PB051

Výpočetní metody v bioinformatice a systémové biologii

Bioinformatika

FI, MU, jaro 2021, Oto Stanko

Z PDB som pozbieral 15 sekvencií z rodiny vírov, do ktorej patrí SARS-CoV-2. Všetky sekvencie sú uložené v samostatných súboroch. Názov súboru je zložený z rodiny viru, do ktorej patrí, organizmu z ktorého bola sekvencia získaná a krajiny, kde sa organizmus vyskytoval.
Všetky sekvencie som vložil do jedného súboru covid_sequences.fa.
Pomocou príkazu: perl -ne 'if (/^>/){$seq=~s/\r?\n(.)/$1/g;print $seq; $seq=q{};print;}else{$seq.=$_}END{$seq=~s/\r?\n(.)/$1/g;print $seq;}' covid_sequences.fa > covid_sequences_single.fa
som všetky sekvencie upravil na jeden riadok každú.
Viacnásobné zarovnanie som spravil pomocou programu clustalo príkazom: clustalo -i covid_sequences_single.fa -o covid_sequences.aln --outfmt clu
Hľadanie konsenzuálnej sekvencie som zvolil online nástroj emboss https://emboss.bioinformatics.nl/cgi-bin/emboss/cons a opäť zarovnal na jeden riadok do súboru consensus_single.
consensus_single som následne použil na hľadanie ORF pomocou online nástroja: https://www.bioinformatics.org/sms2/orf_find.html Nastroj identifikoval rôzne ORF v zavislosti na zadaných parametroch. Ani jeden ORF nevyhovoval Spike proteinu.
Na základe preštudovania primárnej štruktúry spike proteínov v SARS-CoV-2, SARS-cCV a MERS-CoV som prešiel vlastnú konsenzuálnu sekvenciu ako aj viacnásobne zarovnanie a došiel som k záveru, že pri zarovnaní sa časť sekvencie pre S-protein pri MERS sekvenciách nezarovnala spolu s ostatnými a teda vo výslednej konsenzuálnej sekvencii ORF pre S-protein je výsledkom len SARS-CoV-2 a SARS-CoV. Vďaka tejto infotmácii som bol schopný identifikovať ORF pre S-proteín.
Začiatok je na pozícii 22478 aminokyselinou M (ATG) končí na pozícii 26804 AMK T (ACA) za ktorou následuje stop kodon TAA.
Celá sekvencia sa nachádza v súbore spikeFromCons.

Následne som zozbieral 12 primárnych sekvencií Spike proteínov do súboru spike_sequences.fa a zarovnal ich na jeden riadok každú do súboru spike_sequences_single.fa
Stiahol som si program MEGA, v ktorom som si spravil zarovnanie sekvencií a to následne použil na vytvorenie fylogenetického stromu uloženého v súbore phylogenetic_tree
PNG súbory phylogenetic_tree.png a phylogenetic_tree_circle.png sú grafické výstupy z MEGA. Jeden ako časový strom, druhý v kruhovom prevedení.
Zo získaného stromu je názorne vidno odlišnosť spike proteínov u MERS-CoV a SRAS-CoV/SARS-CoV-2, kde SARS-CoV a SARS-CoV-2 sú si fylogeneticky bližšie ako MERS-CoV. Spike proteín získaný z netopiera sa približoval najviac MERS-u.
Viacnásobným zarovnaním cez clustalo a pomocou online nástroja som získal konsenzuálnu sekvenciu spike_sequence_consensus.fasta Nástrojom: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome
Som namapoval získanú konsenzuálnu sekvenciu na najbližší proteín PDB. Ako najbližši proteín s najvyšším skóre vyšiel 7WCL_A, čo je Chain A, Spike glycoprotein (SARS-CoV-2)
v súbore PDB_BLAST_7cwl.png je grafické znázornenie rozdielnosti v aminokyselinách. Odlišnost jednotlivých AMK si vysvetlujem tým, že môj konsenzus pozostáva z proteínov pochádzajúcich zo SARS/SARS-CoV-2 ako aj MERS. Kde pri zarovnaní mohlo dôjsť k nepresnostiam kvôli odlišnosti jednotlivých AMK v rôznych viroch.
Toto podporuje aj výskyt úsekov xxxxxxxxx (programom na výrobu konsenzu nerozhodnutými pozíciami).

FI, MU, jaro 2021, Oto Stanko