„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 04: Generalna ocena jakości modeli przestrzennych RNA

Maciej Antczak

11 i 18 kwietnia 2024

Spis treści

Cele czwartego cyklu zajęć laboratoryjnych są następujące:

1. Zapoznanie się z udostępnionymi zbiorami danych obejmującymi struktury przestrzenne stosunkowo niewielkich oraz powtarzalnych motywów RNA wyekstrahowanych zarówno z (1) eksperymentalnie określonych struktur 3D RNA (tzw. struktury natywne) oraz (2) modeli 3D RNA uzyskanych z wykorzystaniem najnowocześniejszych aktualnie metod obliczeniowych, na podstawie podanej sekwencji RNA struktury natywnej, biorących udział w zakończonych rundach modelowania konkursu RNA-Puzzles. Struktury przedstawione zostały w dwojaki sposób zarówno w przestrzeni kartezjańskiej (w postaci zbioru składających się na nie atomów zapisanych w formacie PDB) jak i przestrzeni kątów torsyjnych (w postaci zbioru wybranych kątów torsyjnych opisujących każdy nukleotyd wchodzący w skład struktury przestrzennej rozpatrywanego motywu RNA). Struktura 3D dla każdego rozpatrywanego motywu RNA wyekstrahowana z każdego rozpatrywanego modelu została oceniona w kontekście odpowiadającej jej struktury natywnej z wykorzystaniem miary RMSD (Root-Mean-Square Deviation).
2. Opracowanie, wytrenowanie, walidacja i ewaluacja własnych podejść, wykorzystujących szeroko stosowane techniki sztucznej inteligencji np. głębokie sieci neuronowe, SVM, RandomForest, itd., pozwalających na przewidywanie wartości miary RMSD na podstawie struktur przestrzennych stosunkowo niewielkich, powtarzalnych motywów RNA reprezentowanych w przestrzeniach kartezjańskiej albo kątów torsyjnych lub obu. Dopuszczalne, a nawet wskazane jest czerpanie z rozwiązań zaprezentowanych na wykładzie, ale oczywiście można również zaproponować nowe rozwiązania o charakterze autorskim.

1 Wprowadzenie

1.1 Motywacja i biologiczna natura problemu.

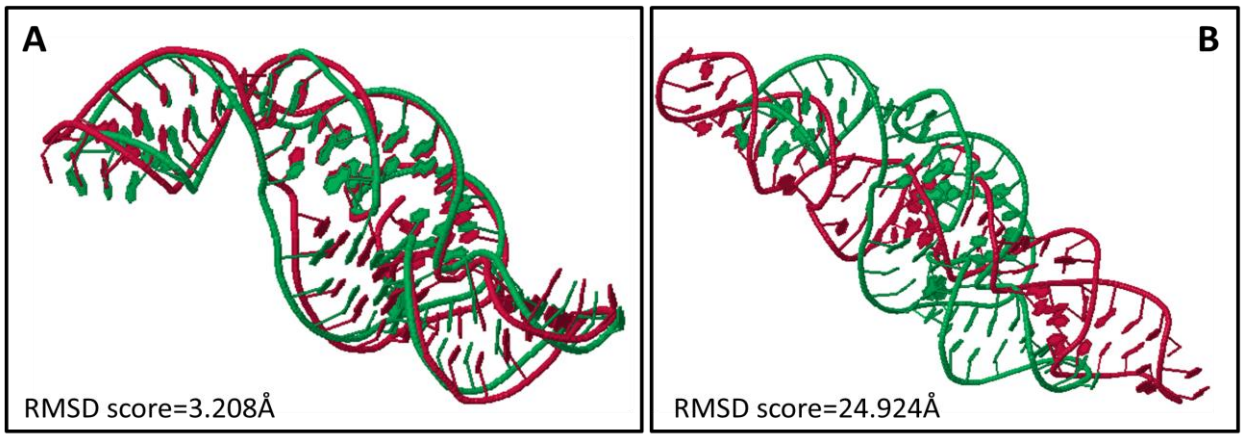
Modele 3D RNA przewidywane z wykorzystaniem aktualnie dostępnych metod obliczeniowych są często niedokładne i wykazują znaczne odchylenia od natywnej struktury przestrzennej [1]. Jakość struktury przestrzennej zwykle jest wypadkową wielu różnych aspektów strukturalnych. Wiarygodne oszacowanie *a priori* jakości modelu 3D RNA, uzyskanego technikami obliczeniowymi, jest szczególnie oczekiwane przez eksperymentatorów, ponieważ pozwala określić ograniczenia rozpatrywanego modelu, co jest kluczowe podczas szacowania jego praktycznej użyteczności w eksperymentach biochemicznych lub procesach projektowania leków. Niestety eksperymentatorzy często nie potrafią wybrać jednego najbardziej obiecującego modelu 3D RNA spośród wielu predykcji, uzyskanych metodami obliczeniowymi, którymi dysponują. Co więcej, rezultaty konkursu RNA-Puzzles (RNA-Puzzles to międzynarodowa inicjatywa, której celem jest kompleksowa ocena najnowocześniejszych metod *in silico* przewidywania struktur 3D RNA) jednoznacznie wskazują, że nawet projektanci metod obliczeniowych, biorących udział w konkursie, często nie potrafią trafnie wskazać ich własnej predykcji, która w wyniku ewaluacji okaże się najbliższa natywnej strukturze przestrzennej.

Podsumowując, wciąż nie jest znane jedno standardowe podejście pozwalające trafnie rozpoznawać struktury o wysokiej rozdzielczości, szczególnie jeśli eksperymentalnie określona struktura referencyjna nie jest znana.

1.2 Jakość modelu 3D RNA.

Strukturę przestrzenną RNA określamy jako wysokorozdzielczą (tzn. charakteryzującą się wysoką jakością), jeżeli jej kształt w przestrzeni kartezjańskiej nie odbiega znacząco od struktury natywnej (tzn. struktury eksperymentalnie określonej). Innymi słowy model 3D RNA jest dobrej jakości jeżeli odległość pomiędzy nim, a strukturą referencyjną jest stosunkowo niewielka. Na Rysunku nr 1 zaprezentowano wizualizację przestrzenną dwóch przykładowych predykcji obliczeniowych (zaprezentowanych kolorem czerwonym) nałożonych na tą samą strukturę

natywną (zaprezentowaną kolorem zielonym) wraz z odpowiadającymi im wartościami miary RMSD (Root-Mean-Square Deviation).



Rysunek 1. Przykładowe wizualizacje modeli 3D RNA (zaprezentowanych kolorem czerwonym), których kształt przestrzenny odpowiada (A) lub odbiega (B) od eksperymentalnie określonej struktury natywnej (zaprezentowanej kolorem zielonym).

1.3 Jakie warianty problemu oceny jakości struktur 3D RNA są rozważane?

Następujące warianty problemu oceny jakości struktur 3D RNA są standardowo rozważane:

- Ze względu na dostępność struktury referencyjnej (natywnej):
  - ✓ Ocena jakości modelu(i) w kontekście struktury referencyjnej.
  - ✓ Ocena jakości modelu(i) nie dysponując strukturą referencyjną.
- Ze względu na skalę oceny:
  - ✓ Ocena globalna, gdzie zwracana jest dokładnie jedna wartość dla modelu.
  - ✓ Ocena lokalna, gdzie ocenie poddawane są lokalne otoczenia wyznaczone dla kolejnych reszt wchodzących w skład rozpatrywanego modelu.
- Ze względu na liczbę rozpatrywanych modeli:
  - ✓ Ocena pojedynczego modelu.
  - ✓ Ocena bazująca na konsensusie wyznaczonym dla wielu analizowanych modeli.

Naszym celem jest opracowanie metody, która będzie potrafiła ocenić wiarygodność pojedynczego modelu 3D RNA w perspektywie lokalnej (bazujemy na stosunkowo niewielkich, powtarzalnych motywach konstruowanych w otoczeniu kolejnych nukleotydów wchodzących w

skład ocenianej struktury 3D RNA) jak i globalnej (poprzez agregację np. uśrednienie ocen cząstkowych), bez potrzeby posiadania struktury referencyjnej.
Do rozwiązania tego problemu zaproponowane zostały już dwie metody wykorzystujące techniki uczenia maszynowego, a mianowicie RNA3DCNN [3] oraz ARES [4]. Ta druga uzyskała zdecydowanie lepsze wyniki niż konkurencja.

1.4 Odchylenie średniokwadratowe (ang. Root-Mean-Square Deviation)

Odchylenie średniokwadratowe (RMSD) to miara odległości wyznaczana pomiędzy obiektami w przestrzeni *n*-wymiarowej (n≥2). W przypadku struktur cząsteczek biologicznych reprezentuje odległość pomiędzy dwoma równolicznymi zbiorami atomów optymalnie nałożonych na siebie w przestrzeni trójwymiarowej np. dla struktur A, B wartość miary wyznaczana jest na podstawie poniższego wzoru:

$RMSD(A,B) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d(a_i, b_i)^2}$, gdzie *d(ai,bi)* jest odległością Euklidesową między parą atomów a_i, b_i (*i=1..N*). Generalnie, im wyższa wartość tej miary tym struktury są do siebie bardziej niepodobne (tzn. oddalone od siebie w przestrzeni). W przypadku stosowania miary RMSD należy mieć świadomość jej ograniczeń:

- Nadaje się do porównywania struktur podobnych z założenia (np. wielu modeli przewidzianych na podstawie tej samej wejściowej sekwencji RNA).
- Nie uwzględnia charakterystycznych elementów struktur 3D RNA:
  - ✓ Nie uwzględnia interakcji przestrzennych wynikających z parowania zasad (ang. base-pairing).
  - ✓ Nie uwzględnia wzajemnych orientacji przestrzennych sąsiadujących ze sobą w sekwencji nukleotydów (ang. base-stacking).
- Nie jest normalizowana względem rozmiaru cząsteczki.

1.5 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

2 Dane

Pobrano pierwsze dziesięć wyzwań (od pz01 do pz10), o rozmiarach między 46, a 188 nukleotydów, opublikowanych w międzynarodowym konkursie RNA-Puzzles [2], na które nadesłano od 12 do 52 modeli 3D RNA. Każde spośród rozważanych wyzwań dysponowało dokładnie jedną strukturą eksperymentalnie określoną, która następnie była wykorzystywana w celu ewaluacji najnowocześniejszych metod obliczeniowych przewidywania struktur 3D RNA biorących udział w konkursie. Podstawowymi problemami podczas oceny struktur cząsteczek biologicznych, których należy mieć świadomość to (1) wrażliwość motywów na rotacje i przesunięcie w przestrzeni trójwymiarowej, (2) często duże różnice w rozmiarach analizowanych struktur 3D RNA oraz (3) niewystarczająco różnorodny zbiór danych wykorzystywany w procesie uczenia. Wszystkie te problemy można w pewnym sensie zaadresować jeśli reprezentuje się strukturę wejściową jako zbiór mniejszych motywów przestrzennych zbudowanych w lokalnym otoczeniu kolejnych nukleotydów wchodzących w skład rozpatrywanej cząsteczki RNA. W przeprowadzonej analizie każdy nukleotyd był reprezentowany przez atom centralny o nazwie C5’. Następnie wyszukiwane były wszystkie inne nukleotydy znajdujące się w kontakcie przestrzennym z określonym nukleotydem centralnym (tzn. odległość pomiędzy odpowiadającymi sobie atomami C5’ nie mogła przekroczyć 16Å). W taki sposób dla każdego nukleotydu (oprócz dwóch pierwszych i dwóch ostatnich reszt w cząsteczce) w rozpatrywanej strukturze referencyjnej utworzone zostało sąsiedztwo przestrzenne, które obejmuje od jednego do kilku nieciągłych fragmentów zwanych segmentami tworzących stosunkowo niewielki, powtarzalny motyw przestrzenny RNA. Następnie dla każdego modelu zgłoszonego w ramach danego wyzwania wyekstrahowane zostały struktury przestrzenne odpowiadające wszystkim motywom zidentyfikowanym w ramach bieżącej struktury referencyjnej. W końcu dla każdego rozpatrywanego motywu ocenione zostały struktury przestrzenne pochodzące z rozpatrywanych modeli w ramach wyzwania w kontekście odpowiadających im motywów pochodzących ze struktury referencyjnej poprzez wyznaczenie wartości miary RMSD z wykorzystaniem pakietu RNAQUA

[2]. Ostatecznie dla każdej struktury 3D motywu, przechowywanej w formacie PDB, wygenerowano wartości wszystkich kątów torsyjnych szkieletu oraz po jednym reprezentancie dla rybozy (δ) i zasady (χ).

Udostępnione zbiory danych obejmują:

1. Lista motywów (*filter-results.txt*) zidentyfikowanych dla struktury referencyjnej/natywnej określonego wyzwania np. dla wyzwania nr 1 (pz01) dysponujemy następującymi motywami, gdzie każdy z nich jest opisany przez nazwę pliku bez rozszerzenia, liczbę nieciągłych segmentów, liczbę reszt wchodzących w jego skład, zakresy nukleotydów i odpowiadające im sekwencje:

```
1_solution_0_rpr_A_3_G       2       15      A1-A7, B11-B18            CCGCCGC, AUGCCUGU
1_solution_0_rpr_A_4_C       3       20      A1-A8, B10-B16, B19-B23   CCGCCGCG, CAUGCCU, GGCGG
…
1_solution_0_rpr_A_20_G      2       16      A15-A23, B3-B9            CUGUGGCGG, GCCGCGC
1_solution_0_rpr_A_21_C      2       14      A17-A23, B2-B8            GUGGCGG, CGCCGCG
1_solution_0_rpr_B_3_G       2       17      A10-A18, B1-B8            CAUGCCUGU, CCGCCGCG
1_solution_0_rpr_B_4_C       3       20      A10-A16, A19-A23, B1-B8   CAUGCCU, GGCGG, CCGCCGCG
…
1_solution_0_rpr_B_20_G      2       15      A3-A9, B16-B23            GCCGCGC, UGUGGCGG
1_solution_0_rpr_B_21_C      2       14      A2-A8, B17-B23            CGCCGCG, GUGGCGG
```

   UWAGA! Warto ograniczyć analizy do motywów strukturalnie złożonych składających się z przynajmniej dwóch, a najlepiej trzech lub więcej segmentów.

2. Strukturę przestrzenną motywu w formacie PDB pochodzącą ze struktury referencyjnej np. *1_solution_0_rpr_B_6_G.pdb*.

```
ATOM      1  P     C A   7       8.267  -8.375  43.687  1.00 29.79           P
ATOM      2  OP1   C A   7       7.102  -9.153  44.095  1.00 29.57           O
…
```

3. Zbiór kątów torsyjnych dla kolejnych nukleotydów wchodzących w skład motywu pochodzącego ze struktury referencyjnej np. *1_solution_0_rpr_B_6_G.tor*.

```
A   7       -       C       -       165.734  42.144   87.83    -152.87  -72.211  -155.851
A   8       -       G       -64.037 175.025  49.996   81.011   -144.913 -72.567  -161.947
…
A   22      -       G       -63.64  170.296  58.638   85.349   -158.751 -70.986  -162.775
A   23      -       G       -72.327 -175.431 55.452   81.464   -        -        -160.009
B   1       -       C       -       174.367  53.309   73.698   -148.241 -81.313  -171.612
B   2       -       C       -61.367 161.154  47.146   76.102   -149.199 -77.342  -163.78
…
B   9       -       C       -66.887 171.792  50.498   83.131   -149.572 -68.426  -157.499
B   10      -       C       -65.639 172.832  56.562   76.467   -        -        -162.329
```

Każdy nukleotyd jest opisany przez identyfikator łańcucha/nici (np. ‘A’), numer nukleotydu/reszty (np. 7), iCode (zwykle ‘-‘), nazwa reszty (np. ‘C’ – cytozyna), wartości kolejnych kątów torsyjnych (α, β, γ, δ, ε, ζ, χ).

4. Zbiór katalogów, w którym znajdują się struktury przestrzenne w formacie PDB oraz wartości kątów torsyjnych wyekstrahowane z modeli dla określonego motywu pochodzącego ze struktury referencyjnej np. *1_solution_0_rpr_B_6_G* (nazwa katalogu to zawsze nazwa pliku PDB motywu pochodzącego ze struktury referencyjnej bez rozszerzenia).
5. Wartości miary RMSD wyznaczone dla wszystkich rozpatrywanych struktur motywów pochodzących z modeli (w katalogu z punktu 4) w kontekście struktury referencyjnej (plik PDB) i zapisane w formacie XML np. *1_solution_0_rpr_B_6_G-rmsd.xml*

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<measureScores>
  <structure>
    <description>
      <filename>1_bujnicki_1_rpr.pdb</filename>
      <errors/>
    </description>
    <score>3.376</score>
  </structure>
  …
  <structure>
    <description>
      <filename>1_chen_1_rpr.pdb</filename>
      <errors/>
    </description>
    <score>3.05</score>
  </structure>
  <structure>
    <description>
      <filename>1_das_1_rpr.pdb</filename>
      <errors/>
    </description>
    <score>3.213</score>
  </structure>
  …
  <structure>
    <description>
      <filename>1_dokholyan_1_rpr.pdb</filename>
      <errors/>
    </description>
    <score>6.128</score>
  </structure>
  <structure>
    <description>
      <filename>1_major_1_rpr.pdb</filename>
      <errors/>
    </description>
    <score>4.041</score>
  </structure>
  <structure>
    <description>
      <filename>1_santalucia_1_rpr.pdb</filename>
      <errors>
        <error>There is no residue C_ _15_ _A in model 1_santalucia_1_rpr.pdb.</error>
        <error>There is no residue U_ _18_ _A in model 1_santalucia_1_rpr.pdb.</error>
      </errors>
    </description>
    <score>3.513</score>
  </structure>
</measureScores>
```

Niekiedy mogą się pojawić informacje, że danej reszty w modelu nie było, tak jak np. „There is no residue U_ _18_ _A in model 1_santalucia_1_rpr.pdb”. To nie jest problem, ponieważ wartość miary RMSD jest wyznaczana zawsze na podstawie odwzorowania zbiorów nukleotydów, które występują zarówno w strukturze referencyjnej jak i modelu.

3 Zadanie

Zadanie obejmuje następujące kroki:

1. Zapoznanie się z udostępnionymi zbiorami danych i ewentualne przetransformowanie ich do postaci ułatwiającej zastosowanie technik sztucznej inteligencji np. integracja danych składowych przechowywanych w różnych formatach z wykorzystaniem jednej spójnej reprezentacji.
2. Krótkie zapoznanie się z dostępnymi przestrzeniami reprezentacji struktur 3D RNA (przestrzenie kartezjańska i kątów torsyjnych) i ich formatami zapisu. Wybór obiecującej przestrzeni na której będziecie Państwo bazować wraz z uzasadnieniem.
3. Określenie procentowych progów pozwalających podzielić dostępny zbiór danych na część treningową, walidacyjną i ewaluacyjną. Czy rozmiar dostępnego zbioru jest wystarczający? Czy należy go rozbudować? Jeśli tak to w jaki sposób?
4. Określenie sposobu reprezentacji wiedzy, którą dysponujemy (tzn. wektora cech). Czy stosowane będą techniki identyfikacji najistotniejszych cech? Jeśli tak to jakie?
5. Wybór obiecujących technik uczenia maszynowego, które uważacie Państwo, że powinny się sprawdzić podczas rozwiązywania postawionego problemu wraz z uzasadnieniem (np. głębokie sieci neuronowe, SVM, RandomForest ,itd.).
6. Iteracyjne przeprowadzenie procesu uczenia, określenie wartości parametrów kluczowych dla tego procesu (np. zastosowana funkcja straty, learning rate, optimizer, itd.) i wskazanie czy natrafiliście Państwo na jakieś problemy podczas tego procesu np. przeuczenie i jak Państwo sobie z tymi problemami poradziliście o ile rzeczywiście wystąpiły?
7. Optymalizacja wartości hiperparametrów – czy warto je optymalizować w przypadku rozpatrywanego problemu? Jeśli tak to w jaki sposób?

8. Wybór i uzasadnienie zastosowanych miar oceny, przeprowadzenie procesu ewaluacji uzyskanego(ych) modelu(i), podsumowanie i analiza uzyskanych wyników.
9. (opcjonalnie) Porównanie najlepiej sprawdzającego się modelu w kontekście innych dostępnych rozwiązań opublikowanych w literaturze [3-5].

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

Zadanie proszę zrealizować przed kolejnym spotkaniem, które odbędzie się 18 kwietnia.

4 Literatura

1. Carrascoza, F., Antczak, M., Miao, Z., Westhof, E. & Szachniuk, M. (2022). Evaluation of the stereochemical quality of predicted RNA 3D models in the RNA-Puzzles submissions, *RNA*, 28(2), 250–262, https://rnajournal.cshlp.org/content/early/2021/11/24/rna.078685.121
2. Magnus, M., Antczak, M., Zok, T., Wiedemann, J., Lukasiak, P., Cao, Y., ... & Miao, Z. (2020). RNA-Puzzles toolkit: a computational resource of RNA 3D structure benchmark datasets, structure manipulation, and evaluation tools. *Nucleic Acids Research*, 48(2), 576-588, https://academic.oup.com/nar/advance-article-pdf/doi/10.1093/nar/gkz1108/31196661/gkz1108.pdf
3. Li, J., Zhu, W., Wang, J., Li, W., Gong, S., Zhang, J., & Wang, W. (2018). RNA3DCNN: Local and global quality assessments of RNA 3D structures using 3D deep convolutional neural networks. *PLoS Computational Biology*, 14(11), e1006514, https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006514
4. Townshend, R. J., 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, https://www.science.org/doi/10.1126/science.abe5650
5. Zablocki, M. , Szachniuk, M., & Antczak, M. (2019). Machine learning approach for general quality assessment of tertiary RNA structures, *PTBI'19: 12th Symposium of the Polish Bioinformatics Society*, 19-21.09.2019, Cracow, Poland.

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