Instytut Informatyki i Matematyki Komputerowej UJ,
opracowanie: mgr Ewa Matczyńska, dr Jacek Śmietański

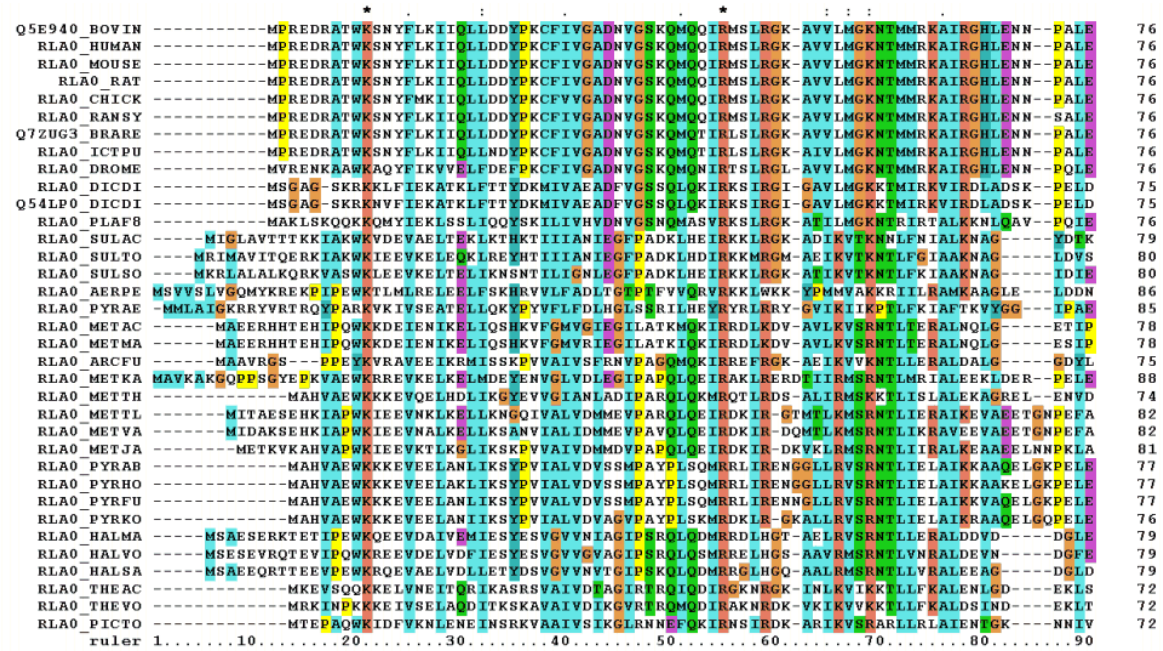
MSA

1. Dopasowania wielosekwencyjne - wprowadzenie

Dopasowanie wielosekwencyjne (MSA - *multiple sequence alignment*) pozwala na porównanie większej liczby sekwencji w celu wyszukania zależności ewolucyjnych między nimi.

Możemy dzięki temu uzyskać informacje o:
- homologii sekwencji, czyli o pochodzeniu od wspólnego przodka
- konserwatywności sekwencji: sekwencje konserwatywne wyróżniają się bardzo dużym podobieństwem lub wręcz identycznością wśród różnych organizmów, ponieważ mutacje, które zdarzają się w rejonach konserwatywnych sekwencji powodują dużo gorsze przystosowanie do środowiska. W konsekwencji, pod działaniem doboru naturalnego nie utrzymują się w populacji.

Informacja o konserwatywnych fragmentach MSA pomaga wnioskować o strukturze 3D białka.



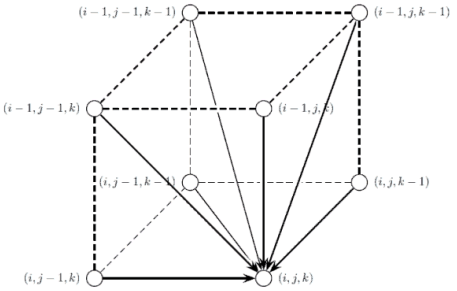
Rysunek 1: Dopasowanie wielosekwencyjne genu białka rybosomowego L1OE (pierwsze 90 pozycji, Clustal)

2. Dopasowanie wielosekwencyjne – podejścia

A) dopasowanie za pomocą programowania dynamicznego

Jeśli mielibyśmy zestaw sekwencji, które podejrzewamy o ewolucyjne podobieństwo moglibyśmy próbować zastosować rozszerzenie algorytmu dynamicznego dla dopasowania globalnego. Na przykład dla dopasowania trzech sekwencji moglibyśmy używać trójwymiarowej tablicy i przemieszczać się po niej w następujący sposób:

$$s_{i,j,k} = \max \begin{cases} s_{i-1,j,k} & +\delta(v_i, -, -) \\ s_{i,j-1,k} & +\delta(-, w_j, -) \\ s_{i,j,k-1} & +\delta(-, -, u_k) \\ s_{i-1,j-1,k} & +\delta(v_i, w_j, -) \\ s_{i-1,j,k-1} & +\delta(v_i, -, u_k) \\ s_{i,j-1,k-1} & +\delta(-, w_j, u_k) \\ s_{i-1,j-1,k-1} & +\delta(v_i, w_j, u_k) \end{cases}$$



Nietrudno zauważyć, że złożoność zwiększałaby się ekspotencjalnie wraz z liczbą dopasowywanych sekwencji, jest to więc podejście dość mało efektywne.

B) dopasowanie progresywne

Aby przyspieszyć czas dopasowania poszukiwano pewnego uproszczenia względem programowania dynamicznego. Pomysł jest taki, aby użyć dopasowania pary najbardziej podobnych sekwencji jako bazy – punktu startowego, ponieważ dopasowanie sekwencji najbliżej spokrewnionych będzie najprawdopodobniej najlepiej reprezentować optymalne dopasowanie wielosekwencyjne. Potem iteracyjnie dokładając kolejne najbardziej podobne sekwencje, uzyskuje się końcowe dopasowanie. W ten sposób działa najbardziej popularny program do wyznaczania dopasowań wielosekwencyjnych - Clustal.

Aby rozpocząć procedurę progresywnego dopasowania musimy najpierw dla każdej pary sekwencji wyznaczyć jak bardzo są podobne, aby wiedzieć od których rozpocząć konstrukcję MSA. W tym celu oblicza się dopasowanie sekwencji, a następnie przyjmuje się jakąś miarę odległości między nimi. Najprościej można np. pominąć wszystkie pozycje na których wystąpiły indele i obliczyć odsetek pozycji na których aminokwasy, bądź nukleotydy się różniły. Na podstawie tej odległości tworzy się tzw. drzewo przewodnie (*guide tree*), zgodnie z którym progresywnie tworzy się dopasowanie. Drzewo przewodnie określa nam kolejność według której dołączamy kolejne sekwencje do dopasowania MSA typu progresywnego.

Drzewo przewodnie tworzone jest zwykle metodami:
- metodą średnich połączeń (UPGMA);
- metodą łaczenia sąsiadów (*neighbour joining*).

Obie metody wykorzystują macierz odległości sekwencji. Niech:
d(x, y) – odległość sekwencji x i y.

Metoda **UPGMA** (*Unweighted Pair Group Method with Arithmetic Mean*) - to poprostu klastrowanie hierarchiczne *bottom up* z obliczaniem odległości klastrów jako średniej odległości wszystkich punktów należących do klastrów *(average linkage hierarchical clustering*):

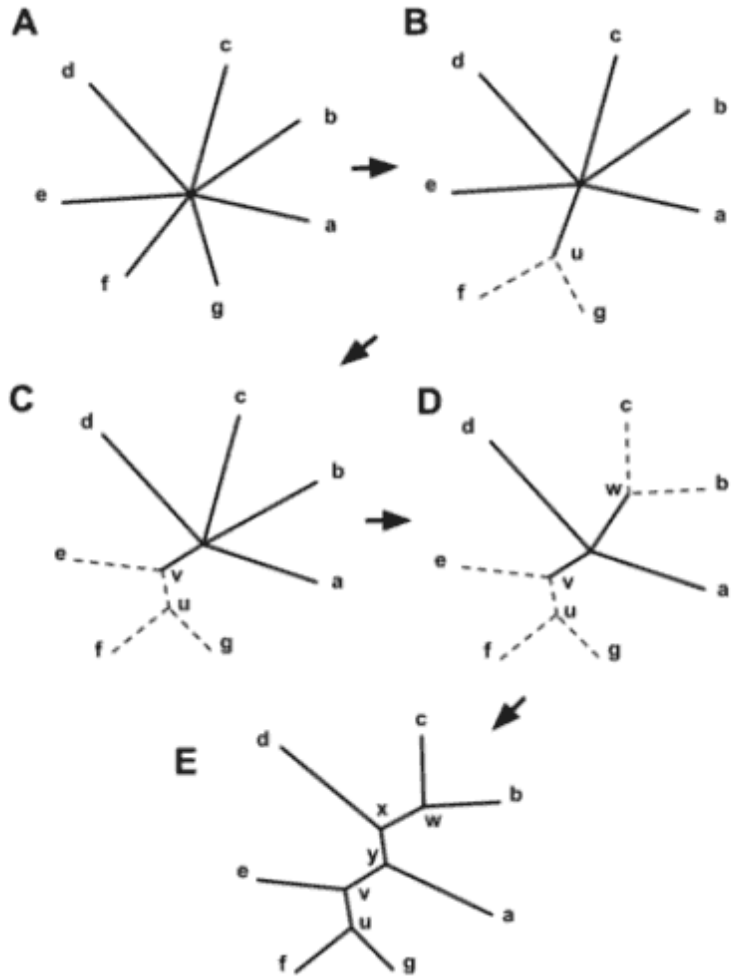
$$\frac{1}{|\mathcal{A}| \cdot |\mathcal{B}|} \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{B}} d(x, y)$$

Metoda ***neighbour joining*** to algorytm typu zachłannego, który dąży do zbudowania drzewa o takiej topologii, która minimalizowałaby *tree length* - określony jako średnia ważona odległości między sekwencjami z wagami określonymi przez topologię drzewa.
W każdym kroku przekształcamy macierz d odległości między sekwencjami na macierz Q, określoną jako:

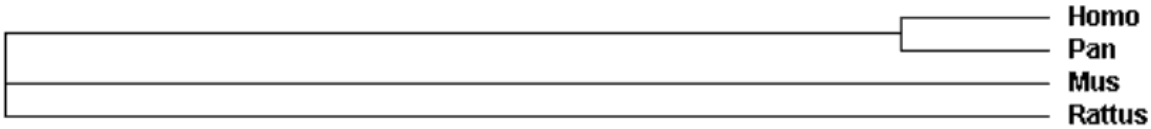
$$Q(i,j) = (r-2)d(i,j) - \sum_{k=1}^{r} d(i,k) - \sum_{k=1}^{r} d(j,k)$$

Wybieramy do połączenia w jeden węzeł te dwie sekwencje, dla których wartość *Q* była najniższa. Można to interpretować tak, że chcemy łączyć te sekwencje, które są bliskie tzn. odległość *d* jest mała, jednocześnie są daleko od wszystkich innych sekwencji, za co odpowiadają odejmowane sumy po wszystkich wierzchołkach we wzorze powyżej.

Następnie wyznaczamy odległości *d* nowego wierzchołka, który jest połączeniem dwóch wybranych sekwencji, do innych wierzchołków i znowu wyznaczamy macierz *Q*.



Używając jednej z powyższych metod otrzymujemy drzewo przewodnie – *guide tree*, które w przybliżeniu obrazuje zależności ewolucyjne między sekwencjami.



Dopasowanie sekwencji do dopasowania (zbioru już dopasowanych sekwencji we wcześniejszych krokach), bądź dopasowanie dopasowań do siebie przebiega w analogiczny, dynamiczny sposób, tak jak dopasowanie globalne sekwencji. Załóżmy, że w naszym dopasowaniu w pierwszej grupie sekwencji już dopasowanych w danej kolumnie mamy aminokwasy Y i P, a w drugiej grupie sekwencji aminokwasy F i Y. Jak dobrze dopasowane są te kolumny obliczamy np. w ten sposób:

$$(\delta(Y;F)+\delta(Y;Y)+\delta(P;F)+\delta(P;Y))/4$$

Czyli obliczamy dopasowania „każdy z każdym" z obydwu grup i uśredniamy. Przerwy wstawiane są analogicznie jak dla przypadku dwóch sekwencji. Zwróćmy jednak uwagę, że jeśli w dopasowaniu pojawia się przerwa, to zostaje ona wstawiona w każdą sekwencję składającą się na to dopasowanie, w myśl zasady „*once a gap, always a gap*". W związku z tym widać, że generowane dopasowania wielosekwencyjne niekoniecznie będą optymalne:

```
A-VKND
AMEKAD
```

```
A-VK-ND
AMEK-AD
TVEKTAD
```

Więc aby uzyskać rezultat biologicznie sensowny czasami trzeba ręcznie poprawiać dopasowania wielosekwencyjne.

Zadanie (4pkt)

Rozwiązanie zadania prześlij mailem do wtorku, **26.11.2019** włącznie, na adres:
jacek.smietanski@ii.uj.edu.pl
Temat wiadomości proszę opatrzyć przedrostkiem **[Bio] Lab 07**. Rozwiązaniem ma być **tylko jeden plik** – dokument PDF. Proszę o nazwanie pliku wg schematu**: Imie.Nazwisko.07.pdf**.

Przeprowadź dopasowania wielosekwencyjne dla zadanego zbioru danych metodami Clustal, Muscle, T-Coffee oraz MAFFT. Zidentyfikuj różnice pomiędzy zestawieniami.

1. Dane sekwencyjne: gen MHC.

W pliku **mhc.fa** znajdują się sekwencje aminokwasowe dla jednego z białek MHC, czyli białek zgodności tkankowej.
Geny MHC kodują białka lokalizowane na powierzchni komórek, które są odpowiedzialne za wykrywanie niebezpieczeństw grożących naszemu organizmowi (np. wirusów). Dlatego też, przed przeszczepami sprawdza się ich zgodność w komórkach dawcy i biorcy, aby szanse przyjęcia się przeszczepu były większe.

We wskazanym pliku (mhc.fa) znajdują się sekwencje tego samego białka, pochodzące od różnych organizmów: człowieka szympansa, szczura i myszy.

Metody (narzędzia on-line):
- Clustal Omega: https://www.ebi.ac.uk/Tools/msa/clustalo/
- Muscle: https://www.ebi.ac.uk/Tools/msa/muscle/
- T-Coffee: https://tcoffee.crg.cat/
- MAFFT: https://www.ebi.ac.uk/Tools/msa/mafft/

a) Przeklej zawartość pliku do głównego okienka (proszę w tytule sekwencji zostawić tylko nazwę organizmu, wynik będzie wtedy bardziej czytelny)

b) Zobacz wyniki dopasowania, zidentyfikuj i wypisz kilka pozycji konserwatywnych, kliknij na opcję ***show colors*** (Muscle, Clustal), aby zobaczyć czy aminokwasy w obrębie kolumn mają podobne właściwości

c) Kliknij u góry na ***Result Summary***, zobacz macierz ***Percent Identity Matrix*** otrzymaną przez dopasowanie par sekwencji.
Zwróć uwagę, które pary sekwencji mają najwyższą punktację. Czy jest to zgodnie z oczekiwaniami?

d) W tym samym widoku wybierz ***Jalview*** - program do wizualizacji dopasowania.
Obejrzyj dopasowanie jeszcze raz, zwróć uwagę na pozycje konserwatywne oraz sekwencję konsensusową (wyświetlona na dole okna). Co oznaczają plusy w sekwencji konsensusowej?

e) Z menu ***Colour*** wybierz ***Hydrophobicty***, aby zobaczyć kolorowanie dla hydrofobowości aminokwasów:
- czerwony: hydrofobowy – grupa aminokwasów o ogonie węglowodorowym, która nie lubi przebywać w otoczeniu wody;
- niebieski: hydrofilowy – grupa aminokwasów o ogonie, który posiada ładunek - częściowy lub całkowity, która lubi przebywać w otoczeniu wody.
Zidentyfikuj i wypisz kilka pozycji gdzie zaszła zmiana hydrofobowości.

f) Wróć do wyników dopasowania. Obejrzyj drzewo przewodnie użyte do konstrukcji tego dopasowania (zakładka ***Guide Tree***)

2. Rodopsyna.

W pliku **rhodopsin.fa** znajdziesz sekwencje białka rodopsyny, czyli światłoczułego barwnika występującego w siatkówce oka (kolejno: człowiek, szympans, pies, byk, mysz, szczur, kurczak, rybka danio). Prześledź kolejne punkty jak w punkcie 1.

3. Który z tych dwóch genów jest bardziej konserwatywny? jak myślisz dlaczego ?

3. MSA w biopythonie

Biopython wspiera korzystanie z narzędzi ClustalW oraz Muscle tworząc interfejs dla zapytania wywoływanego przez commandline – aby zadziałało, trzeba mieć zainstalowany odpowiedni program (Clustal lub Muscle) lokalnie.

ClustalW2:

```
>>> from Bio.Align.Applications import ClustalwCommandline
>>> help(ClustalwCommandline)
...

>>> cline = ClustalwCommandline("clustalw2", infile="opuntia.fasta")
>>> print(cline)
clustalw2 -infile=opuntia.fasta
```

Muscle:

```
>>> from Bio.Align.Applications import MuscleCommandline
>>> help(MuscleCommandline)
...

>>> cline = MuscleCommandline(input="opuntia.fasta", out="opuntia.txt")
>>> print(cline)
muscle -in opuntia.fasta -out opuntia.txt
```