Bioinformatyka — laboratorium 3

Czas na oddanie finalnego raportu: 6 tygodni

Nazwa pliku: imie_nazwisko_3_bio.pdf

Typ ćwiczenia: jednotygodniowe

Cel: Celem ćwiczenia jest zaznajomienie studenta z podstawami genetyki poprzez realizację zadań związanych z genealogią.

Podstawy biologii dla bioinformatyków — zadania

1. Zapoznaj się z opisem na guides.loc.gov i wymień, jakie są 3 metody używane w genealogii [**task 1**].

Format FASTA to sposób zapisu sekwencji nukleotydów lub aminokwasów, który jest często używany w bioinformatyce. Nagłówek (header): rozpoczyna się od znaku ">" i zawiera informacje o sekwencji, takie jak jej nazwa, identyfikator, źródło lub opis. Nagłówek może zajmować kilka linii, ale każda linia powinna zaczynać się od znaku ">". Sekwencja: rozpoczyna się od pierwszej linii, która nie jest nagłówkiem i zawiera sekwencję nukleotydów lub aminokwasów, zapisaną w jednej lub kilku linii. Linie z sekwencją powinny zawierać tylko litery reprezentujące nukleotydy lub aminokwasy, bez dodatkowych znaków lub spacji.

2. Pobierz cały genom mitochondrialny w formacie FASTA z www.phylotree.org: `Amati-Bonneau et al. 2008 publication 1` (kliknij –> download FASTA)

3. Czy genom ten zawiera jakieś nukleotydy oznaczone jako "N" [**task 2**]? Co oznacza "N" i co by to oznaczało gdyby plik zawierał tak oznaczone nukleotydy [**task 3**] (wnioskowanie możesz oprzeć na tym materiale: www.hgmd.cf.ac.uk)

4. Czy genom zawiera jakieś inne oznaczenia niż A lub C lub T lub G, o czym to świadczy [**task 4**]? Gdyby w sekwencji genomu znalazło się oznaczenie "Y" - co to konretnie oznacza [**task 5**]?

Haplogrupa, w bardzo dużym uproszczeniu, to rodzaj mapy genetycznej, która pokazuje pochodzenie naszych przodków i naszego dziedzictwa genetycznego. Jest to również grupa ludzi, którzy dziedziczą ten sam zestaw genów (lub alleli czyli wersji genów) od swoich przodków i są ze sobą powiązani. Haplogrupy są używane do badania naszej historii i pochodzenia, a także do wyjaśniania różnic w naszym zdrowiu.

Na przykład wyobraźmy sobie nieistniejącą haplogrupę A22 – i to, że w mtDNA reprezentowana jest w pozycji 24 nukleotydu przez allele AA, 126 TT, 373 C lub G, 1200 TT, 12001 TA. Te allele w tych pozycjach mogą być nazywane markerami genetycznymi tej haplogrupy. Po obecności ich w tych pozycjach możemy rozpoznać osobę należącą do tej haplogrupy.

5. Przeprowadź analizę uprzednio pobranego genomu w narzędziu dna.jameslick.com (Choose file -> Upload).

    - Jakie kompletne regiony mtDNA zawiera twój plik (czy są to regiony HVR? CR? Czy inne?) [**task 6**]? Podpowiedź możesz też znaleźć tutaj: help.familytreedna.com.

- Jakie markery genetyczne określił algorytm (Markers found) [**task 7**]. Co to jest marker genetyczny [**task 8**]?
- Jaką haplogrupę wykazał algorytm [**task 9**]?
- Co to jest rCRS (zobacz FAQ rozpisany przy tym narzędziu) [**task 10**]? Jaką haplogrupę ma rCRS [**task 11**]?

6. Za pomocą narzędzia predict.yseq.net (Choose file -> wybierz plik w formacie FASTA, który już posiadasz na dysku) znajdź rekord z bazy danych NCBI (NCBI to The National Center for Biotechnology Information czyli organizacja, która utrzymuje m.in. wiele biologicznych baz danych), który wykazuje najwyższe podobieństwo do twoich danych, podaj jego numer, oraz narodowość naukowców, którzy prowadzili te badania [**task 12**].

7. Za pomocą narzędzia **mtDNA nomenclature tool** (mtprofiler.yonsei.ac.kr) przeprowadź porównanie między twoimi danymi a rCRS. Na tej podstawie podaj jaka mutacja występuje między pozycjami w mtDNA 12120…..12130 i co ona oznacza [**task 13**].

8. Za pomocą narzędzia phylogeographer.com wygeneruj mapę frekwencji haplogrupy, którą posiada analizowana sekwencja. Zrób screena jako dowód wykonania tego zadania [**task 14**].

Y-DNA jest rodzajem DNA, który znajduje się na chromosomie Y i jest przekazywany z ojca na syna. Jest to jedyny rodzaj DNA, który jest przekazywany w linii męskiej. Y-DNA jest bardzo przydatne w badaniach genealogicznych, ponieważ pozwala nam śledzić linie męskiej rodziny i zrozumieć nasze pochodzenie. W ramach badań genealogicznych można porównać Y-DNA różnych osób, aby zobaczyć, czy istnieją pokrewieństwa i w jakim stopniu. Jeśli dwie osoby mają podobne Y-DNA, może to oznaczać, że są one spokrewnione w linii męskiej. Im więcej podobieństw w Y-DNA, tym bliższe jest pokrewieństwo. Markerem genetycznym może być również liczba powtórzeń jakiejś wybranej sekwencji DNA, np. AATC (w określonym miejscu!). Wyobraźmy sobie, że w spotykamy osobnika, który na chromosomie Y (jest tylko jeden taki chromosom w genomie więc allel jest tylko jeden) posiada allel AATCAATCAATC czyli 3 powtórzenia. Inny osobnik może mieć 5 powtórzeń: AATCAATCAATCAATCAATC. Jeśli w analizie weźmiemy pod uwagę kilka takich markerów, to możemy się sporo dowiedzieć na temat tego, czy dwa osobniki mogą być ze sobą spokrewnione. Ogólna zasada dla genealogii mówi o tym, że im więcej tych samych alleli dzieli dwóch osobników to tym większe prawdopodobieństwo, że wspólny przodek był między nimi gdzieś „bliżej" niż „dalej".

9. Uruchom narzędzie www.nevgen.org służące do analizy Y-DNA i w bocznym menu zaznacz "Order of 17 markers".

   - Co to jest np. DYS456 [**task 15**]?
   - Jaka sekwencja ulega powtórzeniu w tym markerze [**task 16**]?
   - Ile alleli tego markera jest znanych [**task 17**]?
   - Jaką analizę można wykonać z wykorzystaniem tego narzędzia [**task 18**]?

10. Wejdź na yhrd.org, a następnie:

    - wyjaśnij pokrótce co to za baza [**task 19**],
    - jakie zestawy danych (`datasets`) są publikowane w tej bazie (Database statistics) [**task 20**],
    - jakie allele można zaobserwować dla markera DYS19 [**task 21**],

- ile haplotypów Y27 zawiera ta baza dla próbek pochodzących od Aleutów i w którym roku wykonano ostatnie badanie tej populacji opublikowane w bazie yhrd.org [**task 22**]

11. Z podstrony yhrd.org przenieś do pliku `xlsx` tabelę z haplotypami Y12 (kliknij właściwy przycisk), a następnie:
    - wykryj za pomocą yhrd.org/pages/tools/validator błędy w pliku i napraw je – jako odpowiedź na to zadanie wklej do raportu poprawioną tabelę [**task 23**],
    - za pomocą narzędzia yhrd.org/search wydedukuj, czy do tych analiz użyto zestawu (kitu) www.promega.com czy innego? Jeśli innego podaj jego nazwę [**task 24**].
12. Za pomocą narzędzia www.moseswalker.com policz, jakie jest prawdopodobieństwo posiadania wspólnego przodka przez dwóch osobników różniących się 5 markerami STR w obrębie chromosomu Y spośród 30 markerów badanych w 18 pokoleniu licząc od aktualnego [**task 25**]

Dla chętnych

Zadanie 1 [**task26**] Napisz program w Pythonie, który porównuje dwie sekwencje DNA (`A` i `B`) o jednakowej długości. W przypadku różnych długości sekwencji wejściowych program wyświetla komunikat `Sekwencje różnej długości`. Wynik działania programu dla sekwencji o jednakowych długościach to lista krotek zawierających trzy elementy: pozycję w sekwencji, na której nukleotydy nie są takie same, nukleotyd na wskazanej pozycji z sekwencji `A`, nukleotyd na wskazanej pozycji z sekwencji `B`. Sposób wprowadzania danych do programu jest dowolny.

Przykładowe dane wejściowe

```
Podaj sekwencję DNA 1: AACTTCAGCTTGCGACGTGGGTGCA
Podaj sekwencję DNA 2: AACTTCAGCATGCGACCTGGGAGCA
```

Wynik działania programu:

```
[(10, 'T', 'A'), (17, 'G', 'C'), (22, 'T', 'A')]
```

Przykładowe dane wejściowe:

```
Podaj sekwencję DNA 1: AACTTCAGCTTGCGACGTGGGTGCA
Podaj sekwencję DNA 2: AACTTCAGCATGCGACCTGGGAGC
```

Przykładowy wynik:

```
Sekwencje są różnej długości
```

Zadanie 2 (***) [**task27**] Napisz w Pythonie program, który dla zadanej sekwencji DNA obliczy maksymalną ilość komplementarnych par w strukturze "hairpin" (bez uwzględniania tak zwanego "head").

Przykłady:

Sekwencja: `ACTTGAGTAGCTAT` Wynik działania programu: `3`

Podpowiedź:

```
******ACTT\
      ||| |
TATCGATGAG/
```

Sekwencja: `ACCATAAACGATGACTTTATGACA` Wynik działania programu: `4`

Podpowiedź:

```
**ACCATAAACGA\
    | |||   ||
ACAGTATTTCAGT/
```

W tym przypadku nie jest to jedyne poprawne rozwiązanie.