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

Semestrální práce z předmětu Molekulární biologie a genetika

Přírodovědecká fakulta Univerzity Karlovy, 29. 8. 2021

Markéta Bařinková, Lukáš Daněk, Pavlína Koutecká, Matěj Kudrna, Pavel Myšička

Úvod

Cílem této práce bylo analyzovat sekvence genomů SARS-CoV-2, jmenovitě odhalení vztahů a souvislostí mezi mutacemi jednotlivých lokusů. Zpracování proběhlo ve třech základních krocích, které jsou shrnuty níže.

1. Data pre-processing – Tento krok sestával ze zarovnání sekvencí (tzv. multiple alignment), jejich ořezání na stejnou délku a výběru části dat pro následnou analýzu.
2. Primární analýza – Během tohoto kroku jsme zkoumali základní atributy sekvencí – četnost mutací přes sekvenace, počet variant nukleotidů u jednoho lokusu atd.
3. Statistická analýza – V posledním kroku bylo provedeno vlastní statistické zpracování dat.

Použitá data

Pro analýzu byla použita data z internetové knihovny NCBI library. V době začátku tohoto projektu bylo dostupných přibližně 96 000 kompletních sekvencí genomů SARS-CoV-2. Pro účely tohoto semestrálního projektu byla použita data obsahující 100, 1000 a 10000 sekvencí. Pro případné budoucí použití jsou k dispozici rovněž data obsahující více sekvencí.

Data pre-processing

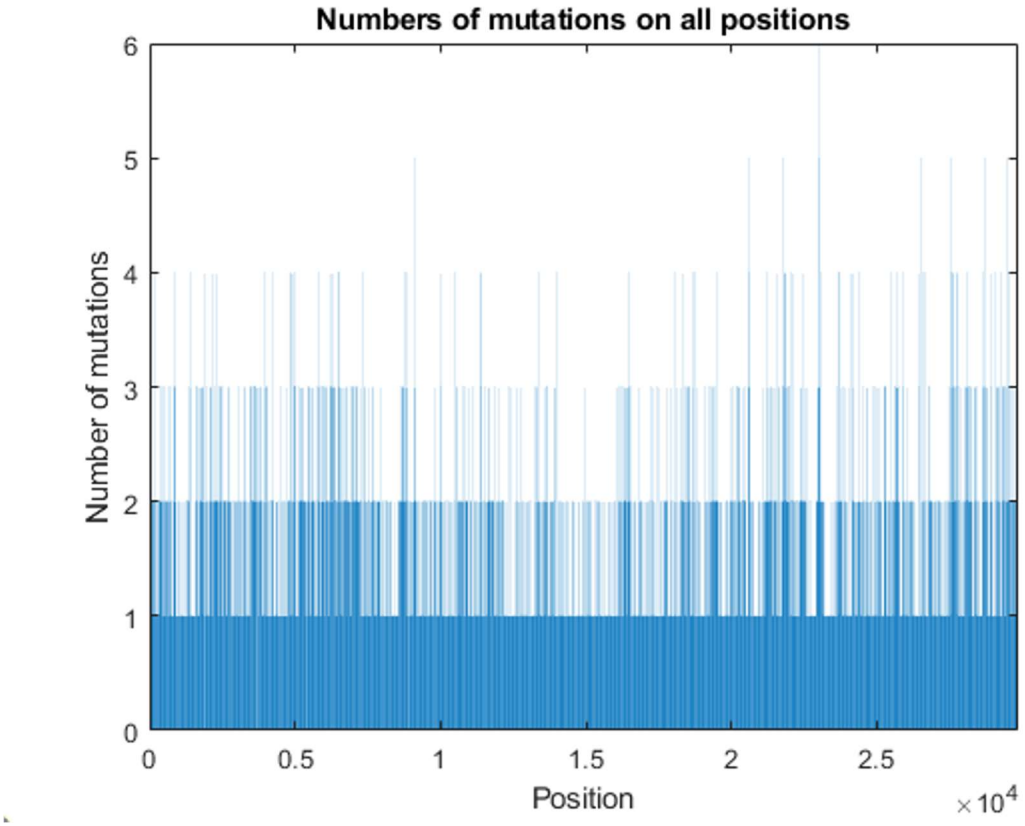
V první řadě byly sekvence před další analýzou zarovnány (byl proveden tzv. multiple alignment) pomocí softwaru MAFFT (Pro velký počet sekvencí byl použit RCI Cluster). Jako referenční sekvence byl použit kompletní izolovaný genom **NC_045512.2.** Pro další zpracování sekvencí byl vytvořen skript v Pythonu, pomocí kterého bylo možné vybrat ze souboru ve formátu FASTA náhodně vybrat určitý počet seřazených sekvencí a ten na krajích oříznout. Bylo totiž vypozorováno, že na okrajích segmentů často docházelo k delecím, či neúplnému sekvenování. Tyto konce by narušovaly výslednou statistickou analýzu. Pro ořezávání byly vytvořeny dva různě restriktivní mechanismy. První využil maximální nalezené mezery na 3‘ a 5‘ genomu a o tento počet oříznul všechny sekvence. Druhý využil průměrné mezery na 3‘ a 5‘ konci prodloužené o odchylku a podle této délky zkrátil všechny sekvence. Takto seřazené a oříznuté sekvence byly následně vyexporotvány do dalšího FASTA souboru, který byl použit v následných analýzách.

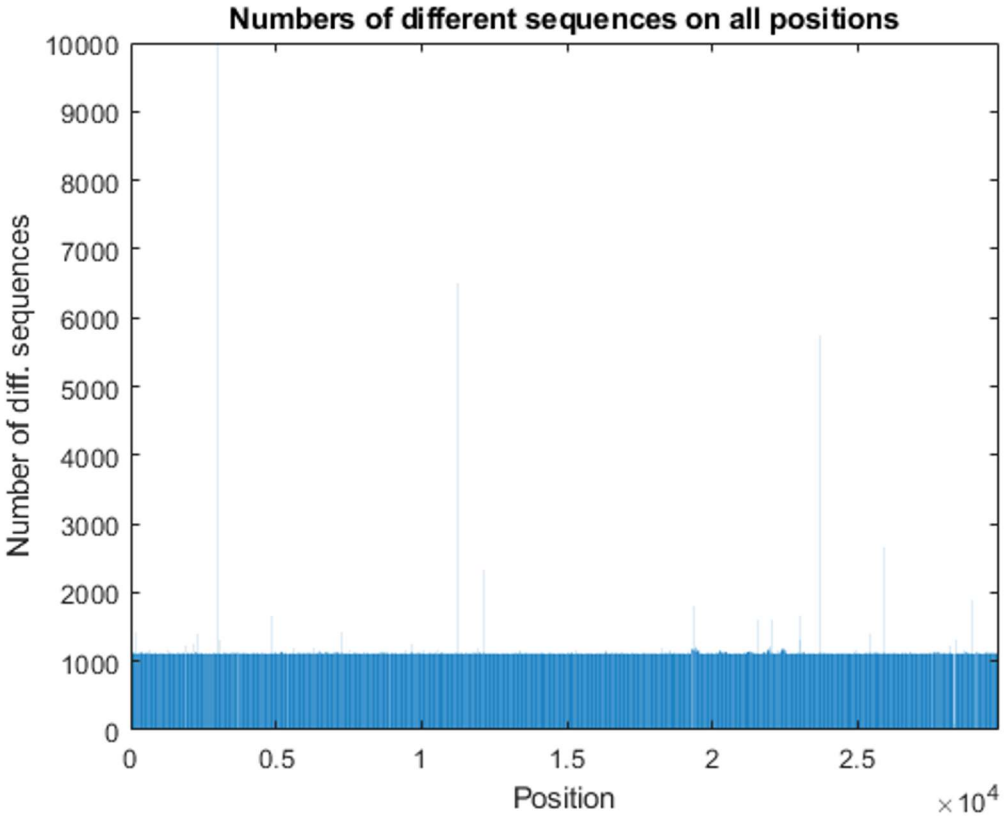
Primární analýza

Pomocí software vytvořeného v jazyce Python byly spočítány mutace na všech pozicích zarovnaných a oříznutých sekvencí. Dále byly určeny počty sekvencí, které se na dané pozici liší od referenční sekvence. Předpokládalo se, že v sekvencích existují určité skupiny pozic, na kterých se často vyskytují mutace, ale také skupiny, na kterých se mutace téměř nevyskytují.

V programu Matlab byla proto provedena analýza za účelem ověření těchto předpokladů. Pro analýzu byla vybrána skupina 10000 sekvencí. Ukázalo se, že alespoň jedna mutace se v tomto počtu sekvencí vyskytuje prakticky na každé pozici. Proto bylo použito filtrování na základě počtu sekvencí lišících se od reference. Mutace na dané pozici byly prohlášeny za významné, pokud se na této pozici lišilo od referenční sekvence alespoň 12 % celkového počtu sekvencí (tedy 1200 sekvencí). Tím bylo získáno 279 pozic z celkového počtu 29851. Tento práh byl stanoven empiricky jako určitý zlom ve výsledném počtu pozic, neboť nižší práh nevedl k významné filtraci (pro práh 11 % bylo pozic stále více než 29000). Tento výrazný zlom v počtu pozic ukazuje, že existuje mnoho pozic, na kterých jsou mutace velmi málo časté.

Na následujících grafech jsou zobrazeny nejprve počty mutací na všech pozicích, poté počty sekvencí, které se na dané pozici liší od reference, a na posledním grafu pouze mutace na pozicích označených za významné. Z druhého grafu můžeme získat představu o četnosti mutací přes všechny zpracované sekvence. Zároveň můžeme pozorovat i lokace, u kterých se většina sekvencí liší od té referenční. Na posledním grafu jsou patrné určité skupinky blízko ležících bodů, což potvrzuje původní předpoklady o existenci lokusů, u kterých se mutace vyskytují častěji.




Numbers of different sequences on all positions
Number of diff. sequences
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0
0.5
1
1.5
2
2.5
Position
×10^4

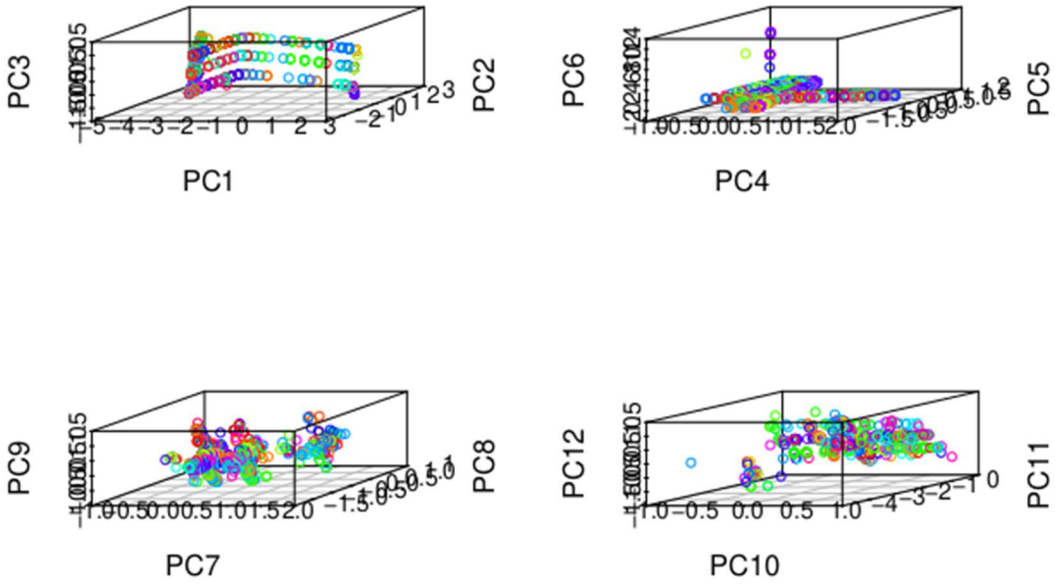

Numbers of mutations on positions
Number of mutations
6
5.5
5
4.5
4
3.5
3
2.5
2
0
0.5
1
1.5
2
2.5
3
Position
×10^4

Statistická analýza

Metoda založená na PCA

Pro analýzu pomocí PCA byla použita data ve formátu binární matice, kde hodnota na pozici *(m,n)* indikovala, zda se sekvence *m* na pozici *n* liší od referenční sekvence (1), či nikoliv (0). Původní myšlenkou pro využití PCA byla identifikace korelací pomocí vektorů výsledné transformační matice. Tento přístup se ovšem ukázal jako nedostatečně teoreticky prozkoumaný a nevhodný kvůli jeho nárokům na RAM paměť.

Výsledky PCA byly ale použity aspoň jako sanity check. Ukazují totiž, že i po provedeném předzpracování dat jsou si pozorování (DNA sekvence viru) sbíraná v jednom místě a čase bližší, než data sbíraná na jiných místech a v jiný čas. Na následujících ilustracích jsou vidět data zpracovaná pomocí PCA a jejich barva odpovídá pozici pozorování v datasetu. Vidíme, že minimálně pro několik prvních komponent se jednotlivá pozorování shlukují podle míst sběru a času (shlukují se podle barev). Až pro další komponenty se v jejich uspořádání projevuje větší náhodnost.



Metoda založená na korelaci

Pro tuto metodu byla vytvořena funkce v prostředí MATLAB. Nejprve byla vytvořena fiktivní referenční sekvence tvořená ze sekvence dané modem každého lokusu. Takto byla reference utvořena z toho důvodu, že původně používaná referenční sekvence se zřejmě v čase výrazně změnila a pro následující kroky by vznikla velmi složitá matice pro kros-korelaci. Data byla následně binarizována – v každé sekvenci se vyskytla 1 tam, kde se daný nukleotid lišil od referenčního, a 0 tam, kde ke změně nedošlo. Před samotným zpracováním je možné odebrat data z lokací, kde došlo k minimálnímu počtu mutací přes všechny lokusy (v našem případě

jsme zvolili hranici 0.5 %, ale pomocí parametrů funkce je toto číslo snižovat až na 0, která odebere jen lokusy, kde nedošlo k vůbec žádné mutaci).

Následný první korelační krok byl proveden na matici takto binarizovaných dat, přičemž hladina významnosti alfa byla korigování Bonferroniho korekcí pro opakované korelační kroky. Tímto jsme získali lokusy, kde jsou často korelované mutace přes všechny sekvence. Druhý krok také používal korelaci, ale na nebinární data. Na matici, která obsahovala již jen silné korelované lokusy a nukleotidy nacházející se v každé sekvenci na těchto pozicích byl proveden Chi-kvadrát test dobré shody, pomocí kterého bylo odhaleno, zda se na lokacích, které spolu často mutují, vyskytují s větší pravděpodobností mutace stejné, či nikoliv. Hladina významnosti alfa byla opět upravena pomocí Bonferroniho korekce. Výsledné dvojice lokusů byly kvůli velkému počtu dat dále filtrovaný takovým způsobem, aby výsledky zahrnuly jen konkrétní mutace, ne degenerované nukleotidy.

Výsledné korelované lokusy včetně typu mutace byly zapsány do tabulek, které jsou výstupem celé funkce. Tabulky si lze prohlédnout na úložišti ownCloud. Výsledné počty se pohybují ve stovkách dvojic lokusů. Tento počet by bylo možné dále snížit omezením hladiny významnosti alfa u jednotlivých statistických testů, omezením na minimální vzdálenost použitých lokusů (vyskytují se zde i korelace mezi sousedními nukleotidy, což by mohlo poukazovat na mutaci v rámci jedné aminokyseliny a nepřineslo by to z hlediska korelace lokusů větší informační hodnotu), omezením na větší počet mutovaných sekvencí (zde byla za hladinu považována hodnota > 0.5 %) a dalšími metodami.