Semestrální práce

Analýza sekvenčních dat SARS-CoV-2 a vzájemných vztahů mutací/proteinů v jednotlivých variantách

# Osnova

1. **průzkum dat**

* z jakých dat můžeme vyjít
* prozkoumat hezky osekvenovaná data

1. **výběr vhodných dat**

* odfiltrování nevhodných dat
  + vyfiltrovat si sekvence označené pouze u lidí
  + odstřihnout nezajímavé/obecné části sekvencí

1. **zarovnání dat**

* pouze v případě, že budeme používat sekvence nukleotidů
* máme dvě možnosti, jak k zarovnání přistoupit:
  + chybějící části (delece), neboli to, jestli tam nukleotid je či není, může korelovat s nějakou mutací, tudíž se jich zbavit nechceme! (v tom případě ale hledat mutace pouze pro delece na začátku a konci sekvence, jinak by nám to mohlo hodně zkreslit krátké delece uvnitř sekvence)
  + chybějící části (delece) na okraji sekvence nám akorát zkreslují vliv delecí uvnitř sekvence, proto se jich zbavíme, k čemuž nám mohou sloužit dva postupy:
    - delece z okrajů jednoduše ořízneme podle největší nalezené delece ve skupině sekvencí
    - ze sekvencí nejdříve vyřadíme ty s extrémně velkou delecí na začátku a konci, potom teprve ořežeme dle bodu výše

1. **lokalizace příznaků**
2. **korelování jednotlivých příznaků**

# Dotazy

* sekvence nukleotidů vs. sekvence proteinů?
  + lépe nukleotidů, osekáním na proteiny bychom se připravili o podstatnou část genomu, která nás ale zajímá
* můžeme sekvence nějak seskupovat?
  + pro začátek je ideální nastahovat co nejvíce dat, zarovnat je a najít příznaky, až v další fázi případně rozdělit na dílčí skupiny podobných sekvencí a v nich hledat konkrétnější příznaky

# Tipy

* bude potřeba ošetřovat co nejvíce false positives případů
* kontrolovat duplicity sekvencí
* rozhodně se neomezovat na jednu laboratoř/místo/čas, ale nechat viru jeho vývoj, brát v co nejširším časovém pásu
* deleci i inzerci ideálně zahrnout jako jednu z featur dané sekvence, která také může způsobovat určité mutace

# Odkazy

<https://www.ncbi.nlm.nih.gov/sars-cov-2/>

<https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/variant-surveillance.html>

<https://nextstrain.org/sars-cov-2/>