

Organizacja

Na zajęciach będzie zadane jedno zadanie do wykonania samodzielnie, z którego należy sporządzić raport i wysłać do dnia **17 kwietnia**. Proszę tylko uważać z ostatecznym terminem. Zadanie jest zależne od zewnętrznych narzędzi, publicznie dostępnych i w ciągłym użyciu. Wyniki być może będą praktycznie natychmiast, a być może zadanie utknie w kolejce na wiele godzin, trzeba to wziąć pod uwagę.

Wprowadzenie

Cząsteczki porównujemy żeby wyznaczyć poziom ich podobieństwa (globalnie) lub wskazać na fragmenty najbardziej/najmniej zbliżone (lokalnie). Otrzymane wyniki pozwalają na klastrowanie (zebranie razem cząsteczek o podobnej budowie), odkrywanie motywów, śledzenie zmian i inne.

Nie istnieje jednak pojedyncza, najlepsza metoda porównywania. Sam cel nie zawsze jest jednoznaczny - czy podobne znaczy o zbliżonym rozkładzie atomów w przestrzeni? o podobnym poziomie energetycznym? o podobnym kształcie?

Miary podobieństwa

- **RMSD** (Root Mean Square Deviation)
 - Cząsteczki reprezentowane są jako atomy - punkty w przestrzeni 3D
 - Obliczana jest odległość uśredniona oparta na euklidesowej odległości między odpowiadającymi sobie atomami
 - Przed samym obliczeniem RMSD należy struktury na siebie nałożyć (obrócić i przesunąć) by wartość RMSD była minimalna
- **MCQ** (Mean of Circular Quantities)
 - MCQ reprezentuje cząsteczki jako tablice kątów torsyjnych dla każdej reszty
 - Dzięki zdefiniowaniu różnicy i sumy kątów, możliwe jest wyznaczenie średniej różnicy (MCQ)
 - Odpowiadające sobie reszty w strukturze są wykorzystywane do obliczenia różnic kątów między nimi
 - Średnia z nich może być obliczana dla całych struktur albo dla fragmentów

Porównywanie

- Globalne:
 1. Ściągnij pliki PDB dla struktur 1EHZ, 1EVV, 4TNA, 4TRA i 6TNA
 2. Uruchom aplikację do **RMSD** i **MCQ**
 3. Na zakładce *Global comparison* dodaj wszystkie wspomniane pliki

4. Porównaj je korzystając z RMSD, następnie zwizualizuj i klastruj korzystając z metody k-medoids ($k=2$ oraz $k=3$)
 5. Powtórz to dla metody MCQ. Można zauważyć, że choć międzycząsteczkowe różnice są inne (na wizualizacji jest inne rozłożenie) to klastrowanie przebiega tak samo. Dla większej liczby struktur tak już jednak nie będzie
- Lokalne:
 1. Na zakładce *Torsion local comparison* dodaj struktury 1EHZ oraz 1EVV
 2. Zaznacz, że interesujące są kąty dla *Nucleotides* i dokładniej tylko *MCQ*
 3. Po porównaniu pokaże się wykres. Dla każdej reszty będzie widać ile wynosi średnia różnica kątów
 4. Warto spróbować wygenerować wykresy dla innych kątów i zobaczyć, który najbardziej wpływa na wysokie piki w MCQ. Taka analiza pozwala wykryć nie tylko fragmenty niedopasowane, ale w ich obrębie jeszcze doprecyzować które atomy mogą być inaczej ułożone.

Dopasowanie (alignment)

- Dla RNA:
 - Wykorzystamy struktury [1U6B](#) oraz [1Y0Q](#):
 1. Otwórz strony z podsumowaniem obu struktur, w szczególności sprawdź struktury: pierwszorzędową, drugorzędową, trzeciorzędową
 2. Na stronie ze strukturą pierwszorzędową (sekwencją) zaznacz opcję *FASTA string*
 3. Wejdź na stronę [NeedleN](#), serwera do dopasowywania sekwencji algorytmem Needlemana-Wunscha. Wpisz tam obie sekwencje FASTA jedna pod drugą tzn.:


```
> . . .
              FASTA1
              > . . .
              FASTA2
```
 4. Opcjonalnie zmień *Job name* na inny niż domyślny
 5. Uruchom zadanie przyciskiem *Launch*
 6. Przejdź do strony z wynikami i odnajdź swoje zadanie. Czy **na poziomie sekwencji** można powiedzieć, że cząsteczki 1U6B oraz 1Y0Q są podobne?
 - [ARTS](#) (Alignment of RNA Tertiary Structures) [[Dror2006](#)]
 - * Algorytm skupia się na centralnych atomach fosforu
 - * Traktuje je jako punkty w przestrzeni trójwymiarowej
 - * Heurystycznie rozwiązuje problem *Largest Common Point Set*
 1. Wejdź na stronę serwera [ARTS](#)

2. Wpisz PDB id struktur 1U6B oraz 1Y0Q
 3. Sprawdź wynikową tabelę:
 - Score:** Ocena jakości dopasowania
 - Core Size:** Ile reszt dopasowano
 - BP Core Size:** Ile wśród nich jest nukleotydów sparowanych
 - P-Value:** Prawdopodobieństwo, że istnieje co najmniej jedna losowa para fragmentów takiej samej długości, taka że ocena podobieństwa ich dopasowania będzie wyższa (w uproszczeniu: statystyczna istotność wyniku)
 4. Dla dowolnego wyniku sprawdź podstronę z linku *BP Core Size* albo *Core Size*. Czy zgodne są dopasowane nukleotydy? Jaki wniosek można wyciągnąć z tego wyniku?
 5. Na stronie z wynikami można też obejrzeć konkretne dopasowania. Warto sprawdzić najlepsze i najgorsze ze znalezionych rozwiązań
- **DIAL** (Dihedral Alignment) [Ferre2007]
1. Metoda opiera się na trzech kryteriach zgodności: sekwencja, parowanie, kąty torsyjne.
 2. Warto sprawdzić stronę projektu oraz zlecić zadanie dla tych samych PDB id co poprzednio

- Dla białek:

Zadanie 1

1. Wyszukaj w NCBI w bazie *protein* sekwencji białka BcnI
2. Zleć wykonanie automatycznego przewidywania struktury trzeciorzędowej na podstawie sekwencji. Wykorzystaj dowolny publicznie dostępny serwer np. [SWISS-MODEL](#) albo [CPHmodels](#)
3. Ściągnij też wzorcową strukturę tego białka zdeponowaną jako PDB [2ODH](#)
4. Wykorzystaj dowolne 2 z 3 poniżej opisanych narzędzi do porównania tych dwóch struktur. W raporcie opisz krótko przebieg ćwiczenia oraz wyciągnięte wnioski.

– **DALI** (Distance Matrix Alignment) [Holm1993]

- * Białka reprezentowane są jako macierze odległości między atomami C^α
- * Z macierzy ekstrahuje się wzorce połączeń i metodą Monte Carlo uogólnia na dopasowanie globalne struktur

– **CE** (Combinatorial Extension) [Shindyalov1998]

- * W białkach poszukiwane są ciągłe fragmenty dobrze dopasowanych regionów
- * Algorytm ma za zadanie połączyć te regiony zgodnie z pewną funkcją oceny (kara za wprowadzenie luki, itd.)
- * Uwaga! Narzędzie umożliwiające podanie własnych plików PDB (czyli tego wygenerowanego przez serwer przewidujący strukturę) [jFATCAT](#)

- [LGA \(Local Global Alignment\)](#) [[Zemla2003](#)]
 - * Metoda zwraca Longest Continuous Segments (LCS) oraz Global Distance Test (GDT)
 - * LCS to największy wspólny segment (przy założeniu pewnego maksymalnego dopuszczalnego błędu)
 - * GDT to zbiór wyników, które obrazują przy jakim progu dopuszczalnego błędu ile reszt się dopasowuje aż do osiągnięcia poziomu, na którym struktury nakładają się w całości

[Dror2006] Dror, O., N.R.W.H. The ARTS web server for aligning RNA tertiary structures *Nucleic Acids Research*, 2006, Vol. 34, pp. W412-W415

[Ferre2007] Ferre, F., Ponty, Y., Lorenz, W.A. & Clote, P. DIAL: a web server for the pairwise alignment of two RNA three- dimensional structures using nucleotide, dihedral angle and base-pairing similarities *Nucleic Acids Research*, 2007, Vol. 35(2), pp. W659-W668

[Holm1993] Holm, L., S.C. Protein structure comparison by alignment of distance matrices *Journal of Molecular Biology*, 1993, Vol. 233(1), pp. 123-138

[Shindyalov1998] Shindyalov, I.N., B.P. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path *Protein Engineering*, 1998, Vol. 11(9), pp. 739-747

[Zemla2003] Zemla, A. LGA: A method for finding 3D similarities in protein structures *Nucleic Acids Research*, 2003, Vol. 31(13), pp. 3370-3374