RNAsolo* [1] (katalog *PDB*). Każda cząsteczka RNA opisana jest przez sekwencje (w formacie *FASTA*) i strukturę drugorzędową (w formacie *BPSEQ*) przechowywane w plikach odpowiednio *.fa oraz *.bpseq. Format *FASTA* obejmuje dwa typy rekordów:

- Nagłówek (np. >1A9N_1_Q), w którym po znaku '>' często występuje identyfikator lub nazwa rozpatrywanej cząsteczki biologicznej.
- Sekwencja (np. cCUGGUAUUGCAGUACCUCCAGGU) będąca ciągiem znaków nad alfabetem {A,C,G,U,a,c,g,u}, gdzie małe litery zwykle oznaczają reszty modyfikowane. W naszym przypadku wielkość litery w sekwencji RNA nie ma żadnego znaczenia. Często dłuższe sekwencje zapisywane są w kilku kolejnych liniach, przy czym maksymalna długość pojedynczej linii jest wtedy określana liczbą 80 znaków. Tutaj sekwencja to zawsze jeden rekord o zmiennej długości.

Format *BPSEQ* obejmuje zbiór rekordów, przy czym każdy rekord opisuje pojedynczy nukleotyd (np. 1 c 23). Każdy nukleotyd opisywany jest z wykorzystaniem wartości dla trzech poniższych kolumn:

- Numer bieżącego nukleotydu w sekwencji (liczony od 1 do długości sekwencji).

- Jednoliterowy kod bieżącego nukleotydu (A/a – Adenina, G/g – Guanina, C/c – Cytozyna, U/u – Uracyl).
- Numer innego nukleotydu w sekwencji z którym bieżący nukleotyd wchodzi w interakcje przestrzenne (tzn. tworzy wiązania wodorowe) lub 0 jeśli bieżąca reszta nie jest sparowana.

3 Zadanie

Zadanie obejmuje następujące kroki:

1. Zapoznanie się z udostępnionymi zbiorami danych i ewentualne przetransformowanie ich do postaci ułatwiającej zastosowanie technik uczenia maszynowego np. integracja danych składowych z wykorzystaniem jednej spójnej reprezentacji.
2. Wybór odpowiedniej reprezentacji wejścia np. *one-hot encoding* wraz z uzasadnieniem.
3. Określenie sposobu reprezentacji wiedzy, którą dysponujemy.
4. Wybór obiecujących technik uczenia maszynowego, które zgodnie z oczekiwaniami powinny sprawdzić się podczas rozwiązywania postawionego problemu wraz z uzasadnieniem (np. *Residual Networks* [3], *2D-Bidirectional Long Short-Term Memory* [4,5], *U-net* [6], itd., a także ich dowolne kombinacje).
5. Określenie procentowych progów pozwalających podzielić dostępne zbiory danych na część treningową, walidacyjną i ewaluacyjną.
6. Przeprowadzenie i ewaluacja wpływu uczenia z transferem (ang. *transfer learning*) na skuteczność procesu predykcji poprzez:
    a. Przygotowanie modeli wytrenowanych najpierw na zbiorze *ArchiveII*, a w kolejnym kroku dotrenowanych również na zbiorze *PDB*.
    b. Przygotowanie modeli wytrenowanych niezależnie na zbiorze *ArchiveII* jak i *PDB*.
    c. Porównanie skuteczności procesu predykcji wszystkich trzech modeli z wykorzystaniem miary *Interaction Network Fidelity* (*INF*) [12], precyzja (*PPV*), czułość (*TPR*), specyficzność (*TNR*).
7. Iteracyjne przeprowadzenie procesu uczenia oraz określenie wartości kluczowych parametrów dla tego procesu.

8. Optymalizacja wartości hiperparametrów – czy warto je optymalizować w przypadku rozpatrywanego problemu? Jeśli tak to w jaki sposób?
9. Podsumowanie i analiza uzyskanych wyników.

Po wykonaniu zadania każda grupa przygotowuje krótką prezentację (~3 slajdy) przedstawiającą zaproponowane rozwiązanie wraz z uzasadnieniem oraz uzyskane wyniki.

Zadanie proszę zrealizować i przekazać prowadzącemu najpóźniej 6 czerwca.

4 Literatura

1. Adamczyk, B., Antczak, M., & Szachniuk, M., (2022). RNAsolo: a repository of clean, experimentally determined RNA 3D structures, *submitted*, https://rnasolo.cs.put.poznan.pl/
2. Sloma, M. F., & Mathews, D. H. (2016). Exact calculation of loop formation probability identifies folding motifs in RNA secondary structures. *RNA*, 22(12), 1808-1818, https://rnajournal.cshlp.org/content/22/12/1808.full.pdf+html
3. He, K., Zhang, X., Ren, S., & Sun, J. (2016, October). Identity mappings in deep residual networks. *In European conference on computer vision* (pp. 630-645). Springer, Cham, https://link.springer.com/chapter/10.1007/978-3-319-46493-0_38
4. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780, https://ieeexplore.ieee.org/abstract/document/6795963
5. Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11), 2673-2681, https://ieeexplore.ieee.org/abstract/document/650093
6. Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. *In International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Springer, Cham, https://link.springer.com/chapter/10.1007/978-3-319-24574-4_28

7. Singh, J., Hanson, J., Paliwal, K., & Zhou, Y. (2019). RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nature Communications*, 10(1), 1-13, https://www.nature.com/articles/s41467-019-13395-9
8. Fu, L., Cao, Y., Wu, J., Peng, Q., Nie, Q., & Xie, X. (2022). UFold: fast and accurate RNA secondary structure prediction with deep learning. *Nucleic Acids Research*, 50(3), e14-e14. https://academic.oup.com/nar/article/50/3/e14/6430845?login=true
9. Sato, K., Akiyama, M., & Sakakibara, Y. (2021). RNA secondary structure prediction using deep learning with thermodynamic integration. *Nature Communications*, 12(1), 1-9, https://www.nature.com/articles/s41467-021-21194-4
10. Leontis, N. B., & Westhof, E. (2001). Geometric nomenclature and classification of RNA base pairs. *RNA*, 7(4), 499-512, https://www.cambridge.org/core/journals/rna/article/abs/geometric-nomenclature-and-classification-of-rna-base-pairs/0D16AEB8C228306F9003E43276B91AAC
11. Puton, T., Kozlowski, L. P., Rother, K. M., & Bujnicki, J. M. (2013). CompaRNA: a server for continuous benchmarking of automated methods for RNA secondary structure prediction. *Nucleic Acids Research*, 41(7), 4307-4323, https://academic.oup.com/nar/article/41/7/4307/1072710?login=true
12. Parisien, M., Cruz, J. A., Westhof, É., & Major, F. (2009). New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *RNA*, 15(10), 1875-1885, https://rnajournal.cshlp.org/content/15/10/1875.full.pdf+html