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

# Nevyřešené problémy pro 11.5

* Detekce mutaci?
  + Najít správný nástroj pro detekci mutací (nebo napsat mismatch funkci), ale to moc
* Vyřazovnání málo četných mutací
  + Omezení práce na mutace které mají aspoň n-výskytů (hlavně kvůli omezení přiznakového prostoru)
* Vyřazovnání příliš četných mutací ( C→T)
  + vyřazení, neboť se vyskytují tak často, že se nemohou s ničím párovat
  + jak je identifikovat?

# Odkazy:

Python mutace:

https://towardsdatascience.com/python-and-numpy-for-sars-cov-2-gene-mutation-analysis-805833fbb23f

Pro nás příliš složitá analýza:

<http://staden.sourceforge.net/mutations/index.html>

https://docs.gdc.cancer.gov/Data/Bioinformatics\_Pipelines/DNA\_Seq\_Variant\_Calling\_Pipeline/

Vlákno na RG:

https://www.researchgate.net/post/What\_is\_the\_most\_useful\_program\_to\_detect\_mutation\_in\_DNA\_SEQUENCES