Cześć!

W związku ze zbliżającym się terminem Biohack Hackathon Bioinformatyczny stworzyliśmy dla Was specjalne repozytorium GitHub:

Link do repozytorium: *https://github.com/BioHack2018*

Jeżeli jeszcze nie masz konta GitHub, załóż je już dziś na stronie: *https://github.com/*

Będziecie pracować w obszarze czterych głównych dziedzin, w ramach których możecie tworzyć również swoje koncepcje i problemy.

I. Analizy genetyczne bakteriofagów:

- Bakteriofagi są coraz częściej rozpatrywane jako alternatywa dla antybiotyków,
- Dalej pozostaje wiele niewiadomych w tej dziedzinie,
- Istnieje wiele praktycznych problemów związanych z ich wykorzystaniem w medycynie,

W tym temacie możesz zająć się problemami takimi jak:

- Analiza lityczności i lizogenności fagów,
- Planowanie mieszanek terapeutycznych,
- Klasyfikacja i analiza nieznanych genów fagowych,

II. Organizacja i przetwarzanie danych NGS

- Sekwencjonowanie oraz analizy danych NGS generują wielkie ilości danych, które trzeba przechowywać i przetwarzać w uporządkowany sposób,
- Nowe narzędzia wspierające zarządzanie projektami i analizami NGS są niezwykle potrzebne zarówno w nauce jak i biznesie,

W tym temacie możesz zająć się problemami takimi jak:

- Narzędzie do tworzenia projektów analiz,
- Narzędzia wspierające przechowywanie i archiwizację danych NGS,
- Narzędzia wspierające projekty na podstawie danych z sekwencjonowania Oxford Nanopore,

III. Narzędzia do analiz genomów prokariotycznych:

- Istnieje wiele możliwości analiz funkcjonalnych jak i filogenetycznych możliwych do wykonania na genomach prokariotycznych,
- Wytworzenie lokalnego narzędzia do tego typu analiz może ułatwić i przyśpieszyć pracę naukowców,
- Przyjazne dla użytkownika rozwiązania poprawią dostępność tego typu analiz dla przeciętnego użytkownika,

W tym temacie możesz zająć się problemami takimi jak:

- Narzędzie charakteryzujące genomy na podstawie baz COG, Pfam itp.
- Narzędzia wspierające analizy porównawcze gatunków np. w oparciu o ANI, ortoANI, GGD,

IV. Przewidywanie koloru oczu w oparciu o dane genetyczne:

- Analizy wykorzystujące przewidywanie fenotypu na podstawie genotypu znajdują wykorzystanie w kryminalistyce oraz innych dziedzinach,
- Obecne modele predykcyjne typu IrisPlex są niedoskonałe i mogą zostać poprawione,
- Wykorzystanie metod uczenia maszynowego w analizach genetycznych znajduje coraz szersze zastosowanie i budowane są dzięki nim coraz lepsze narzędzia,

W tym temacie możesz zająć się problemami takimi jak:

- Wykorzystanie metod uczenia maszynowego do przewidywania koloru oczu na podstawie danych genetycznych,

Jeżeli wraz z zespołem macie ciekawą koncepcję, nad którą chcecie pracować - śmiało możecie zgłosić swój temat.

Jeżeli potrzebujecie inspiracji, przedstawiamy zadania proponowane przez Biobank:

1. Analiza genomów bakteriofagowych pod kątem ich lityczności/lizogenności

Celem zadania będzie opracowanie ścieżki analitycznej w oparciu o oprogramowanie open source oraz samodzielnie przygotowane skrypty pozwalającej na maksymalnie dokładne

zidentyfikowanie bakteriofaga jako litycznego bądź lizogennego w oparciu o sekwencję jego DNA. Ważnym etapem będzie także zautomatyzowanie procesu analiz np. w oparciu o framework Luigi.

2. Implementacja algorytmu „balance tree" dla analizy różnic obfitości w badaniach metagenomicznych

Analiza różnorodności taksonów znajdujących się w różnych próbkach jest istotnym problemem w badaniach metagenomicznych. Obecnie najpopularniejszą metodą analiz tego typu jest ANCOM (ANalysis of Composition Of Microbiomes) która posiada wiele dobrych implementacji. Z drugiej jednak strony wykazano, że lepszą metodą do analiz różnic taksonomicznych jest metoda „balance tree" zapożyczona z geologii. Celem zadania jest odpowiednie dostosowanie oraz implementacja metody „balance tree" dla analizy różnic obfitości w badaniach metagenomicznych.

3. Analiza danych GWAS w oparciu o algorytmy uczenia maszynowego.

Celem zadania będzie przygotowanie ścieżki analitycznej wykorzystującej metody uczenia maszynowego pozwalającej na wyszukanie interesujących wzorców w dostarczonym zestawie danych z analiz GWAS. Uczestnicy poszukiwać będą zależności pomiędzy cechami fenotypowymi takimi jak np. kolor oczu, kolor włosów, różnego rodzaju choroby a ok. 500 000 wariantów genetycznych badanych na mikromacierzach. Możliwe do przeprowadzenia analizy obejmują także badania powiązania występowania poszczególnych wariantów z danymi geograficznymi dotyczącymi osób od których pobrane były próbki.

4. Opracowanie algorytmu do tłumaczenia nazw zwyczajowych/słownych jednostek chorobowych na kody ICD-10 (preferowane podejście NLP).

W świecie medycyny istnieje wiele określeń na opisanie jednej i tej samej jednostki chorobowej. Jest to istotny problem z punktu widzenia analizy danych pochodzących z wielu ośrodków pomiędzy którymi nie określono jednolitego nazewnictwa jednostek chorobowych. Problem ten dotyka w dużej mierze także badań populacyjnych, gdzie choroby dotykające pacjentów wprowadzane są przez nich samodzielnie lub przez ankieterów, którzy nie zawsze mają przygotowanie medyczne. Sposobem na rozwiązanie tego problemu jest stworzenie translatora na jednorodny system nazewnictwa jednostek chorobowych a mianowicie ICD-10. Z racji na wielorakość stosowanych nazw zwyczajowych optymalnym rozwiązaniem jest zastosowanie metod NLP.

5. Opracowanie lokalnego narzędzia do adnotacji genomów bakteriofagowych.

Adnotacja genomów bakteriofagowych jest problemem w dzisiejszej biotechnologii. Istnieje wiele aplikacji pozwalających na wykonanie tego zadania. Istnieje jeszcze jednak duże pole do rozwoju tego typu narzędzi poprzez lepsze metody poszukiwania ORF dostosowane do bakteriofagów oraz przygotowanie bogatszych baz danych genów, na podstawie których genomy będą adnotowane.

6. Opracowanie metody uporządkowanego przechowywania danych pochodzących z sekwencjonowania NGS.

Przechowywanie i przetwarzanie danych powstałych w wyniku sekwencjonowania NGS stanowi istotny problem w świecie badań wykorzystujących tą metodę. Każda tura sekwencjonowania generuje dziesiątki gigabajtów danych które należy przechowywać w sposób uporządkowany i umożliwiający ich sprawne wykorzystanie w późniejszym czasie. Celem zadania jest opracowanie systemu pozwalającego na przechowywanie danych w formie łatwej do wykorzystania oraz pozwalającego na archiwizację próbek które zostały zanalizowane i prawdopodobnie nie będą wykorzystywane w najbliższym czasie.

Pamiętajcie, że będą Was wspierać

nasi Mentorzy :)