„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 02: Odkrywanie konserwatywnych wzorców w rezultatach próbkowania miejsc wiązań RNA-białko

Maciej Antczak

14 i 21 marca 2024

Spis treści

Cele drugiego cyklu zajęć laboratoryjnych są następujące:

1. Zapoznanie się z udostępnionymi zbiorami danych obejmującymi rezultaty eksperymentalnego próbkowania transkryptów metodą fSHAPE w kontekście dwóch białek wiążących RNA, a mianowicie HNRNPC (*Heterogeneous Nuclear Ribonucleoprotein C*) oraz HNRNPA2B1 (*Heterogeneous Nuclear Ribonucleoprotein A2/B1*).
2. Przeprowadzenie analizy skupień zbioru wybranych motywów spełniających założoną charakterystykę miejsca wiązania dla rozpatrywanego białka w celu odkrycia nieznanych, konserwatywnych wzorców w danych reprezentujących potencjalnie obiecujące miejsca wiązania RNA-białko.

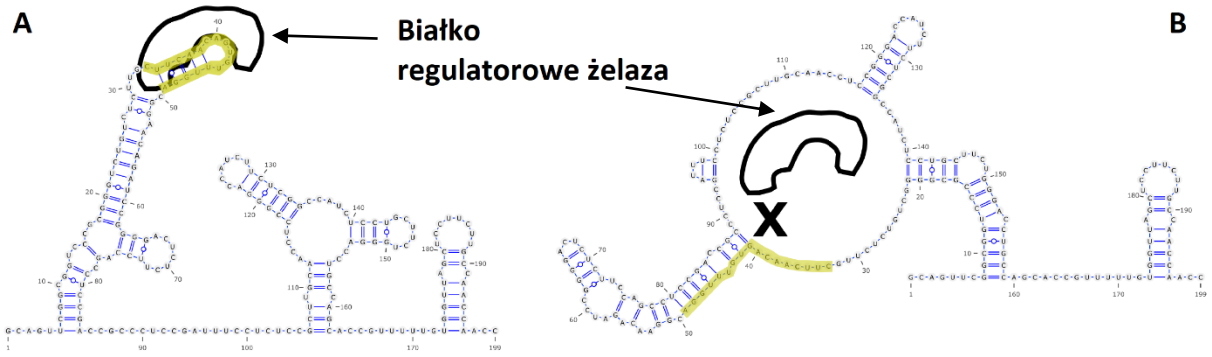
# 1 Wprowadzenie

## 1.1 Motywacja i biologiczna natura problemu.

Białka wiążące RNA (RBP) odgrywają kluczową rolę w procesie regulacji potranskrypcyjnej. Zrozumienie mechanizmów rządzących procesem wiązania białek i RNA oraz umiejętność lokalizowania potencjalnie obiecujących miejsc wiązań ma kluczowe znaczenie dla zrozumienia procesu regulacji ekspresji genów. W ludzkim genomie wyróżniamy setki białek tworzących kompleksy tego typu. Niestety zaledwie dla niewielkiego odsetka z nich dostępna jest

szczegółowa wiedza na temat interakcji międzycząsteczkowych (tzn. wiązań wodorowych) zachodzących w miejscu wiązania (ang. *binding site*) pomiędzy białkiem i określonym, specyficznym motywem w strukturze RNA. W znakomitej większości przypadków aktualna wiedza dotycząca zarówno sekwencji jak i struktury drugorzędowej cząsteczek wchodzących w interakcję jest niewystarczająca do precyzyjnego przewidywania potencjalnych miejsc wiązań białek tego typu w obrębie różnych, znanych transkryptów RNA [1,2].

Jako przykład dobrze poznanego kompleksu tego typu warto tutaj przywołać białko (IRP) wiążące specyficzny motyw IRE (*Iron Response Element*) – spinkę do włosów (ang. *hairpin*) – zlokalizowaną w nieulegającym translacji regionie 5' (tzw. 5'-UTR) transkryptu FTL (mRNA). Celem tej interakcji jest regulacja translacji genu FTL. Przykładowa wizualizacja właściwego wiązania RNA-białko została zaprezentowana na Rysunku 1A.




*Rysunek 1. Schematyczna wizualizacja miejsca wiązania białka (IRP) ze spinką zlokalizowaną w regionie 5'-UTR transkryptu FTL (źródło: [1,3]).*

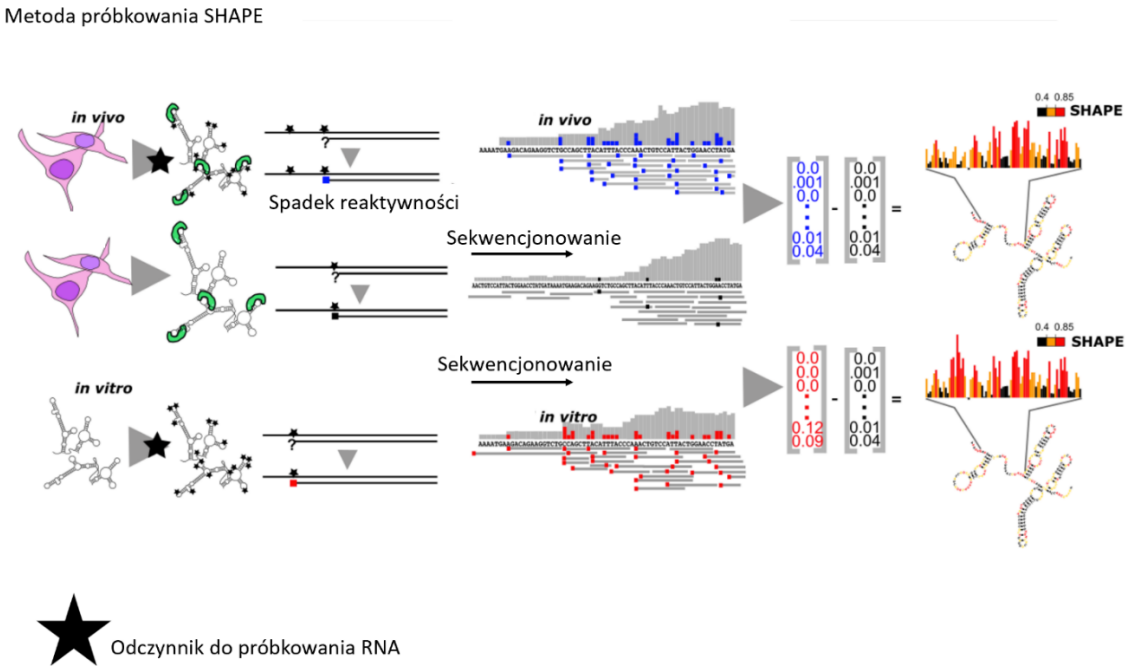
Polimorfizm pojedynczych nukleotydów (ang. *Single-Nucleotide Polymorphism*) w 5'-UTR genu łańcucha lekkiego ferrytyny (FTL) powoduje hiperferrytynemię zaburzając proces translacji poprzez zmianę konformacji miejsca wiązania, która uniemożliwia właściwą interakcję białka regulatorowego żelaza (IRP) z mRNA FTL 5'-UTR [5]. Przykładowa wizualizacja konformacji uniemożliwiającej wiązanie RNA-białko została zaprezentowana na Rysunku 1B.

## 1.2 Metody eksperymentalnego próbkowania miejsc wiązań RNA-białko.

Zaprezentowane poniżej metody umożliwiają badanie i w dalszej perspektywie zrozumienie specyficznych interakcji zachodzących pomiędzy kwasami rybonukleinowymi i białkami tworzącymi kompleksy.

### 1.2.1 Charakterystyka metody SHAPE.

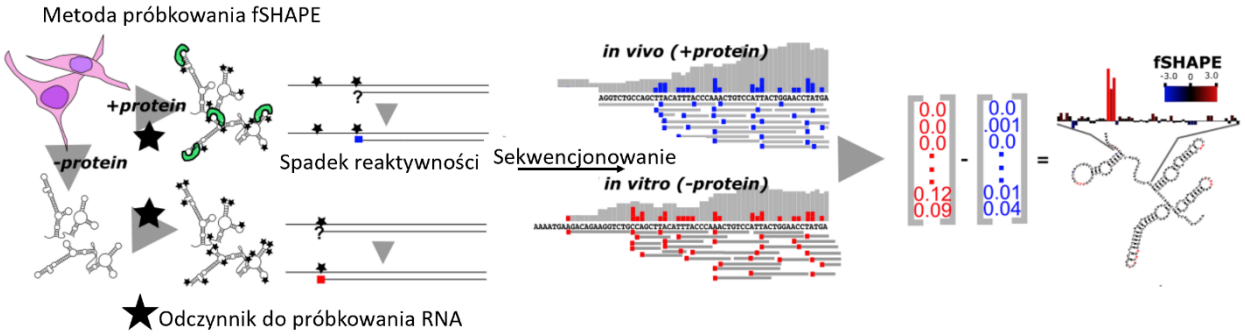
Próbkowanie struktury metodą SHAPE pozwala określić skłonność każdej zasady w transkrypcie do parowania RNA. Niskie wartości reaktywności (<0.4) dotyczą nukleotydów sparowanych (tzn. związanych wiązaniami wodorowymi) z innymi resztami transkryptu lub aminokwasami białka, z którym tworzą kompleks. Natomiast wysokie wartości reaktywności (>0.85) oznaczają nukleotydy niesparowane wchodzące w skład pojedynczych nici lub pętli. W ramach analizowanej metody rozpatrywane są trzy różne eksperymenty (zaprezentowane na Rysunku nr 2). W pierwszym z nich sekwencjonowaniu poddawany jest materiał w komórce (in vivo) obejmujący transkrypt, związane z nim białko oraz charakterystyczny odczynnik próbkowania RNA. W ostatnim przypadku sekwencjonowaniu poddaje się jedynie transkrypt wyekstrahowany z komórki (in vitro) uzupełniony odczynnikiem próbkowania RNA. W końcu drugi eksperyment, gdzie materiał ponownie badany w komórce obejmuje transkrypt oraz związane z nim białko, pozwala określić bazowe wartości reaktywności nukleotydów, które wykorzystywane są następnie w celu normalizacji wyników uzyskanych w dwóch poprzednich eksperymentach. Reaktywność nukleotydów w komórce (in vivo) obniża się zauważalnie dla tych nukleotydów spośród nich, które wchodzą w interakcję (tworzą wiązania wodorowe) z aminokwasami białka.




*Rysunek 2. Koncepcja próbkowania struktur 3D RNA metodą SHAPE (źródło: [1,3]).*

### 1.2.2 Charakterystyka metody fSHAPE (*footprinting SHAPE*).

Próbkowanie struktury metodą fSHAPE identyfikuje nukleotydy w transkrypcie oddziałujące z białkiem. Analizy strukturalne wskazują, że fSHAPE precyzyjnie wykrywa zasady azotowe w transkrypcie tworzące wiązania wodorowe z aminokwasami białka. W przypadku tej metody przeprowadzane są dwa eksperymenty sekwencjonowania (zaprezentowane na Rysunku nr 3), gdzie materiał badany jest odpowiednio w komórce (in vivo) i poza nią (in vitro). W obu przypadkach stosowany jest charakterystyczny odczynnik próbkowania RNA. W przypadku eksperymentu kontrolnego (in vitro), przeprowadzanego w celu normalizacji wartości reaktywności nukleotydów, badaniu poddawany jest transkrypt RNA wyizolowany z komórki. Wysokie wartości reaktywności nukleotydów (> 1.0) raportowane w ramach fSHAPE wskazują jednoznacznie miejsca wiązań RNA-białko.



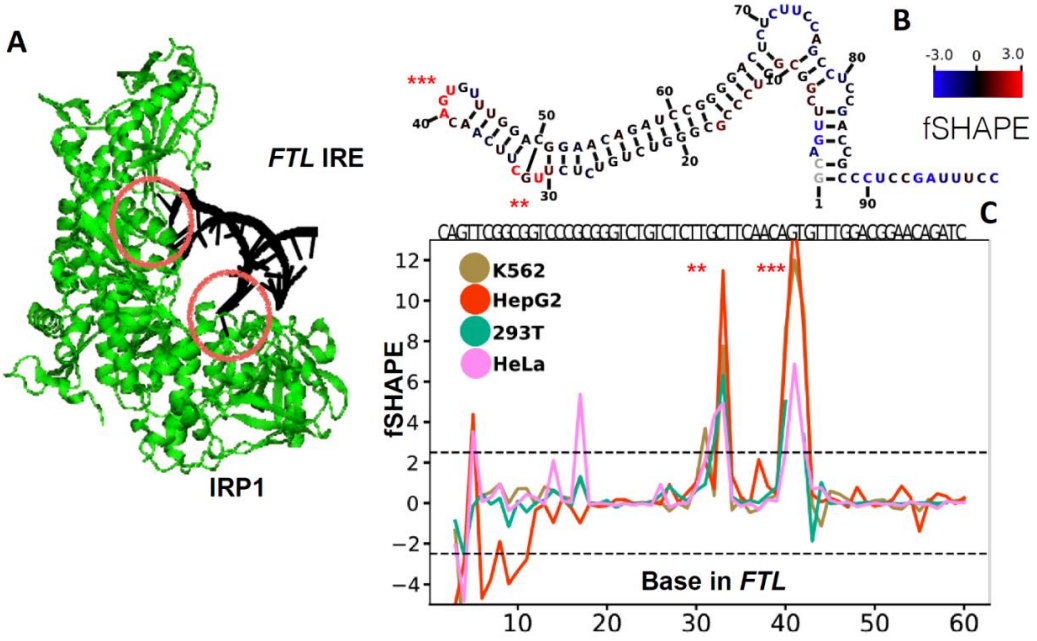

*Rysunek 3. Koncepcja próbkowania struktur 3D RNA metodą fSHAPE (źródło: [1,3]).*

W przypadku metody fSHAPE zakłada się, że wiązanie białka nie zmienia drastycznie struktury drugorzędowej RNA, a zatem zmiany, które obserwowane są w materiale pozbawionym białka, wynikają jednoznacznie z utraconych z nim interakcji.

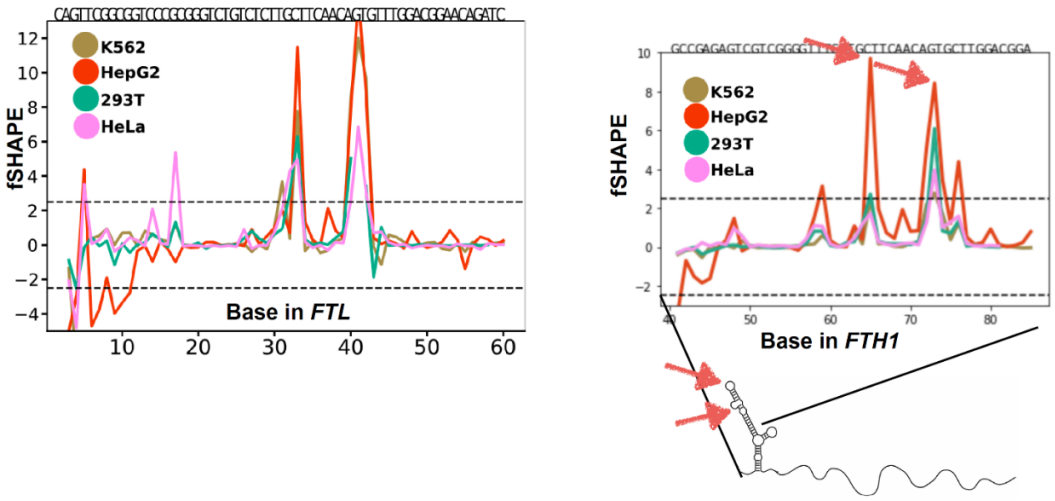
## 1.3 W jaki sposób identyfikować białka wiążące RNA?

Profil danych fSHAPE określony dla transkryptu jednoznacznie wskazuje miejsce wiązania RNA-białko. Dla przykładu (Rysunek 4), nukleotydy, a w szczególności ich zasady azotowe, wchodzące w skład pętli apikalnej oraz wybrzuszenia utworzonych w ramach IRE transkryptu FTL tworzą interakcje z białkiem regulującym metabolizm żelaza (IRP1) co łatwo zaobserwować w postaci ich wysokich wartości reaktywności uzyskanych z próbkowania RNA metodą fSHAPE.

Badania strukturalne potwierdziły, że jeżeli dysponujemy niezależnymi transkryptami i wiadomo, że wiążą one to samo białko to eksperymentalnie określone profile danych fSHAPE dla zidentyfikowanych w ramach nich miejsc wiązania (obejmujących zwykle ok. 10 nukleotydów) będą charakteryzowały się wysoką korelacją (> 0.8). Przykładowe profile danych fSHAPE dla dwóch transkryptów wiążących to samo białko regulujące metabolizm żelaza zaprezentowano na Rysunku 5. W związku z powyższym możliwe jest odkrywanie potencjalnie obiecujących miejsc wiązań znanych białek regulatorowych dla transkryptów, dla których nie potwierdzono dotąd eksperymentalnie ich występowania poprzez wyszukiwanie konserwatywnego wzorca w sekwencji (np. [CG]NNNNNCAG[AT]G, gdzie N reprezentuje dowolny nukleotyd) oraz weryfikację stosunkowo wysokiej korelacji profili danych fSHAPE pomiędzy znanym i potencjalnie obiecującym miejscem wiązania.




*Rysunek 4. Charakterystyka miejsca wiązania przykładowego kompleksu RNA-białko. A: Wizualizacja struktur 3D odpowiednio białka (oznaczonego kolorem zielonym) oraz fragmentu RNA (oznaczonego kolorem czarnym) tworzących kompleks. B: Wizualizacja struktury drugorzędowej fragmentu RNA, gdzie kolorem czerwonym oznaczono nukleotydy (** - wybrzuszenie, *** - pętla apikalna), których reszty azotowe wchodzą w interakcje z białkiem. C: Wykres przedstawiający rezultat próbkowania fragmentu struktury RNA z wykorzystaniem metody fSHAPE (źródło: [1,3]).*




*Rysunek 5. Profile danych fSHAPE dla fragmentów transkryptów odpowiednio FTL i FTH1, które wchodzą w interakcje z białkiem regulującym metabolizm żelaza (źródło: [1,3]).*

Co więcej dysponując zbiorem zestawów danych fSHAPE dla transkryptów, o których wiadomo, że wiążą białko, ale nie wiadomo do końca jakie to białko, można zastosować, m. in., metody uczenia maszynowego w celu odkrywania obiecujących, konsensusowych motywów o długości między 6, a 14 nukleotydów o założonej charakterystyce miejsca wiązania w oparciu o współczynnik reaktywności uzyskany metodą fSHAPE (tzn. wartość reaktywności dla przynajmniej jednej reszty w ramach motywu musi przekraczać wartość 1.0), które następnie mogą być weryfikowane w kontekście eksperymentalnie potwierdzonych miejsc wiązań RNA-białko dla których dysponujemy profilami danych fSHAPE.

## 1.4 Metody odkrywania wzorców w danych z wielu szeregów czasowych.

Zbiór danych uzyskany z próbkowania pojedynczego transkryptu metodą fSHAPE jest często traktowany jako szereg czasowy, gdyż obejmuje wartości reaktywności przypisane kolejnym nukleotydom wchodzącym w skład analizowanej struktury RNA. Dlatego eksploracja wielu szeregów czasowych (bo zwykle dysponujemy wieloma potencjalnymi transkryptami) powinna pozwolić nam odkrywać obiecujące, konsensusowe motywy w nich występujące, które po eksperymentalnym potwierdzeniu mogą okazać się rzeczywistymi, dotąd nieznanymi miejscami wiązania RNA-białko.

Odkrywanie motywów ma na celu poznanie nowej, wartościowej wiedzy na podstawie istniejących danych. Pod pojęciem motywu rozumiemy powtarzalny wzorzec zidentyfikowany w

analizowanym szeregu czasowym bez posiadania informacji o jego lokalizacji lub kształcie. W przypadku szeregów czasowych, zwykle odkrywane są reguły lub określone zdarzenia w analizowanym sygnale, które często dostarczają informacji przydatnych podczas modelowania lub analizy danych. Problemy tego typu są rozważane w różnych obszarach, takich jak np. telekomunikacja, medycyna, IoT [4]. Głównym celem jest odkrywanie w danych motywów konsensusowych charakteryzujących się najmniejszą odległością zachodzącą między nimi.

## 1.5 W jaki sposób porównywać motywy?

Aby określić odległość pomiędzy profilami danych fSHAPE porównywanych ze sobą motywów należy wykorzystać miarę *z-normalized Euclidean distance* (*znEd*). Jeżeli wartość tej miary dla określonej pary motywów nie przekracza wartości 2.5 to można je uznać za obiecujące i wtedy rankingować motywy po zgodności sekwencyjnej. W kolejnym kroku wyznaczany jest *współczynnik podobieństwa sekwencyjnego* (*ssf*), który jest sumą wartości cząstkowych, wyznaczanych dla każdego rozpatrywanego nukleotydu w motywie zakładając: *2 pkt* za identycznie odwzorowane nukleotydy tzn. A<->A, G<->G, C<->C, U<->U, T<->T, *1 pkt* za odwzorowanie w zakresie odpowiednio puryn (tzn. A<->G, G<->A) jak i pirymidyn (tzn. C<->U, U<->C, C<->T, T<->C), oraz 0 pkt w przeciwnym wypadku, znormalizowaną przez długość motywu. Im wyższa wartość współczynnika tym większe podobieństwo między motywami. Na przykład (N oznacza dowolny nukleotyd):

NTTTTN

TGATTT

Podobieństwo sekwencyjne = (2 + 0 + 0 + 2 + 2 + 2)/6 = 8/6 = 1.33(3). Jeżeli chcemy zastosować jedną wartość opisującą podobieństwo to możemy skorzystać z poniższego wzoru *αS = 10 * znEd - ssf*. Oczywiście im mniejsza wartość wyznaczona dla *αS* tym lepiej.

## 1.6 Użyteczne biblioteki.

MatrixProfile: https://github.com/matrix-profile-foundation/matrixprofile

STUMPY: https://github.com/TDAmeritrade/stumpy

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

# 2 Dane

Udostępnione zbiory danych obejmują rezultaty eksperymentalnego próbkowania transkryptów metodą fSHAPE w kontekście dwóch białek wiążących RNA, a mianowicie HNRNPC (*Heterogeneous Nuclear Ribonucleoprotein C*) oraz HNRNPA2B1 (*Heterogeneous Nuclear Ribonucleoprotein A2/B1*). W ramach każdego rozpatrywanego białka rozpatrywane są trzy pliki (jeden plik tekstowy i dwa archiwa):

- *[hnrnpc|hnrnpa2b1]_expected_pattern* – plik tekstowy zawierający oczekiwany motyw skojarzony z miejscem wiązania RNA-białko wyznaczony dla określonego kompleksu na podstawie eksperymentalnie określonej struktury przestrzennej (uśredniona liczba wiązań wodorowych dla nukleotydów rozpatrywanych w ramach miejsc wiązań z białkami koreluje z ich wysoką reaktywnością wyznaczoną metodą fSHAPE). Plik zawiera dane w postaci tabelarycznej (bez nagłówka) obejmując następujące kolumny (oddzielone tabulatorami), a mianowicie wartość współczynnika fSHAPE oraz nazwę zasady (A/C/T/G).
- *[hnrnpc|hnrnpa2b1]_binding_sites_fshape.zip* – archiwum obejmujące zestaw krótkich fragmentów transkryptów służących do odkrywania powtarzalnych wzorców stanowiących konsensus, które mogą stanowić nieznane dotąd, potencjalnie obiecujące miejsca wiązań z rozpatrywanym białkiem. Każdy plik zawiera dane w postaci tabelarycznej (bez nagłówka) obejmując następujące kolumny (oddzielone tabulatorami), a mianowicie wartość współczynnika fSHAPE, nazwę zasady (A/C/T/G) oraz opcjonalnie wartość współczynnika SHAPE.
- *[hnrnpc|hnrnpa2b1]_ search _fshape.zip* – archiwum obejmujące zestaw transkryptów, które będą przeszukiwane z wykorzystaniem konserwatywnego(ych) motywu(ów) odkrytego(ych) w poprzednim kroku w celu identyfikacji motywów charakteryzujących się zbliżonym profilem danych fSHAPE (*z-normalized Euclidean distance* nie powinien przekraczać wartości 2.5) oraz wysokim współczynnikiem podobieństwa sekwencyjnego.

Wszystkie pliki powinny zawierać co najmniej jeden długi, ciągły fragment nici RNA obejmujący określone wartości współczynnika fSHAPE. Pliki mogą zawierać również długie

ciągi wartości nieokreślonych (*NaN*) zarówno w ramach kolumn SHAPE jak i fSHAPE, z którymi należy radzić sobie podczas identyfikacji obiecujących motywów.

# 3 Zadanie

Zadanie obejmuje następujące kroki:

1. Wybór białka (HNRNPC/HNRNPA2B1) analizowanego podczas eksperymentów, co jednoznacznie identyfikuje wykorzystywany zestaw danych.
2. Zapoznanie się z motywem oczekiwanym, wyekstrahowanym na podstawie eksperymentalnie określonej struktury 3D kompleksu, utworzonego przez analizowane białko wiążące z RNA, w celu identyfikacji liczby wchodzących w jego skład nukleotydów (*len*). Załóżmy, że nasz motyw oczekiwany to TTTT, wtedy *len* = 4.
3. Określenie długości potencjalnie obiecujących motywów *w* = {*len*, *len + 1*, *len + 2*}. Załóżmy, że nasz motyw oczekiwany to TTTT (*len* = 4), wtedy *len + 1* = {NTTTT, TTTTN}, a *len + 2* = {NTTTTN}, gdzie N to dowolny nukleotyd.
4. Ekstrakcja wszystkich obiecujących, ciągłych motywów (reprezentowanych przez odpowiadające im profile danych fSHAPE) o długości *w* = *len* … *len + 2* spełniających założoną charakterystykę miejsca wiązania, w oparciu o wartości współczynnika reaktywności uzyskane metodą fSHAPE (tzn. wartość reaktywności dla przynajmniej jednego nukleotydu w ramach motywu musi przekraczać wartość 1.0), spośród krótkich fragmentów transkryptów zawartych w archiwum *identyfikator_wybranego_białka_binding_sites_fshape.zip*, które prawdopodobnie zawierają nieodkryte dotąd miejsca wiązań RNA z rozpatrywanym białkiem.
5. Przeprowadzenie analizy skupień z wykorzystaniem przynajmniej dwóch metod (np. KMEANS++, DBSCAN, itd.) dla zbioru motywów wyekstrahowanych w poprzednim kroku dla każdej długości motywu (*w*) analizowanej niezależnie.
6. Wyznaczenie obiecującego motywu konsensusowego (np. z wykorzystaniem biblioteki STUMPY) na podstawie profili danych fSHAPE zawartych w trzech najbardziej licznych klastrach zidentyfikowanych w poprzednim kroku (przy czym minimalna moc klastra, dla którego należy wyznaczać motyw konsensusowy nie może być niższa niż 3).

7. Przeprowadzenie przeszukiwania transkryptów (np. z wykorzystaniem biblioteki STUMPY), przechowywanych w archiwum *identyfikator_wybranego_białka_ search _fshape.zip*, na podstawie profili danych fSHAPE wyekstrahowanych dla obiecujących motywów konsensusowych zidentyfikowanych w poprzednim kroku oraz motywu oczekiwanego. Opracowanie rezultatów w postaci tabelarycznej, gdzie każdy motyw będzie opisany przez jego sekwencję, zakresy numerów nukleotydów, nazwę pliku transkryptu, w którym został zidentyfikowany oraz wartości następujących miar, a mianowicie *znEd*, *ssf*, *aS* wyznaczonych w kontekście motywu oczekiwanego. Rekordy powinny być uporządkowane w porządku niemalejącym według ostatniej kolumny (*aS*).

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

**Zadanie proszę zrealizować najpóźniej do dnia 28 marca 2024 r.**

# 4 Literatura


1. Corley, M., Antczak, M., Zok, T., Elhajjajy, S., Bernetti, M., & Adler J., (2021) "Predicting hnRNPA2/B1 binding sites from fSHAPE data", RBP Footprint Challenge @ RNA Society 2021 Meeting, 04.06.2021
2. Corley, M., Burns, M. C., & Yeo, G. W. (2020). How RNA-binding proteins interact with RNA: molecules and mechanisms. *Molecular Cell*, 78(1), 9-29, https://doi.org/10.1016/j.molcel.2020.03.011
3. Corley, M., Flynn, R. A., Lee, B., Blue, S. M., Chang, H. Y., & Yeo, G. W. (2020). Footprinting SHAPE-eCLIP reveals Transcriptome-wide hydrogen bonds at RNA-Protein interfaces. Molecular cell, 80(5), 903-914, https://doi.org/10.1016/j.molcel.2020.11.014
4. Torkamani, S., & Lohweg, V. (2017). Survey on time series motif discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(2), e1199, https://doi.org/10.1002/widm.1199
5. Pulos-Holmes, M. C., Srole, D. N., Juarez, M. G., Lee, A. S., McSwiggen, D. T., Ingolia, N. T., & Cate, J. H. (2019). Repression of ferritin light chain translation by human eIF3. Elife, 8, e48193, https://doi.org/10.7554/eLife.48193.001

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