Заузеће задатака

Слободни задаци: 1, 2, 8, 14, 18, 21, 27

Заузети задаци: 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 15, 16, 17, 19, 20, 22, 23, 24, 25, 26, 29, 30, 31

Задатак 1

Анализирати утицај P-adičnosti на разлике генетског кода SARS-1, SARS-2 и MERS коронавируса. За SARS2 коронавирус урадити анализу и по WHO класификацији. Испитати да ли класификација или кластеровање протеинских секвенци по врсти коронавируса и типу протеина може да се одреди на основу P-adičnost њихових нуклеотидних секвенци.

- Приказати појам P-adičnosti и повезаност са генетским кодом.
- Геноме SARS-1, SARS-2 и MERS коронавируса скинути са
  - SARS1 (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:2901879)
  - SARS2 (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:2697049)
  - MERS (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:1335626)
- Употребу кодона преузети са https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi

За SARS2 коронавирус урадити анализу и по WHO класификацији. Поделу преузети од наставника. Референце о P-adičnosti - радове проф. Бранка Драговића скинути са мреже или преузети од наставника.

Задатак одабрали:

Задатак 2

За геноме коронавируса

- SARS1 (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:2901879)
- MERS (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:1335626)
- BCOV (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:11128)
- BAT SARS (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:442736)

1. одредити директне и индиректне некомплементарне понављајуће секвенце у аминокиселинским секвенцама користећи програм StatRepeats из пакета
   http://bioinfo.matf.bg.ac.rs/home/downloads.wafl?cat=Software&project=RepeatsPlus
2. на основу њих направити модел (класификација, кластеровање или правила придруживања) за одређивање врсте коронавируса.

Обавезно урадити следеће:

- консултовати се са наставником око потребних корака при преузимању материјала и одређивању понављајућих секвенци
- изабрати само протеине чији изолати имају 0 двосмислених карактера и који су нуклеотидно комплетни
- податке за све изолате преузимати са https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid: уз додавање таксономског броја тражене фамилије на крају
- понављајуће секвенце урадити за *spike glycoprotein* и још један протеин
- модел тестирати помоћу бар 3 различита алгоритма

Задатак одабрали:

Задатак 3

За геноме коронавируса

- SARS1 (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:2901879)
- MERS (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:1335626)
- BCOV (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:11128)
- BAT SARS (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:442736)

1. одредити директне и индиректне комплементарне и некомплементарне понављајуће секвенце у кодирајућим секвенцама користећи програм StatRepeats из пакета
   http://bioinfo.matf.bg.ac.rs/home/downloads.wafl?cat=Software&project=RepeatsPlus
2. на основу њих направити модел (класификација, кластеровање или правила придруживања) за одређивање врсте коронавируса.

Обавезно урадити следеће:

- консултовати се са наставником око потребних корака при преузимању материјала и одређивању понављајућих секвенци
- изабрати само протеине чији изолати имају 0 двосмислених карактера и који су нуклеотидно комплетни
- податке за све изолате преузимати са https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid: уз додавање таксономског броја тражене фамилије на крају
- понављајуће секвенце урадити за *spike glycoprotein*
- модел тестирати помоћу бар 3 различита алгоритма

Задатак одабрали: Марија Ристић (188/2019) и Димитрије Петровић (39/2019)

Задатак 4

Испитати постојање нуклеотидних секвенци које карактеришу S-протеин у SARS и EBOLA virusima.

- Приказати карактеристике S-протеина и преглед техника истраживања података које омогућују претрагу секвенци и образаца (део истраживања текста).
- Геноме SARS-1, SARS-2, MERS коронавируса и EBOLA вируса скинути са
  - SARS1 (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:2901879)
  - SARS2 (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:2697049)
  - MERS (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:1335626)
  - EBOLA (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:3044781)
- Изабрати само протеине чији изолати имају 0 двосмислених карактера и који су нуклеотидно комплетни
- Важно: од SARS2 коронавируса узети само по 10 представника сваког соја по WHO класификацији

Задатак одабрале: Симона Јевтовић (139/2019) и Сања Недељковић (166/2019)

Задатак 5

Улазни подаци се састоје од скупа секвенци генома SARS-CoV-2 вируса који су секвенцирани из узорака сакупљених на територији Србије за период 2020-2023. година уз додатак референтне секвенце SARS-CoV-2 вируса (NCBI идентификација NC_045512.2). Податке у Fasta формату преузети у договору са наставником.

- Поравнати нуклеотидне секвенце у односу на референтни изолат. За тако поравнате секвенце одредити могуће границе протеина
- Идентификовати 5 најчешће мутираних места у геному SARS-CoV-2 вируса из узорака са територије Србије. Учесталост одредити као проценат замена у односу на референтни изолат на конкретној позицији.) Израчунавања урадити за следеће опције:
  - N (неидентификовани нуклеотид) рачунати као нуклеотидну замену
  - N рачунати као референтни нуклеотид (нема мутације)
  - Из рачунања учесталости промена за дату позицију искључити све секвенце које имају N на тој позицији.
- Идентификовати 5 најређе мутираних места у геному SARS-CoV-2 вируса из узорака са територије Србије. Учесталост одредити као проценат замена у односу на референтни изолат на конкретној позицији. Израчунавања урадити за следеће опције:
  - N (неидентификовани нуклеотид) рачунати као нуклеотидну замену
  - N рачунати као референтни нуклеотид (нема мутације)
  - Из рачунања учесталости промена за дату позицију искључити све секвенце које имају N на тој позицији.
- Идентификовати регионе од 5 нуклеотида у низу који су најређе мутирани у геному SARS-CoV-2 вируса из узорака са територије Србије. Учесталост одредити као проценат замена у односу на регион референтног изолата на конкретној позицији, а мутираност за регион сабирањем вредности за сваку позицију. Искључити из рачунања учесталости промена за конкретну позицију све секвенце које имају N (неидентификовани нуклеотид) на тој позицији, односно сматрати да на тој позицији нема мутације.
- Идентификовати 5 најдужих региона (нуклеотидних низова) у геному SARS-CoV-2 вируса из узорака са територије Србије, чија је стопа мутираности мања од 15%. Учесталост одредити као проценат замена у односи на референтни изолат на конкретној позицији, а мутираност за сваки регион израчунавати сабирањем вредности за сваку позицију тог региона. Искључити из рачунања учесталости промена за конкретну позицију све секвенце које имају N (неидентификовани нуклеотид) на тој позицији, односно сматрати да на тој позицији нема мутације.

- За протеине чије су позиције одређене након поравнања нуклеотидних секвенци, извршити превођење нуклеотидне у аминокиселинску секвенцу. Употребу кодона за превођење преузети са https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi (користити стандардни код - transl_table=1) Одредити проценат мутација тако добијених аминокиселинских секвенци на свакој од позиција и упоредити добијене резултате са резултатима мутација на нуклеотидном нивоу.

Задатак одабрале: Андријана Ивковић (115/2019) и Ивана Несторовић (130/2019)

Задатак 6

Улазни подаци се састоје од скупа секвенци генома SARS-CoV-2 вируса који су секвенцирани из узорака сакупљених на територији Србије и околних европских земаља.

- Потребно је одредити кретање изолата SARS-CoV-2 коронавируса ка Србији. "Кретање"обухвата упоређивање секвенци и датума појаве одређеног изолата у некој од земаља пре појављивања у Србији, као и приказ њиховог груписања. Одредити и проценат разлика одређеног изолата у Србији који је "дошао"у Србију из околних земаља.
- Приказати укратко анализу временских серија као технику истраживања података.
- Резултате визуелизовати у облику графа.

Податке у Fasta формату преузети у договору са наставником.

Задатак одабрале: Александра Лабовић (150/2019) и Мила Лукић (222/2018)

Задатак 7

За геноме коронавируса

- SARS1 (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:2901879)
- MERS (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:1335626)
- BCOV (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:11128)
- Human coronavirus 229E (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:11137)
- Human coronavirus OC43 (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:31631)

1. Изабрати само протеине чији изолати имају 0 двосмислених карактера и који су нуклеотидно комплетни
2. На основу употребе кодона направити модел (класификација, кластеровање или правила придруживања) за
   (a) класификацију типова протеинских секвенци
   (b) класификацију (одређивање) врсте коронавируса

Обавезно урадити следеће:

- консултовати се са наставником око потребних корака при преузимању материјала и одређивању понављајућих секвенци
- податке за све изолате преузимати са https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:2697049 претраживањем по таксономији вируса
- употребу кодона за превођење преузети са https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi (користити стандардни код - transl_table=1)

Задатак одабрале: Милева Симић (156/2019) и Јелена Максимовић (193/2018)

Задатак 8

Поновити поступак описан у раду
An Integrative Network Approach to Identify Common Genes for the Therapeutics in Tuberculosis and Its Overlapping Non-Communicable Diseases аутора *Aftab Alam, Hala Abubaker Bagabir, Armiya Sultan, Mohd Faizan Siddiqui, Nikhat Imam, Mustfa F Alkhanani, Ahmad Alsulimani, Shafiul Haque and Romana Ishrat* објављеног у *Frontiers in Pharmacology*, 27.01.2022. godine. Рад може да се преузме са линка *doi: 10.3389/fphar.2021.770762*

Нагласак у поступку је формирање мрежа које представљају интеракцију и кластеровање гена и интеракцију лекова. Консултовати се са наставником око потребних корака при преузимању материјала.

Задатак одабрали:

Задатак 9

Поновити поступак описан у раду
Network-medicine approach for the identification of genetic association of parathyroid adenoma with cardiovascular disease and type-2 diabetes аутора *Nikhat Imam, Aftab Alam, Mohd Faizan Siddiqui, Akhtar Veg, Sadik Bay, Md. Jawed Ikbal Khan and Romana Ishrat* објављеног у *Briefings in Functional Genomics, 2023, 22, pp.250–262.* Рад може да се преузме са линка *https://doi.org/10.1093/bfgp/elac054*

Нагласак у поступку је формирање мрежа које представљају интеракцију и кластеровање протеина који представљају везу између кардиоваскуларних болести, дијабетеса типа 2 и паратироидног аденома. Консултовати се са наставником око потребних корака при преузимању материјала.

Задатак одабрали: Марко Никитовић (123/2020) и Бошко Андрић (26/2020)

Задатак 10

Извршити анализу текстова песама на српском језику и на основу ње класификовати песме према ауторима.
Улазни материјал (скуп песама) преузети од наставника.

Изглед дела улазног текста:

Padajte, braćo, plin'te u krvi! Ostav'te sela nek gori plam! Bacajte sami u oganj decu! Stresite s sebe ropstvo i sram!
Mudraci su prinosili dara: Smirnu, zlato, caru judejskome; I ja darak caru nosim svome: Evo primi - pesmu iz nedara!
Mnogi me je dosad zapitkiv'o: sa čega sam srca obolela? Sad nek znade moja družba cela: rana j' ljubav, što sam od njih skriv'o.
Ja sam stena, o koju se zloba mori, svetska čuda i pokori. Mnogi težak oblak, jeka, krš gromova, oganj, kletva i sto čuda neba,
Kroz ponoć nemu i gusto granje vidi se zvezda tiho treptanje, čuje se srca silno kucanje; - O, lakše samo kroz gusto granje!
„Vina, Milo!“ - orilo se, dok je Mila ovde bila. Sad se mila izgubila: Tuđe ruke vino nose. Ana toči, Ana služi, al' za Milom srce tuži.
Jeste li mi rod, siročići mali? Il' su i vas, možda, jadi otrovali? Ili vas je, slabe, progonio svet - pa dođoste samo da, kad ljude
I ovaj kamen zemlje Srbije, što preteć suncu dere kroz oblak, sumornog čela mračnim borama, o vekovečnosti priča dalekoj, pokazujući
Zašto se meni javljaš tajno kada mi duša tiho sniva? I zašto tvoje oko sjajno golemu tugu i jad skriva? Zašto me kroz noć staneš zvati,
Sinoć, kad se vratih iz topla hamama, prođoh pokraj bašte staroga imama; Kad tamo, u bašti, u hladu jasmina, s ibrikom u ruci stajaše
Ponoć je. Ležim, a sve mislim na te - u tvojoj bašti ja te vidjeh juče, gdje bereš krupne raspukle granate. Mila kô zlatno nebo pošlje
Ko poškropi tvoje kose, Jelo, ko orosi tvoje lice b'jelo? "Jutros stajah ispod jorgovana, pa me rosa pokapa sa grana". Pravo kaži, a ne
Mati, mati, mila mati, oh, da mi je samo znati moju ljubav iskazati prema tebi što mi sja! Vidjela bi da j' vrelija od sunašca štono
Mi znamo sudbu i sve što nas čeka, no strah nam neće zalediti grudi! Volovi jaram trpe, a ne ljudi - Bog je slobodu dao za čovjeka.
Ostajte ovdje!...Sunce tuđeg neba, neće vas grijat kô što ovo grije; Grki su tamo zalogaji hljeba gdje svoga nema i gdje brata nije.
Pučina plava spava, prohladni pada mrak. Vrh hridi crne trne zadnji rumeni zrak. I jeca zvono bono, po kršu dršće zvuk; S uzdahom tuge
....
....

Задатак одабрале: Милица Тошић (105/2019) и Јелена Лазовић (288/2019)

Задатак 11

Скуп улазних података садржи податке из *PAIDB* базе (http://www.paidb.re.kr/about_paidb.php) о патогеним острвима бактерија *Escherichia coli* : http://www.paidb.re.kr/browse_genomes.php?m=g#Escherichia%20coli и *Helicobacter pylori*: http://www.paidb.re.kr/browse_genomes.php?m=g#Helicobact

Пронаћи обрасце са и без уметања и брисања појединих карактера у секвенцама које одговарају геномским острвима ове две фамилије бактерија. Задатак одабрали: Вељко Продан (163/2019) и Маја Миленковић (160/2019)

Задатак 12

Скуп улазних података садржи

- секвенце SARS-CoV-2 вируса и одговарајуће метаподатке издвојене из базе GISAID (https://gisaid.org/) уз коришћење опција *Complete* и *High Coverage*. Секвенце изабрати из узорака са територије следећих земаља: Мађарске, Грчке, Хрватске, Јапана и Србије.
- референтну секвенцу SARS-CoV-2 вируса (NCBI идентификација NC_045512.2) и одговарајуће метаподатке.

1. Поравнати нуклеотидне секвенце у односу на референтни изолат. За тако поравнате секвенце одредити могуће границе протеина у секвенцама издвојеним из GISAID базе.
2. Одредити проценат идентичних секвенци за сваки месец почевши од марта 2020. до јануара 2023. године, за сваки могући пар земаља (Мађарска-Грчка, Мађарска-Хрватска...).
3. Одредити проценат различитих секвенци за сваки месец почевши од марта 2020. до јануара 2023. године, за сваки могући пар земаља (Мађарска-Грчка, Мађарска-Хрватска....) у зависности од броја позиција на којима се налазе различити нуклеотиди. Табелу приказати по (претходно одређеним) протеинима.
4. Одредити проценат уникатних секвенци које су присутне у укупним узорцима из Србије у односу на укупан број узорак за сваку од остале 4 земаље. Табелу приказати по (претходно одређеним) протеинима.
5. За узорке из периода 1.3.2020-1.4.2021. године и период 1.12.2022-1.1.2023. одредити који најдужи низ нуклеотида је идентичан за узорке из Србије и Јапана и којим протеинима припадају пронађени низови

Задатак одабрали: Филип Огрењац (275/2019) и Мина Кованџић (127/2019)

Задатак 13

Скуп улазних података садржи

- секвенце SARS-CoV-2 вируса и одговарајуће метаподатке издвојене из базе GISAID (https://gisaid.org/) уз коришћење опција *Complete* и *High Coverage*. Секвенце изабрати из узорака са територије следећих земаља: Аустрије, Немачке, Швајцарске, Шведске и Србије.
- референтну секвенцу SARS-CoV-2 вируса (NCBI идентификација NC_045512.2) и одговарајуће метаподатке.

1. Поравнати нуклеотидне секвенце у односу на референтни изолат. За тако поравнате секвенце одредити могуће границе протеина у секвенцама издвојеним из GISAID базе.
2. Одредити проценат идентичних секвенци за сваки месец почевши од марта 2020. до јануара 2023. године, за сваки могући пар земаља (Аустрија-Немачка, Аустрија-Швајцарска, Аустрија-Шведска, ...).
3. Одредити проценат различитих секвенци за сваки месец почевши од марта 2020. до јануара 2023. године, за сваки могући пар земаља (Аустрија-Немачка, Аустрија-Швајцарска, Аустрија-Шведска, ...) у зависности од броја позиција на којима се налазе различити нуклеотиди. Табелу приказати по (претходно одређеним) протеинима.
4. Одредити проценат уникатних секвенци које су присутне у укупним узорцима из Србије у односу на укупан број узорак за сваку од остале 4 земаље. Табелу приказати по (претходно одређеним) протеинима.
5. Одредити 3 позиције у геному које су најдужи временски интервал остале идентичне у узорцима из Аустрије и Србије.

Задатак одабрали: Владимир Кнежевић (206/2017) и Ивана Вучковић (197/2019)

Задатак 14

Скуп улазних података садржи

- секвенце SARS-CoV-2 вируса и одговарајуће метаподатке издвојене из базе GISAID (https://gisaid.org/) уз коришћење опција *Complete* и *High Coverage*. Секвенце изабрати из периода 1.2.2020 - 28.2.2020. године (група А) и секвенце са територије Србије за период 1.3.2020 - 31.3.2020. године (група Б)
- референтну секвенцу SARS-CoV-2 вируса (NCBI идентификација NC_045512.2) и одговарајуће метаподатке.

1. Одредити број секвенци из групе А и из које земље потичу, а који су идентичне са неком од секвенци из групе Б, и
2. Издвојити 10 секвенци из групе А које се највише разликују у односу на групу Б.
3. Поравнати нуклеотидне секвенце у односу на референтни изолат. За тако поравнате секвенце одредити 10 позиција у групи А које су најчешће мутиране. Учесталост мутације одредити као проценат замена у односи на референтни изолат на конкретној позицији.
4. Одредити најдужи низ нуклеотида који је идентичан за узорке из група А и Б.

Задатак одабрали:

Задатак 15

Скуп улазних података садржи податке из *PAIDB* базе (http://www.paidb.re.kr/about_paidb.php) о патогеним острвима бактерије *Escherichia coli* : http://www.paidb.re.kr/browse_genomes.php?m=g#Escherichia%20coli.

1. Анализирати одступања појаве сва 4 типа понављајућих секвенци у острвима у односу на остатак кода користећи програм StatRepeats из пакета
   http://bioinfo.matf.bg.ac.rs/home/downloads.wafl?cat=Software&project=RepeatsPlus
2. Секвенце *Escherichia coli*
3. На основу анализе направити модел који предвиђа да ли је нека *Escherichia coli* бактерија патогена или не
4. За избор скупа бактерија за формирање модела и његово тестирање обавезно се консултовати са наставником.

Задатак одабрали: Немања Ршумовић (91/2020) и Стефан Митровић (350/2020)

Задатак 16

Скуп улазних података садржи податке из *PAIDB* базе (http://www.paidb.re.kr/about_paidb.php) о патогеним острвима бактерије *Escherichia coli* : http://www.paidb.re.kr/browse_genomes.php?m=g#Escherichia%20coli.

1. Анализирати разлику у употреби кодона у комплетној секвенци *Escherichia coli* и патогеним острвима. За контролу рада користити пресликавање кодона приказано на https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi (за бактерије користити transl_table=11)
2. На основу анализе направити модел који предвиђа да ли је нека *Escherichia coli* бактерија патогена или не
3. За избор скупа бактерија за формирање модела и његово тестирање обавезно се консултовати са наставником.

Задатак одабрали: Михајло Дедић (406/2021) и Анђелија Васиљевић (222/2020)

Задатак 17

Анализирати утицај P-adičnosti на разлике генетског кода SARS-1, MERS, BCOV, Human coronavirus 229E i Human coronavirus OC43 коронавируса. Референце о P-adičnosti - радове проф. Бранка Драговића скинути са мреже или преузети од наставника. Геноме коронавируса скинути са

- SARS1 (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:2901879)
- MERS (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:1335626)
- BCOV (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:11128)
- Human coronavirus 229E (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:11137)
- Human coronavirus OC43 (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:31631)

Употребу кодона преузети са https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi

Нека су

$$a = a_1a_2a_3a_4a_5a_6 \ ... \ a_{3n+1}a_{3n+2}a_{3n+3}$$

и

$$b = b_1b_2b_3b_4b_5b_6 \ ... \ b_{3n+1}b_{3n+2}b_{3n+3}$$

две ниске РНК секвенци a i b sa $n+1$ кодона, где је $n = 0, 1, 2, 3....$

За *surface glycoprotein* секвенце једнаких дужина

- Израчунати растојање између a и b преко 5-адичносг растојања по кодонима

$$d_5(a,b) = |a_1a_2a_3 - b_1b_2b_3|_5 + |a_4a_5a_6 - b_4b_5b_6|_5 + \ ... \ + |a_{3n+1}a_{3n+2}a_{3n+3} - b_{3n+1}b_{3n+2}b_{3n+3}|_5$$

- Израчунати Хамингово растојање $d_H(a,b)$ између РНК a и b, где су елементи кодони, не нуклеотиди.
- Израчунати Хамингово растојање $d_H(a,b)$ између одговарајућих протеина који кодирају РНК a и b
- Упоредити резултате за добијена растојања $d_5(a,b)$, $d_H(a,b)$ и $D_H(a,b)$.

Израчунати Едит растојање $d_H(a,b)$ између нуклеотидних секвенци свака два протеина и на основу израчунатог растојања извршити сакупљајуће хијерархијско кластеровање. За *surface glycoprotein* секвенце једнаких дужина поредити резултате добијене кластеровањем са резултатима добијеним коришћењем 5-адичности.

Задатак одабрали: Јелисавета Гавриловић (188/2020) и Марко Пауновић (104/2020)

Задатак 18

Скуп улазних података садржи секвенце SARS-CoV-2 вируса и одговарајуће метаподатке издвојене из базе GISAID (https://gisaid.org/) уз коришћење опција *Complete* и *High Coverage*. Секвенце изабрати из узорака са територије:

- Комплетан скуп узорака из Србије, Босне и Херцеговине, Црне Горе, Северне Македоније, Мађарске
- Узорке из главних градова следећих земаља: Хрватске, Бугарске, Румуније, Аустрије, Немачке, Швајцарске и Италије.
- референтну секвенцу SARS-CoV-2 вируса (NCBI идентификација NC_045512.2) и одговарајуће метаподатке.

1. Поравнати нуклеотидне секвенце у односу на референтни изолат. За тако поравнате секвенце одредити могуће границе S-протеина (*surface glycoprotein*) у секвенцама издвојеним из GISAID базе.
2. Одредити RSCU карактеристику употребе кодона у пронађеним протеинима. Употребу кодона преузети са https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi. За SARS2 користити стандардни код (transl_table=1)
3. Кластеровати са бар два различита алгоритма добијени скуп протеина на основу одређених RSCU вредности.
4. Анализирати добијене резултате.

Обавезно се консултовати са наставником после издвајања скупа узорака, пре почетка поравнања са референтним изолатом.

Задатак одабрали:

Задатак 19

На основу фотографија Меланом канцер оболења направити класификациони модел за његово предвиђање. Класификациони модел направити на основу бар два различита алгоритма од којих бар један не користи неуронске мреже. Упутство: Погледати поступак описан у раду *Melanoma Classification Using a Novel Deep Convolutional Neural Network with Dermoscopic Images* аутора *Ranpreet Kaur, Hamid Gholam-Hosseini, Roopak Sinha, and Maria Lindén* објављеног у *Sensors (Basel)* 2022 Feb 2;22(3):1134. Рад може да се преузме са линка *doi: https://doi.org/10.3390/s22031134*

Обавезо се консултовати са наставником око избора материјала за тренинг и тест податке.

Задатак одабрали: Игор Золотарев (228/2019) и Бранко Грбић (2/2020)

Задатак 20

На адреси https://www.inf.ed.ac.uk/teaching/courses/dme/html/datasets0405.html се налазе подаци који садрже каталог космичких објеката (*The SuperCOSMOS Sky Survey objects catalogue*). Направити класификациони модел користећи бар два различита алгоритма који класификује звездане објекте на звезде или галаксије.

vspace2ex

Задатак одабрали: Алекса Тороман (84/2018) и Милица Сударˇ (79/2017)

Задатак 21

(Задатак за љубитеље научне фантастике)

Сакупити слике ванземаљаца + људи који се појављују у епизодама и филмовима Звезданих стаза и направити класификациони модел који на основу унете слике врши класификацију врсте ванземаљаца+људи.

Упутство: погледати садржај странице https://en.wikipedia.org/wiki/List_of_Star_Trek_aliens

Задатак одабрали:

Задатак 22

Сакупити слике планета и сателита у Сунчевом систему и на основу њих направити класфиикациони модел који на основу унете слике врши класификацију унетог објекта на планета (име) или сателит (име).

Упутство: за списак сателита и планета у Сунчевом систему погледати садржаје страница

- https://photojournal.jpl.nasa.gov/
- https://www.nasa.gov/wp-content/uploads/2009/12/solar_system_moons_lithograph.pdf
- https://ares.jsc.nasa.gov/interaction/lmdp/documents/solar_system_lithograph_set_h.pdf

Задатак одабрали: Тодор Тодоровић (241/2019(и Петра Игњатовић (63/2020)

Задатак 23

На основу садржаја базе STUD2020 пронаћи правила придруживања која говоре када је најбоље полагати поједини испит (у ком року, која оцена се добија, ...)

Задатак одабрале: Јелена Митровић (357/2020) и Ана Величковић (170/2019)

Задатак 24

База података садржи податке о 75.000 рачуна везаних за продају производа у једном ланцу пекара. Потребно је извршити анализу података и одредити

- Који артикли се продају заједно. Да ли то важи за све продавнице без обзира на локацију?
- Којим артиклима треба спустити цену да би се повећала укупна зарада при продаји?
- Који продавац је имао највећи промет?
- Који су најбољи делови дана за продају (тј. у ком делу дана је био највећи приход)?
- Сва негативно корелисана правила придруживања

Податке и опис материјала преузети од наставника.

Задатак одабрали: Емилија Стевановић (100/2019) и Ива Читлучанин (143/2019)

Задатак 25

База података садржи податке о 75.000 рачуна везаних за продају производа у једном ланцу пекара. Потребно је извршити анализу података:

- Излистати правила придруживања која имају већу подршку од 50%.
- Излистати купце чији низ куповина садржи куповине код којих је максимални јаз између ставки мањи или једнак 3 дана (без обзира на локацију продаје).
- Приказати низ куповина код којих је максимални јаз између ставки мањи или једнак 2 дана за најмање 5 купаца (без обзира на локацију продаје)
- Приказати парове куповина ако је једна куповина непрекидна подниска друге (посматрано у односу на ставке у тим куповинама).

Податке и опис материјала преузети од наставника.

Задатак одабрали: Лука Бура (218/2019) и Лука Вукотић (120/2019)

Задатак 26

Написати програм који

- За унети скуп трансакција и минималну подршку црта решетку, и одређује који су скупови ставки затворени, максимални, затворени и максимални.
- За унети скуп трансакција и минималну подршку црта, корак по корак, конструкцију ФП-дрвета и одређивање честих скупова ставки.

Задатак одабрали: Алекса Костур (44/2018) и Немања Јанковић (213/2018)

Примедба: задатак може да се подели на два дела, као два задатка за по једног студента.

Задатак 27

Улазне датотеке садрже податке добијене из периферних мононуклеарних крвних ћелија (Peripheral blood mononuclear cells, PBMCs). ПБМЦ ћелије укључују ћелије различитих типова: лимфоците (B ћелије, T ћелије, NK ћелије ("natural killer – cells"), моноците и дендритске ћелије. ПБМЦ ћелије се користе у истраживању у различитим областима биомедицине, укључујући инфективне болести, имунологију (укључујући и аутомуне поремећаје), малигнитет, трансплантациону имунлогију, развој вакцина, и скрининг. Мада могу да имају различите функције, главна функција ПБМЦ ћелија је имуна одбрана организма. Сваки тип ћелије има карактеристичне обрасце ('мустре') протеина и гена које их међусобно разликују и могу да се користе за поделу према њиховом типу.

Улазни материјал садржи резултате истраживања скупова различитих хуманих ћелија. Свака од 4 улазне датотеке садржи податке везане за експресију 10800 гена. Скуп гена који су посматрани је идентичан у свим датотекама, и у свим датотекама је уређен по истом редоследу. Називи свих гена имају префикс hg38 који означава да су подаци везани за верзију 38 хуманог генома. Датотеке представљају различите типове ћелија и добијени су из различитих извора. Улазна датотека садржи податке у ЦСВ формату где је прва колона назив изворне датотеке која садржи тип ћелије.

Задатак је извршити кластеровање са бар три различита алгоритма сваке од појединачних датотека (по извору података), и утврдити да ли добијени кластери одговарају типовима ћелија. Унакрсно применити формиране моделе за кластеровање на остала три датотеке и поновити анализу квалитета. При кластеровању не примењивати методе димензионе редукције које уводе нове атрибуте (као нпр. PCA).

Подаци се преузимају од наставника.

Задатак одабрали:

Задатак 28

Улазне датотеке садрже податке добијене из периферних мононуклеарних крвних ћелија (Peripheral blood mononuclear cells, PBMCs). ПБМЦ ћелије укључују ћелије различитих типова: лимфоците (B ћелије, T ћелије, NK ћелије ("natural killer – cells"), моноците и дендритске ћелије. ПБМЦ ћелије се користе у истраживању у различитим областима биомедицине, укључујући инфективне болести, имунологију (укључујући и аутомуне поремећаје), малигнитет, трансплантациону имунлогију, развој вакцина, и скрининг. Мада могу да имају различите функције, главна функција ПБМЦ ћелија је имуна одбрана организма. Сваки тип ћелије има карактеристичне обрасце ('мустре') протеина и гена које их међусобно разликују и могу да се користе за поделу према њиховом типу.

Улазни материјал садржи резултате истраживања скупова различитих хуманих ћелија. Свака од 4 улазне датотеке садржи податке везане за експресију 10800 гена. Скуп гена који су посматрани је идентичан у свим датотекама, и у свим датотекама је уређен по истом редоследу. Називи свих гена имају префикс hg38 који означава да су подаци везани за верзију 38 хуманог генома. Датотеке представљају различите типове ћелија и добијени су из различитих извора. Улазна датотека садржи податке у ЦСВ формату где је прва колона назив изворне датотеке која садржи тип ћелије.

Задатак је извршити класификацију са бар три различита алгоритма сваке од појединачних датотека (по извору података), и утврдити да ли добијени модели добро предвиђају типове ћелија. Улазне податке поделити у односу 70:30 за тест и трениг део, односно 50:30:20 (тест, тренинг и провера) ако алгоритам омогућава рад са три дела улазних података. Унакрсно применити формиране моделе за класификацију на остала три датотеке и поновити анализу квалитета. При класификацији не примењивати методе димензионе редукције које уводе нове атрибуте (као нпр. PCA).

Подаци се преузимају од наставника.

Задатак одабрали: Стефан Керкоч (28/2020) и Богдан Стојадиновић (73/2020)

Задатак 29

Извршити анализу временских серија на:

- Податке о кретању курса динара у односу на Евро у периоду од 2002-2023. године
- Податке о просечној температури у Београду у периоду

Податке о кретњу курса преузети са сајта Нареодне банке СРбије, а податек о просечној темпрератури преузети од наставника.

Задатак одабрале: Ивана Нешковић (167/2019) и Марија Паповић (63/2019)

Задатак 30

За геноме коронавируса

- SARS1
- MERS
- BCOV
- BAT SARS

1. одредити директне и индиректне некомплементарне понављајуће секвенце у аминокиселинским секвенцама користећи програм StatRepeats из пакета
   http://bioinfo.matf.bg.ac.rs/home/downloads.wafl?cat=Software&project=RepeatsPlus
2. на основу њих направити, помоћу бар 3 различита алгоритма, класификациони модел за одређивање врсте коронавируса.

Податке преузети од наставника. Задатак одабрао: Александар Шмигић (28/2019)

Задатак 31

Улазне датотеке садрже податке добијене из периферних мононуклеарних крвних ћелија (Peripheral blood mononuclear cells, PBMCs). ПБМЦ ћелије укључују ћелије различитих типова: лимфоците (B ћелије, T ћелије, NK ћелије ("natural killer – cells"), моноците и дендритске ћелије. ПБМЦ ћелије се користе у истраживању у различитим областима биомедицине, укључујући инфективне болести, имунологију (укључујући и аутомуне поремећаје), малигнитет, трансплантациону имунлогију, развој вакцина, и скрининг. Мада могу да имају различите функције, главна функција ПБМЦ ћелија је имуна одбрана организма. Сваки тип ћелије има карактеристичне обрасце ('мустре') протеина и гена које их међусобно разликују и могу да се користе за поделу према њиховом типу.

Улазни материјал садржи резултате истраживања скупова различитих хуманих ћелија. Свака од 2 улазне датотеке садржи податке везане за експресију 10800 гена. Скуп гена који су посматрани је идентичан у свим датотекама, и у свим датотекама је уређен по истом редоследу. Називи свих гена имају префикс hg38 који означава да су подаци везани за верзију 38 хуманог генома. Датотеке представљају различите типове ћелија и добијени су из различитих извора. Улазна датотека садржи податке у CSV формату где је прва колона назив изворне датотеке која садржи тип ћелије.

Задатак је извршити кластеровање са бар три различита алгоритма сваке од појединачних датотека (по извору података), и утврдити да ли добијени кластери одговарају типовима ћелија. Унакрсно применити формиране моделе за кластеровање на другој (преосталој) датотеци и поновити анализу квалитета. При кластеровању не примењивати методе димензионе редукције које уводе нове атрибуте (као нпр. PCA).

Подаци се преузимају од наставника.

Задатак одабрала: Тамара Миковић (82/2017)