

Istraživanje podataka u bioinformatiči

Teme za seminarske radove

1. Klasterovanje izolata SARS-CoV-2 prema RSCU

Zadatak

- Izračunati ukupan RSCU¹ (Relative Synonymous Codon Usage) SARS-CoV-2 izolata humanog domaćina i svakog izolata ponaosob;
- Izvršiti klasterovanje izolata po RSCU atributima za $k=1..10$ i uporediti kvalitet klasterovanja
- Za najbolji rezultat, sumirati podatke o geografskom poreklu i datumu uzimanja materijala u svakom od klastera uz vizuelni prikaz.

Postupak

1. Skinuti metapodatke o svim kompletnim SARS-CoV2 human-izolatima 9 kladusa (grana) sa GISAID-a (<https://www.gisaid.org/phylogenetics/global/nextstrain/>) i izdvojiti one koji imaju *genbank_accession*
2. Za prethodni skup skinuti kodirajuće regione ili cele genome (FASTA) (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=SARS-CoV-2,%20taxid:2697049)
3. Izbaciti one koji sadrže ne-ACGT slova u CDS-ovima
4. Eliminirati identične (duplikate) (najbolje skladištenjem u bazu pa poređenjem po jednakosti)
5. Koristiti python programski jezik (ili postojeći softver) za računanje RSCU

2. Genomski polimorfizam SARS-CoV-2

Zadatak

- Izračunati odnos Ka/Ks u skupu izolata *human-host*, i u skupu *human-host+outgroup* (outgroup *pangolin* ili *mink*)
- Izračunati odnos Ka/Ks² u skupu izolata *human-host*, i u skupu *human-host+outgroup*³ (outgroup *pangolin* ili *mink*)

Postupak

¹ <https://www.hindawi.com/journals/abi/2009/316936/>

² https://en.wikipedia.org/wiki/Ka/Ks_ratio

³ [https://en.wikipedia.org/wiki/Outgroup_\(cladistics\)](https://en.wikipedia.org/wiki/Outgroup_(cladistics))

1. Skinuti metapodatke o svim kompletnim SARS-CoV2 human-izolatima 9 kladusa (grana) sa GISAID-a (<https://www.gisaid.org/phylogenetics/global/nextstrain/>) i izdvojiti one koji imaju *genbank_accession*
2. Sa sajta (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=SARS-CoV-2,%20taxid:2697049) skinuti kodirajuće regione ili cele genome (FASTA) za prethodni skup i za jedan *pangolin* (jedan od kompletnih genoma sa https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&Completeness_s=complete&VirusLineage_ss=Pangolin%20coronavirus,%20taxid:2708335 koji nema ne-ACGT slova u CDS-ovima) i jedan *mink* (jedan od https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&Completeness_s=complete&VirusLineage_ss=Mink%20coronavirus%201,%20taxid:1913642 koji nema ne-ACGT slova u CDS-ovima)
3. Eliminirati identične (duplikate) (najbolje skladištenjem u bazu pa poređenjem po jednakosti)
4. Odrediti MSA u odnosu na svaki od gena referentnog izolata (strukturni S, M, E, N i nestrukturni ORFs) pomoću CLUSTAL-OMEGA
5. izračunati Ka/Ks odnos za humane izolate po genu i zbirno po protein-kodirajućim sekvencama (hint: [DnaSP v6](#))
6. Izračunati Ka/Ks odnos za humane + 1 outgroup (*pangolin* ili *mink*) - da li je neki od gena, posebno S gen, predmet Darvinijanske selekcije?

3. Predikcija uspešnosti elektrokonverzije

Zadatak

- Pronaći attribute značajne za predikciju uspešnosti elektrokonverzije⁴ u odnosu na kliničku sliku pacijenta, druge indikacije i dodeljenu terapiju
- Konstruisati bajesovsku mrežu koja uključuje relevantne attribute i vrši predikciju uspešnosti elektrokonverzije
- Testirati druge metode istraživanja podataka (neke od: SVM, neuronske mreže, stabla odlučivanja, logistička regresija) i uporediti sa uspešnosti predikcije korišćenjem bajesovske mreže

Uputstva

- Preuzeti tabelu sa podacima (kontaktirati asistenta)
- Eliminirati suvišne attribute (prazne kolone)
- Obraditi kolone sa mešanim numeričkim i nenumeričkim vrednostima
- Analizirati značajnost atributa za predikciju uspešnosti elektrokonverzije
- Konstruisati bajesovsku mrežu i izračunati vrednosti ocena kvaliteta (preciznost, r2 skor, specifičnost i senzitivnost) predikcije dobijenom mrežom i matricu konfuzije
- Pronaći što bolji model mreže, testirati različite strukture kao i učenje strukture mreže iz podataka

⁴ <https://sr.wikipedia.org/sr-el/Kardioverzija>

- Odabrati neke od drugih tehnika istraživanja podataka i uporediti ocene kvaliteta modela sa ocenama bayesovske mreže

4. Predikcija održavanja sinusnog ritma posle godinu dana

Zadatak

- Pronaći attribute značajne za predikciju uspešnosti održavanja sinusnog srčanog ritma godinu dana nakon obavljene elektrokonverzije u odnosu na kliničku sliku pacijenta, druge indikacije i dodeljenu terapiju
- Konstruisati bayesovsku mrežu koja uključuje relevantne attribute i vrši predikciju uspešnosti održavanja sinusnog ritma godinu dana nakon elektrokonverzije
- Testirati druge metode istraživanja podataka (neke od: SVM, neuronske mreže, stabla odlučivanja, logistička regresija) i uporediti sa uspešnosti predikcije korišćenjem bayesovske mreže

Uputstva

- Preuzeti tabelu sa podacima (kontaktirati asistenta)
- Eliminirati suvišne attribute (prazne kolone)
- Obraditi kolone sa mešanim numeričkim i nenumeričkim vrednostima
- Analizirati značajnost atributa za predikciju uspešnosti održavanja sinusnog ritma
- Konstruisati bayesovsku mrežu i izračunati vrednosti ocena kvaliteta (preciznost, r^2 skor, specifičnost i senzitivnost) predikcije dobijenom mrežom i matricu konfuzije
- Pronaći što bolji model mreže, testirati različite strukture kao i učenje strukture mreže iz podataka
- Odabrati neke od drugih tehnika istraživanja podataka i uporediti ocene kvaliteta modela sa ocenama bayesovske mreže