

# Przewidywanie i porównywanie struktur 3D

Tomasz Żok

## 1 RMSD

1. RMSD (root mean square deviation) to miara odległości między dwoma zbiorami atomów:

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^N \delta_{i,j}^2}$$

- $N$  to liczba par atomów
  - $\delta_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$ , to odległość między atomami  $i$  oraz  $j$
2. W prawie każdym kontekście pod nazwą RMSD kryje się (1) superpozycja struktur oraz (2) wyznaczenie wartości ze wzoru z punktu powyżej. **W tym zadaniu skupiamy się tylko na (2)**
  3. Napisz program, który otrzyma na wejściu dwie struktury w formacie PDB oraz nazwę atomu. Program ma za zadanie wyznaczyć RMSD między strukturami tylko dla konkretnego typu atomu lub wypisać -1 jeżeli RMSD było niemożliwe do wyznaczenia (tj. gdy liczba atomów danego typu w porównywanych strukturach się nie zgadza)
  4. Uwagi:
    - Dokumentacja formatu PDB. Proszę poszukać informacji o formatowaniu linii ATOM oraz HETATM
    - Proszę przyjąć bardzo uproszczone założenia. Przykładowo, jeżeli porównujemy wg atomu fosforu (P), to proszę przejrzeć całą strukturę pierwszą, zbierając informację o pozycjach fosforów. Następnie to samo powtórzyć dla drugiej struktury. Na koniec trzeba wyliczyć RMSD wg tak powstałych list (lub wypisać -1). Można zignorować numery nukleotydów, nazwy łańcuchów, itp.

- Wynik będzie sprawdzany z dokładnością do 0.1

```

$ wget http://www.pdb.org/pdb/files/1EHZ.pdb
$ wget http://www.pdb.org/pdb/files/1EVV.pdb
$ ./program 1EHZ.pdb 1EVV.pdb P
52.65766068024471

$ wget http://www.pdb.org/pdb/files/1FFK.pdb
$ ./program 1EHZ.pdb 1FFK.pdb P
-1

```

## 2 INF

1. INF (interaction network fidelity) to miara dedykowana strukturom RNA. Jest to współczynnik podobieństwa listy par zasad w skali od 0 (różne) do 1 (identyczne)
2. W mierze tej jedną strukturę nazywamy referencyjną  $S_r$ , a drugą modelem  $S_m$
3. Wyznaczamy zbiory TP (true positives), FP (false positives) i FN (false negatives):
  - $TP = S_r \cap S_m$
  - $FP = S_m \setminus S_r$
  - $FN = S_r \setminus S_m$
4. Miara INF to średnia geometryczna współczynników PPV (positive predictive value) i STY (sensitivity):
  - $PPV = \frac{|TP|}{|TP| + |FP|}$
  - $STY = \frac{|TP|}{|TP| + |FN|}$
  - $INF = \sqrt{PPV \cdot STY}$
5. Napisz program, który otrzyma na wejściu dwie struktury w formacie BP i wypisze wartość miary INF na wyjściu.

```
target.bp
3 16
4 15
5 14
6 13
9 28
10 27
11 26
```

```
model.bp
3 16
4 15
5 14
8 29
9 28
10 27
11 26
```

```
$ ./program target.bp model.bp
0.8571428571428571
```

### 3 MCQ

1. MCQ (mean of circular quantities) to miara oparta na kątach torsyjnych (obrotów wokół wiązań atomowych)
2. Ściągnij aplikację mcq-cli
3. Przetestuj tryb `global`, w którym miara zwraca po jednej liczbie dla całej struktury: średnią kątową odległość (MCQ), medianę kątowych odległości (MedCQ) oraz RMSD

```
$ ./global -t 1EHZ.pdb -m 1EVV.pdb
MCQ:    0.16 rad, 9.18°
MedCQ:  0.08 rad, 4.52°
RMSD:   1.11348563093234
```

4. Przetestuj tryb `lcs`, w którym algorytm znajduje najdłuższy fragment taki, dla którego miara MCQ jest poniżej zadanego progu (przełącznik `-v`)

```
$ ./lcs -v 5 -t 1EHZ.pdb -m 1EVV.pdb
MCQ value: 4.99°
Number of residues: 35
Coverage: 46.05%
Target name:
First target residue: A.G42
Last target residue: A.A76
Model name:
First model residue: A.G42
Last model residue: A.A76
```

5. Przetestuj tryb `local`, w którym miara szczegółowo analizuje każdy typ kąta torsyjnego dla każdego nukleotydu. W wyniku uruchomienia dostajemy nazwę ścieżki, w której znajdujemy szczegółowe wyniki porównania.

```
$ ./local -t 1EHZ.pdb -m 1EVV.pdb
Results are available in:
/tmp/4a19f2d5-e2c5-4ba1-9c71-a108ac2ecbdb
```

6. Proszę o wykonanie analizy dla modelu z konkursu 13 RNA-Puzzles. Należy ściągnąć strukturę referencyjną `13_0_solution_4XW7_rpr.pdb` oraz dowolny model (najlepiej aby w obrębie grupy każdy wybrał inny). Należy uruchomić `mcq-cli` w trybie `local`. Proszę przesłać do mnie raport zawierający (1) tabelę dla kąta  $P$ , (2) tabelę dla kąta  $\chi$  oraz (3) wykres dla kąta  $\delta$  pokazujący wartość kolumny *Value Difference* dla każdego nukleotydu. Raport powinien również zawierać Państwa komentarz do zaprezentowanych danych (proszę przeanalizować kolumny *Range Difference*, wizualizacje, itd.)

## 4 Przewidywanie struktur 3D RNA

1. Proszę wykorzystać stronę Rosetta FARFAR (RNA De Novo) Protocol do wygenerowania modelu 3D dla danych: (uwaga, może to trwać kilka dni!)

```
>strand_A
gCGGCACCGUCCGCUAAACAAACGG
((((...[[[.]])).....]])
```

2. Dla tych samych danych, skorzystaj z RNAComposera
3. Proszę porównać uzyskane modele ze strukturą referencyjną o PDB id 2A43 i przedstawić wnioski w raporcie.