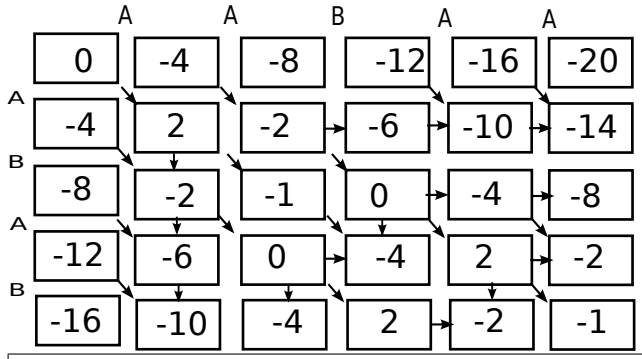
Metody Bioinformatyki, część informatyczna, 19 czerwca 2020 czas: 60 min

- Proszę nie edytować tego pliku. Tutaj są tylko treści zadań.
- Odpowiedzi proszę umieszczać w pliku o nazwie 'mbi-nazwisko.txt'. Proszę pobrać ten plik, zmienić mu nazwę wpisując własne nazwisko bez użycia polskich znaków. Dodatkowo proszę wpisać swoje imię i nazwisko w pierwszej linii pliku.
- Na koniec proszę pobrać i wypełnić oświadczenie (plik 'oswiadczenie-nazwisko.doc').
- Egzamin składa się z 5 zadań. Treść każdego zadania jest na oddzielnej stronie.

Zadanie 1 (8 pkt)

Poniżej przedstawiono macierz dla algorytmu, który dostarcza uliniowienia globalnego (algorytm Needlemana-Wunscha) dwóch sekwencji. Macierz podobieństwa pokazano obok. Stosujemy liniową karę za przerwę, $\gamma(n) = n * d, d = -4$.

	A	**B**
A	2	-3
B	-3	2

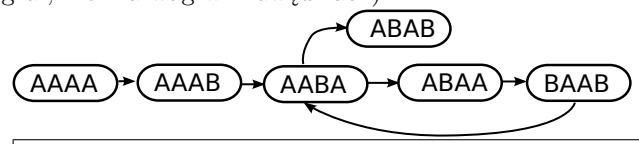


Pytania:

- algorytm uruchomiono dla sekwencji AABAA i ...
- rozwiązań optymalnych jest ...
- rozwiązania optymalne to ...

Zadanie 2 (8 pkt)

Dla zestawu odczytów zbudowano graf de Brujna 5 rzędu pokazany na rysunku (nie jest to multi-graf, ani A-graf, nie ma wag w krawędziach).



Pytania:

- Czy istnieje rozwiązanie (ścieżka Eulera)?
- Jeżeli tak, podaj sekwencję odpowiadającą jednemu z rozwiązań:
- Zbuduj graf de Bruijna 4 rzędu (bez wag w krawędziach). Ile ten graf ma wierzchołków:
- Czy istneje (dla grafu de Bruijna 4 rzędu) rozwiązanie (ścieżka Eulera)?
- Jeżeli tak, to podaj sekwencję odpowiadającą jednemu z rozwiązań:

Zadanie 3 (8 pkt)

Cząsteczka jest biopolimerem i może być reprezentowana przez napis nad alfabetem { A, B }. Obserwujemy oddziaływanie poszczególnych symboli z naszą sondą i oznaczamy je jako '+' lub '-'. Chcemy opisać naszą cząsteczkę oraz oddziaływanie za pomocą ukrytego modelu Markowa. Stanem będzie symbol, więc $Q = \{A, B\}$, obserwacją oddziaływanie, więc $V = \{+, -\}$. Dla cząsteczki $AAAAAAAAAA$ obserwujemy oddziaływanie $+-++++----$, dla cząsteczki $BBBB$ obserwujemy $+-++$. Nasze sekwencje zaczynają się od z takim samym prawdopodobieństwem od A i od B, więc $P_A = \frac{1}{2}, P_B = \frac{1}{2}$. Znamy macierz przejść:

	A	**B**
A	$\frac{4}{5}$	$\frac{1}{5}$
B	$\frac{1}{5}$	$\frac{4}{5}$

Pytania:

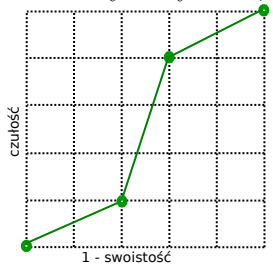
- podaj macierz emisji:

	+	-
A		
B		

- podaj najbardziej prawdopodobną sekwencję cząsteczki dla oddziaływania $+++$:

Zadanie 4 (8 pkt)

Opracowano test binarny X, dla którego krzywa ROC jest pokazana na rysunku. Test zwraca wynik 'tak' lub 'nie'. Używamy testu, gdzie obiekty mają w 50% stan 'tak' i w 50% stan 'nie'.



Pytania:

- czy test może zwracać zawsze wynik 'nie' ?
- jaka jest wtedy czułość?
- jaka jest czułość dla swoistości 0.4?
- Jaki będzie błąd przy czułości 0.2?
- Jaki będzie błąd przy czułości 0.8?

Zadanie 5 (8 pkt)

Mamy dane pokazane poniżej i zakładamy pierwszą ramkę odczytu (pierwszy kodon odpowiada pierwszemu aminokwasowi).

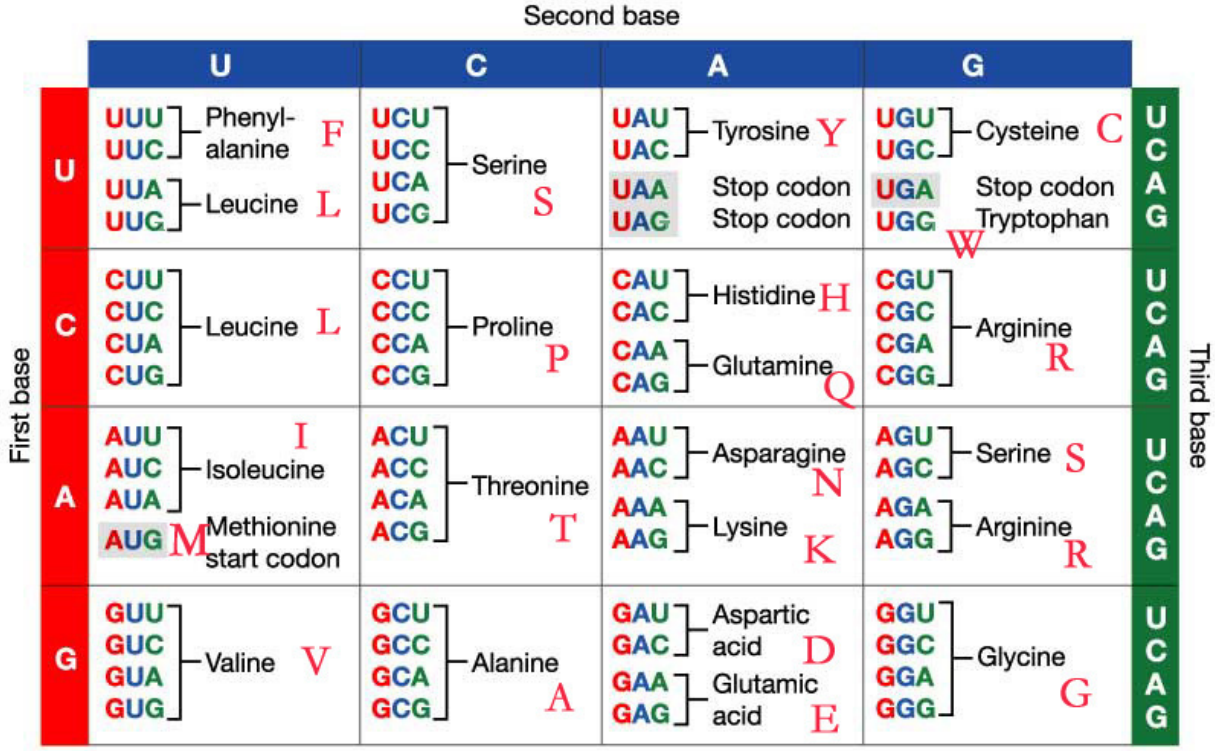
a) fragment sekwencji kodującej genu CLP1 w genomie referencyjnym:

```
>hg38_dna range=chr11:57659477-57659493 5'pad=0 3'pad=0 strand=+
ATGGGAGAAGAGGCT
```

b) sekwencję pacjenta:

```
> patient1
ATGGGAGTAGAGGCC
```

Tablica kodowania aminokwasów:



Polecenia:

- Dokonaj translacji sekwencji referencyjnej DNA na sekwencję aminokwasów
- Dokonaj translacji sekwencji DNA pacjenta na sekwencję aminokwasów
- Znajdź warianty SNV na poziomie sekwencji DNA i zapisz je w formacie:
  - `CHR:POS_REF>ALT, np: chr11:57659477_A>T`
- Określ efekt każdego wariantu jak będzie on miał dla sekwencji białkowej, nadając im jedną z 4 etykiet: 'missense', 'synonymous', 'stopgain', 'stoploss'