



---

Семинарски рад из  
Истраживања Података 2

Задатак број 12

---

Професор: Ненад Митић

Мина Кованцић

Филип Огрењац

јун 2024.

# Садржај

<b>Садржај</b>	<b>1</b>
<b>1 Увод</b>	<b>2</b>
1.1 Скуп података . . . . .	3
1.2 Претпроцесирање . . . . .	3
<b>2 Поравнање нуклеотидне секвенце у односу на референтни изолат</b>	<b>5</b>
<b>3 Проценат идентичних секвенци</b>	<b>6</b>
<b>4 Проценат различитих секвенци</b>	<b>7</b>
<b>5 Проценат уникатних секвенци на територији Србије</b>	<b>18</b>
<b>6 Најдужи низ нуклеотида идентичан за узорке из Србије и Јапана</b>	<b>21</b>
<b>7 Закључак</b>	<b>22</b>
<b>Референце</b>	<b>23</b>

# 1 УВОД

**SARS-Cov-2**, скраћено од **Severe Acute Respiratory Syndrome Coronavirus 2** спада у групу коронавируса познатих по способности да изазивају респираторне болести како код људи, тако и код животиња. Циљ овог истраживања је анализа сличности и разлика секвенци **SARS-Cov-2** вируса из узорака прикупљених у неколико држава у периоду између 2020. и 2022. године.

Уводни део фокусиран је на анализу скупа улазних података који садржи две главне компоненте. Прва компонента обухвата секвенце SARS-Cov-2 вируса издвојене из базе GISAID [1] са опцијама *Complete* и *High Coverage*. Ове секвенце потичу из узорака прикупљених са територија следећих земаља: Мађарска, Хрватска, Србија, Јапан и Грчка. Друга компонента укључује референтну секвенцу SARS-Cov-2 вируса (NCBI идентификација NC\_045512.2 [2]), заједно са одговарајућим метаподацима.

Током израде рада коришћени су следећи алати:

1. **MySQL**: Коришћен је за складиштење и управљање великим скуповима података о секвенцима. Помоћу MySQL-а креиране су табеле, извршени сложени упити и дефинисане процедуре потребне за анализу података.
2. **Python**: Коришћен је за писање скрипти које аутоматски уносе податке у базе, обрађују секвенце и извршавају сложене анализе. Скрипте попут `load_sequence.py` омогућиле су ефикасну обраду података и уклапање са MySQL базом.
3. **Shell skripte**: Shell скрипте су коришћене за аутоматизацију задатака на нивоу оперативног система. Скрипте попут `poravnanje.sh` олакшале су извршавање секвенцијалних команда, покретање других програма (као што је Clustal Omega) и управљање датотекама, чиме је целокупан процес био бржи и ефикаснији.
4. **Clustal Omega**: Clustal Omega је моћан алат за вишеструко поравнање секвенци. Коришћен је за поравнање великог броја протеинских или нуклеотидних секвенци како би се идентификовале сличности и разлике међу њима. Омогућава прецизно и брзо поравнање.

Заједно, ови алати су омогућили ефикасно прикупљање, обраду, анализу и визуализацију биолошких података, чиме су значајно допринели успешном спровођењу нашег истраживања.

## 1.1 Скуп података

Скуп улазних података груписан је у **FASTA** фајлове за сваку од држава, што омогућава једноставан и организован приступ генетским подацима са територије различитих земаља. На [слици 1.1. \(а\)](#) налази се пример улазног **FASTA** фајла који у првој линији садржи додатне информације о времену и локацији где је прикупљен узорак. Референтна секвенца ([слика 1.1. \(б\)](#)) представља **Wuhan-Hu-1** изолат коронавируса у односу на који ће се вршити даље поређење секвенци.

```
>hCoV-19/Serbia/Novi_Sad-NIVNS013007/2020|EPI_ISL_833572|2020-07-29
AGATCTTCTCTAAACGAACTTAAAATCTGTGCGCTGCACTCGGCTGCATGCTAGTCACGCAGTATAATT
ATAACTAATTACTGTGCGTTGACAGGACACGAGTAACCTCGTCTATCTTCGAGGCTGCTTACGGTTCTGCC
AGCCGATCATCAGCACATCTAGGTTTGCCGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTCCCTGGTT
GAAACACACGTCAAACCTAGTTGCGCTGTTTACAGGTCGCGACGTGCTCGTACGTGGCTTGGAGACTCGT
GGAGGTCTTATCAGAGGCACGTCAACATCTAAAGATGGCACTTGTGGCTTAGTAGAAAGTTGAAAAAAGGC
GTTGCACAGCCCTATGTGTTCATCAAACGTTGGATGCTCGAACCTCATGGTCATGTTATGGTTGAGCTGG
AGAACCTGAAGGCATTCACTAGTACGGTGTAGTGGTGGAGACACTTGTGTCCTTGTCCCTCATGTGGCGA
AAATACCAGTGGCTTACGGCAAGGTTCTTCTCGTAAGAACGGAATAAAGGAGCTGGTGGCCATAGTTACGG
CGCCGATCTAAAGTCATTGAGCTTGCAGCTGATCCTTATGAAGATTTCAGAAACACTGAAACACTAAACAT
AGCAGTGGTGTACCGTGAACCTTCTAGCACGTGCTGGTAAGCTTCACTGACTTTGTCGAAACAACGTTG
ACTAAGAGGGGTGATACTGTCGCGTGAACATGAGCATGAAATTGCTGGTACACGGAAACGTTGAAAGAGCT
ATTGCAAGACACCTTTGAAATTAAATTGGCAAGAAATTGACACCTTCAATGGGAATGTCCAATTGGTATT
TAAATTCCATAATCAAGACTATTCAACCAAGGGTGGAAAGAAAAGCTTGTGATGGCTTATGGTAGAATTG
CATCTGTCATGGTGAAGTGTGATCATTGTGGTGAACACTTCAGTGGTGAAGTGTGATCATTGTGGTGAAC
ACTTCAGTGGTGCACCAAATGAATGCAACCAAATGTGCCCTTCACTCTCATGAAGTGTGATCATTGTGGT
GAACACTTC
```

(a) Секвенца

```
>NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome
ATTAAGGTTTACCTTCCAGTAACAAACCAACCTTCGATCTCTTGAGATCTGTTCTCTAA
CGAACCTTAAATCTGTGGCTGTCACTCGGCTGCATGCTAGTCACGCAGTATAATTAAAC
TAATTACTGTGCGTTGACAGGACACGAGTAACCTCGTCTATCTTCGAGGCTGCTTACGGT
GGCTCGTCAACGAGAAAACACACGTCAAACCTAGTTGCGCTGTTTACAGGTCGCGACGTGCTCGTAC
GTGGCTTGGAGACTCGTGGAGGAGCTTACAGAGGCACGTCAACATCTAAAGATGGCACTTGTG
CTTAGTAGAAAGTTGAAAGGCGCTTGGCTCAACTTGAACAGCCCTATGTGTTCATCAAACGTTGG
GCTCGAACCTGACCTCATGGTATGTTGAGCTGGTAGCAGAAACTCGAAGGCATTCACTGGC
GTAGTGGTAGACACTTGGTGTCTTGTCCCTCATGGGGAAATACAGTGGTACCGCAAGGTT
TCTTCGTAAGAACGGTATAAAAGGAGCTGGTGGCCATAGTTACGGCGCGATCTAAAGTCA
GGCGAGGCTGGCACTGTGCTTATGAAGATTTCAGAAAAGCTGGAACACTAAACATGAGTGG
TTACCCGTGAACTCATCGTGCTTAACGGAGGGCATAACACTCGCTATGTGCAATAACAACTTGTG
CCCTGATGGTACCCCTTGTGAGTCATTAAGACCTTCACTGAGCAGTGTGTTAAAGCTTCA
TGGTACCGAACCTGACTTATGACACTAAGAGGGGTGATACTGTCGCGTGAACATGAGCATGAA
ATTGACACCTTCATGGGAATGTCAAATTGATTTCCCTAAATTCCATAATCAAGACTATTCAA
CCAAGGGTGGAAAGAAAAGCTTGTGAGGGTAAATTGATGGTAGAATTGATCTGTCATCCAG
TTGCGTCACCAATGAATGCAACCAAATGTGCCCTTCACTCTCATGAAGTGTGATCATTGTGGT
GAACACTTC
```

(b) Референтна секвенца

Слика 1.1: Пример улазног скupa података

## 1.2 Претпроцесирање

Претпроцесирање података је кључни корак у обради и анализи података који обухвата припрему сирових података како би постали прикладни за даљу анализу или употребу. Овај процес је неопходан јер сирови подаци често садрже неправилности, недоследности или непотпуности које могу негативно утицати на резултате анализе. Припрема улазних података за даљу обраду се састоји из наредних фаза:

- 1. Уклањање дупликата:** Скрипта `sample_rmdup_modheader.py` ће прво уклонити све дупликате из датотеке геномске секвенце.
- 2. Узимање узорка од 1000 секвенци по држави:** Након уклањања дупликата, скрипта ће изабрати наслучично узорак од 1000 секвенци за сваку државу.
- 3. Подела у више фајлова по 200 секвенци са референтном секвенцом на врху:** Свака одабрана група од 200 секвенци биће сачувана у засебном фајлу. Референтна секвенца се поставља на почетак сваког фајла.
- 4. Измена заглавља секвенци:** Заглавља ће бити модификована тако да садрже идентификациони број секвенце, приступни идентификациони број секвенце и датум, раздвојени карактером '|'.

## 2 Поравнање нуклеотидне секвенце у односу на референтни изолат

Након претпроцесирања података, врши се један на један поравнање секвенци у односу на референтни изолат коришћењем алата **Clustal Omega**. Овај алат подржава велике скупове секвенци и омогућава њихово поравнање на рачунарима са различитим архитектурама.

У нашем истраживању, поравнање се врши помоћу скрипте `poravnanje.sh`. Након тога, резултати поравнања се уписују у табеле `poravnato_clob` и `poravnato_clob_ref`, како би се сачувале поравнате секвенце и њихове референтне секвенце.

Важно је да свака секвенца има свој редни број како би се успешно повезала са референтним изолатом. Ове информације се користе за упаривање граница кодирајућих региона. У фајлу `granice_proteini.txt` налазе се границе, имена и приступни бројеви за кодирајуће регионе референтне секвенце, који се уносе у табелу `granice_proteini` коришћењем скрипте `load_granice.py`.

Да би се ускладиле границе кодирајућих региона за референтну секвенцу са границама поравнатах секвенци, потребно је израчунати број карактера ' \_ ' (неусаглашених карактера) и померити границе за тај број карактера надесно. Ове прилагођене границе се затим уписују у табелу `protein_kod_sekv` коришћењем процедуре дефинисаних у скрипти `proteini.sql`.

Ова процедура омогућава ефикасно управљање биолошким подацима, омогућавајући истраживачима да анализирају и интерпретирају еволуцијске и структурне карактеристике протеинских секвенци са високим нивоом прецизности и поузданости.

```
266,13483,ORF1A,YP_009725295.1
266,21555,ORF1AB,YP_009724389.1
21563,25384,surface glycoprotein,YP_009724390.1
25393,26220,ORF3a,YP_009724391.1
26245,26472,envelope protein,YP_009724392.1
26523,27191,membrane glycoprotein,YP_009724393.1
27202,27387,ORF6,YP_009724394.1
27394,27759,ORF7a,YP_009724395.1
27756,27887,ORF7b,YP_009725318.1
27894,28259,ORF8,YP_009724396.1
28274,29533,nucleocapsid phosphoprotein,YP_009724397.2
29558,29674,ORF10,YP_009725255.1
```

Слика 2.1: Пример граница протеина

## 3 Проценат идентичних секвенци

У овом поглављу испитиваће се проценат идентичних секвенци за сваки месец, почевши од марта 2020. до јануара 2023. године, за сваки могући пар земаља (Мађарска-Хрватска, Грчка-Србија...).

За овај задатак коришћени су оригинални подаци (без уклањања дупликата). Коришћени фајлови су `sequence_table.sql`, `load_sequence.py` и `calc_identical.sql`. Направљене су све могуће комбинације месеци и година (од марта 2020. до јануара 2023.), као и парова земаља, како би се одредио проценат секвенци које се поклапају између тих земаља за дати временски период. Почетни подаци се уносе у табелу `sekvence` помоћу скрипте `load_sequence.py`, док се проценти поклапања рачунају помоћу SQL процедуре `Calc_identical` и уносе у табелу `procenat_identicnih`.

На основу резултата из табеле `percentage_results`, долазимо до закључка да је проценат идентичних секвенци већи од нуле у 3 случаја:

- **Јапан - Мађарска**, март 2020, проценат: 0.0313029
- **Јапан - Мађарска**, април 2020, проценат: 0.0011877
- **Грчка - Хрватска**, јул 2021, проценат: 0.0003878

## 4 Проценат различитих секвенци

У овом поглављу испитиваће се проценат различитих секвенци за сваки месец почевши од марта 2020. до јануара 2023. за сваки могући пар земаља (Мађарска-Хрватска, Грчка-Србија...) у зависности од броја позиција на којима се налазе различити нуклеотиди.

Како бисмо одредили све могуће комбинације месеци и година, као и комбинације држава и протеина, користили смо скрипте `kombinacije_zemalja.sql`, `razlicite_table.sql` и `razlicite.sql`. Првобитно се кроз процедуру `CalculateDifferent`, за сваку комбинацију земаља у датом временском периоду и за дати протеин израчунава број разлика и уписује у табелу `razlike_protein`. Након тога, коришћењем процедуре `CalculatePercentage`, за сваку комбинацију и израчунати број разлика, рачуна се проценат тог броја (у односу на све могуће вредности броја разлика за ту комбинацију) и уписује у табелу `percentage_results`.

Анализирањем графикона у наставку дошли смо до закључка:

- **Протеин са доминантним бројем позиција**

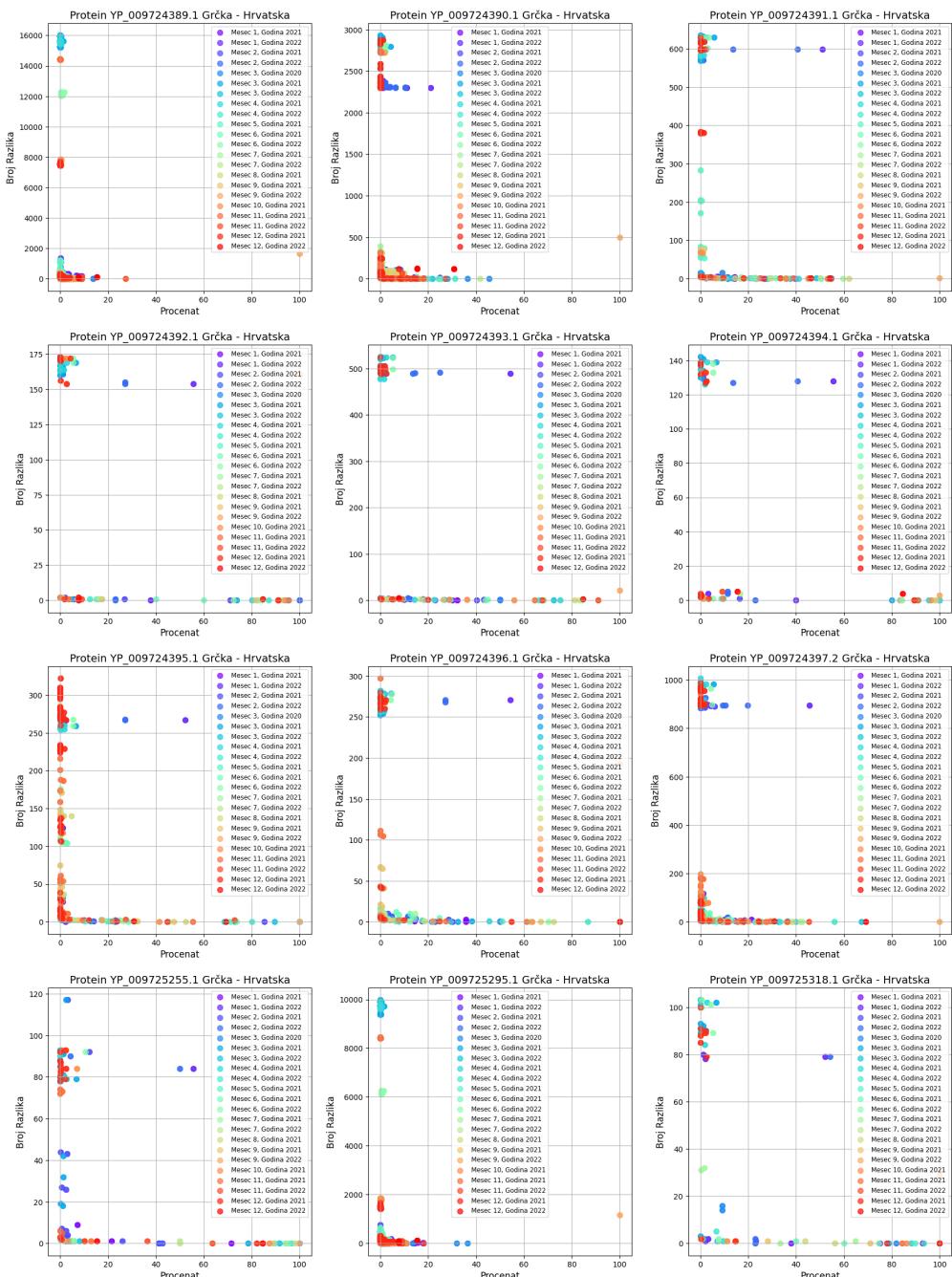
- Запажамо да број позиција на којима се налазе различити нуклеотиди варира и да се горња граница овог броја креће и до 16000 код протеина `YP_009724389.1`, што је забележено у марта 2021. и 2022. године, иако је сам проценат присуства изузетно низак.
- Затим следи протеин `YP_009725295.1` чија горња граница достиже 10000 различитих нуклеотида, али је та појава изузетно ретка (мање од 1%).
- Протеин `YP_009725318.1` такође бележи велику горњу границу (3000 различитих нуклеотида) уз мали број таквих случајева.
- За остале протеине је присутна знатно мања горња граница која се креће између 150 и 1000 различитих нуклеотида.

- **Варијабилност међу паровима држава**

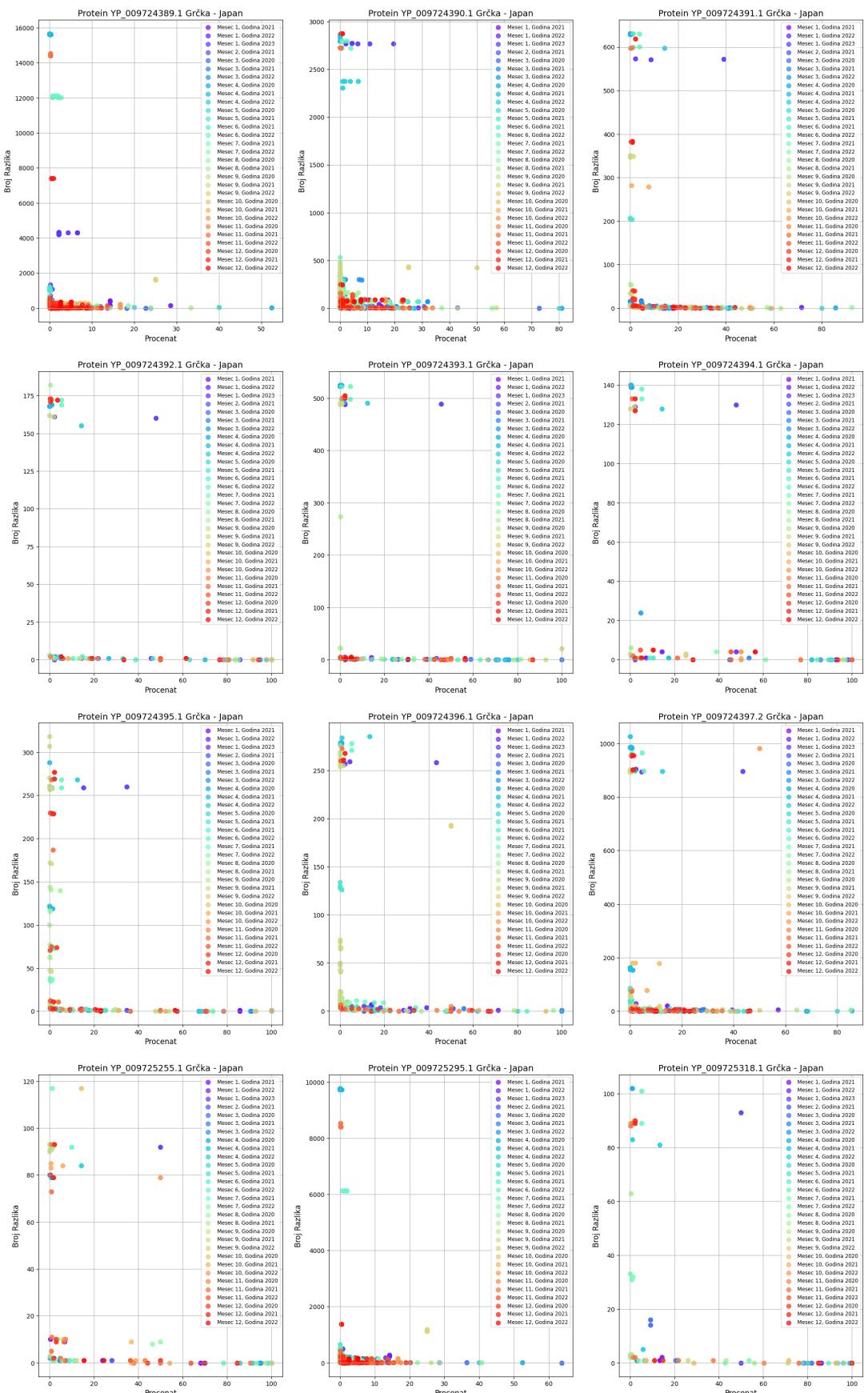
- Присутно је другачије понашање протеина `YP_009724389.1` у паровима држава Јапан - Србија и Мађарска - Србија. У односу на остале парове држава где је горња граница броја позиција на којима се налазе различити нуклеотиди 16000 са изузетно ниским процентом присуства, у случају два наведена пара држава, горња граница је близу 500 са најчешћим процентом појављивања између 3% и 15%.

- **Сличности између земаља**

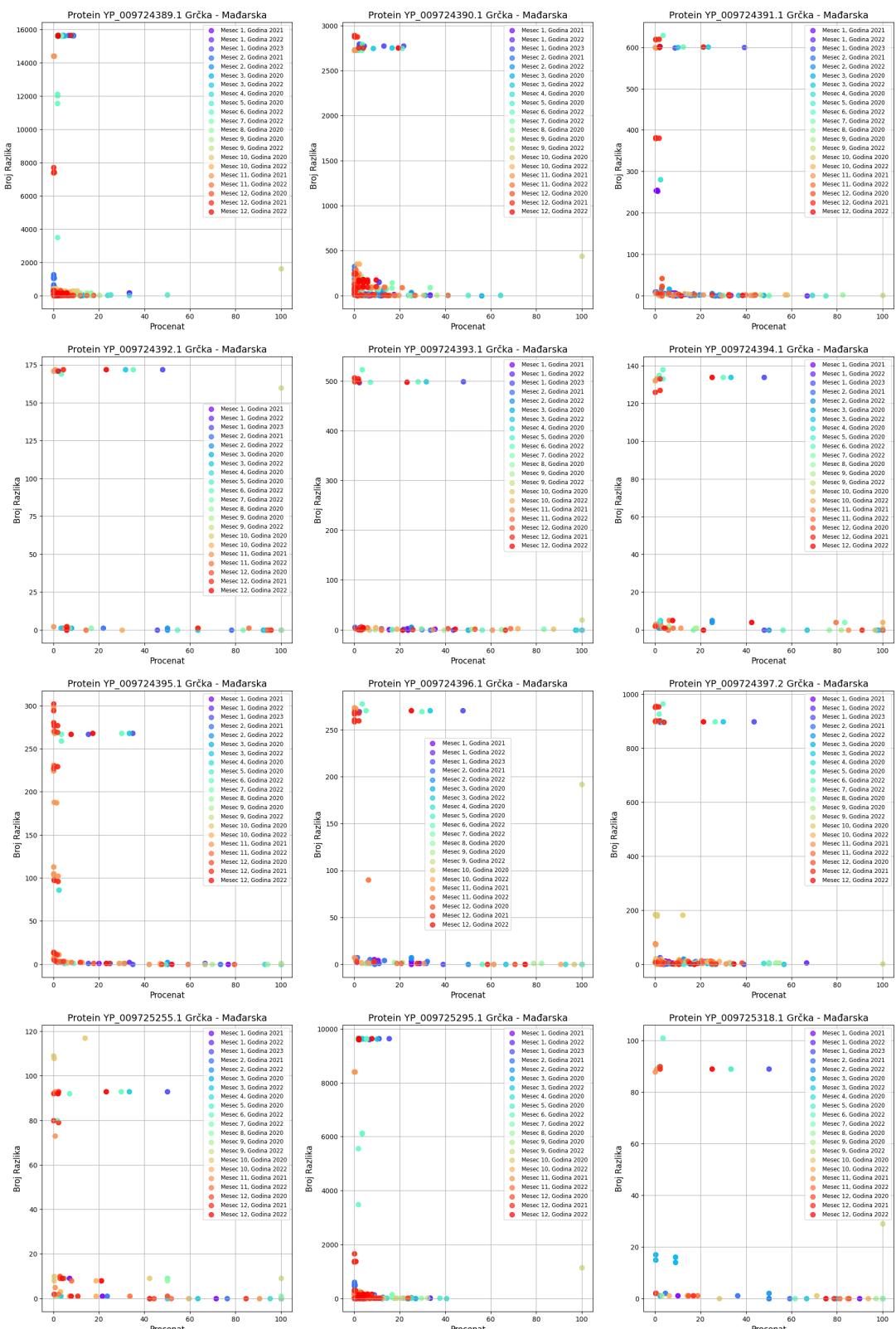
- Примећујемо да је код већине протеина независно од броја позиција на којима се налазе различити нуклеотиди, проценат присуства таквих секвенци најчешће мањи од 50% за различите месеце и године. У највећем броју случајева, број позиција је изузетно мали, при чему за протеине YP\_009724392.1, YP\_009724393.1, YP\_009724394.1, YP\_009725255.1 и YP\_009725318.1 важи да је тај број мањи од 15, независно од месеца и године.



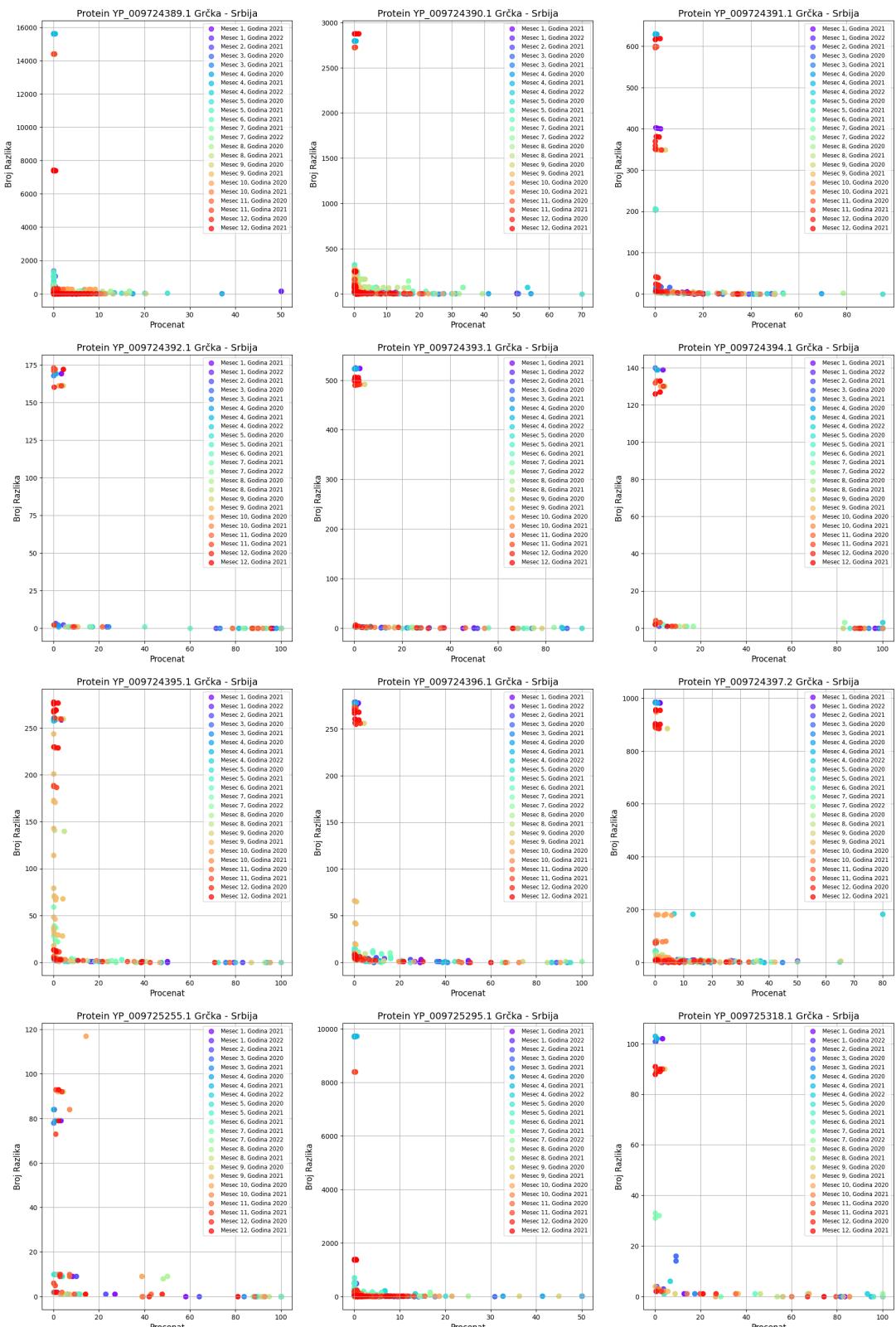
Слика 4.1: Визуелизација процента различитих секвенци за сваки протеин за пар држава: Грчка - Хрватска



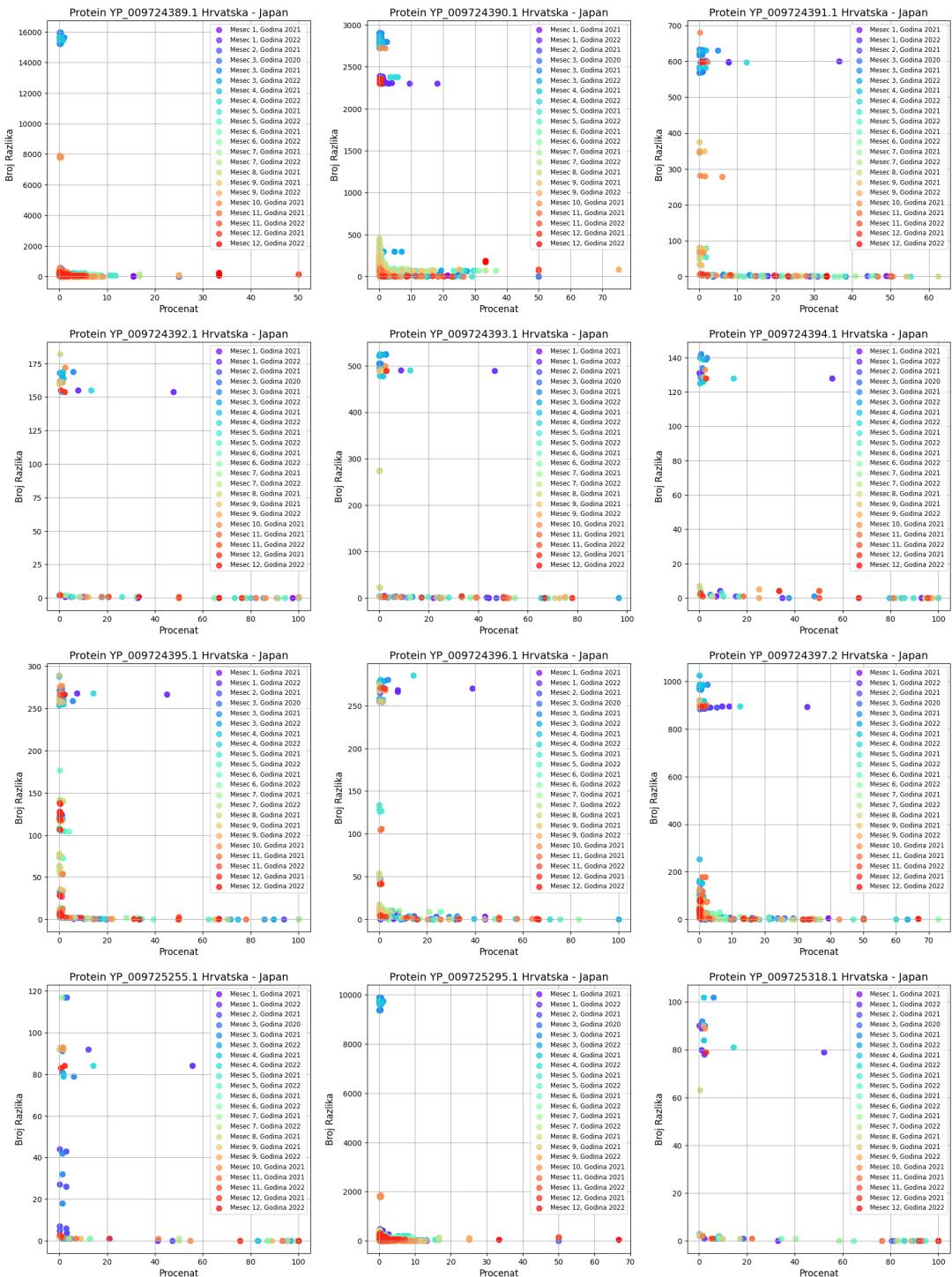
Слика 4.2: Визуелизација процента различитих секвенци за сваки протеин за пар држава:  
Грчка - Јапан



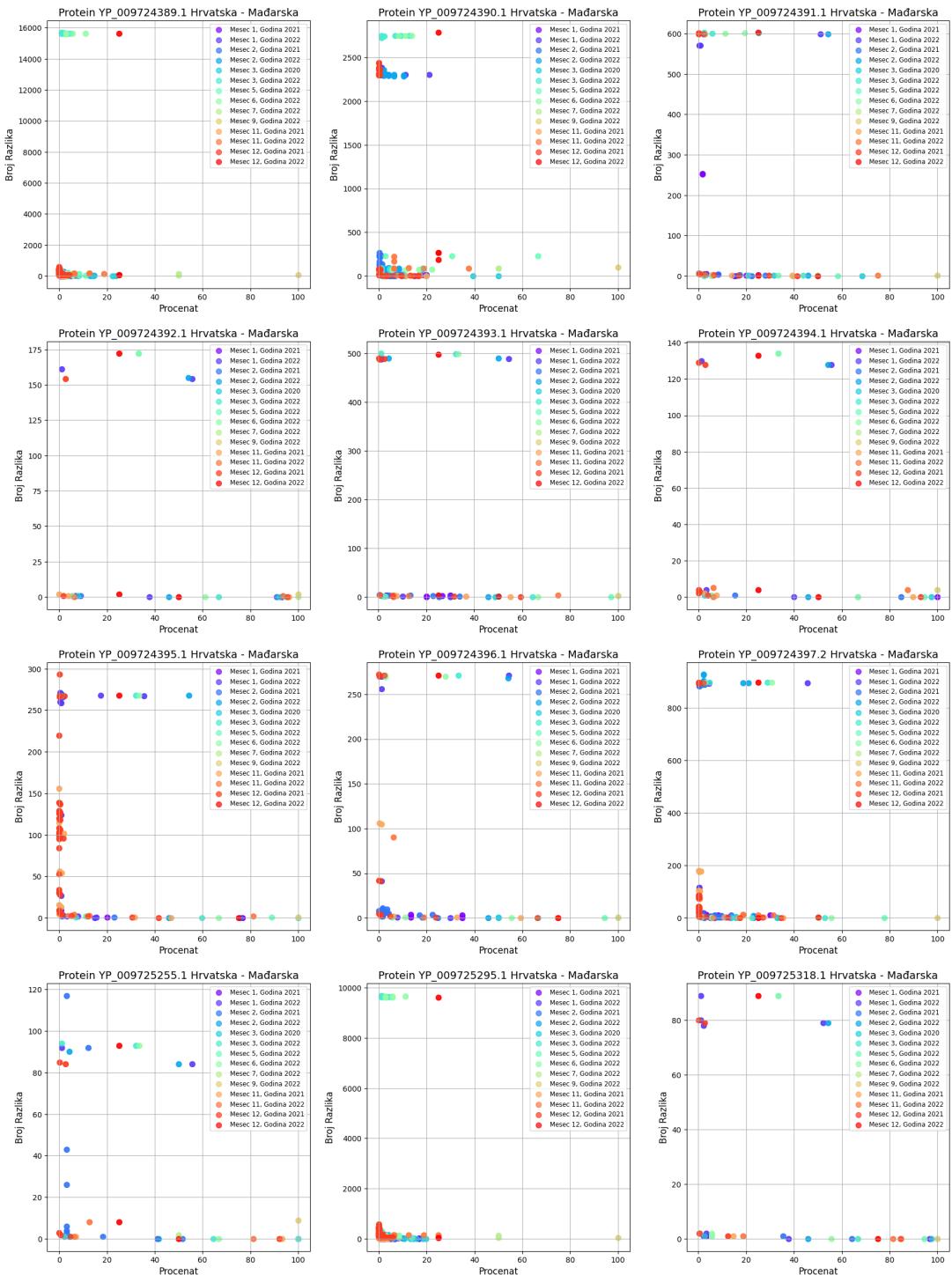
Слика 4.3: Визуелизација процента различитих секвенци за сваки протеин за пар држава: Грчка - Мађарска



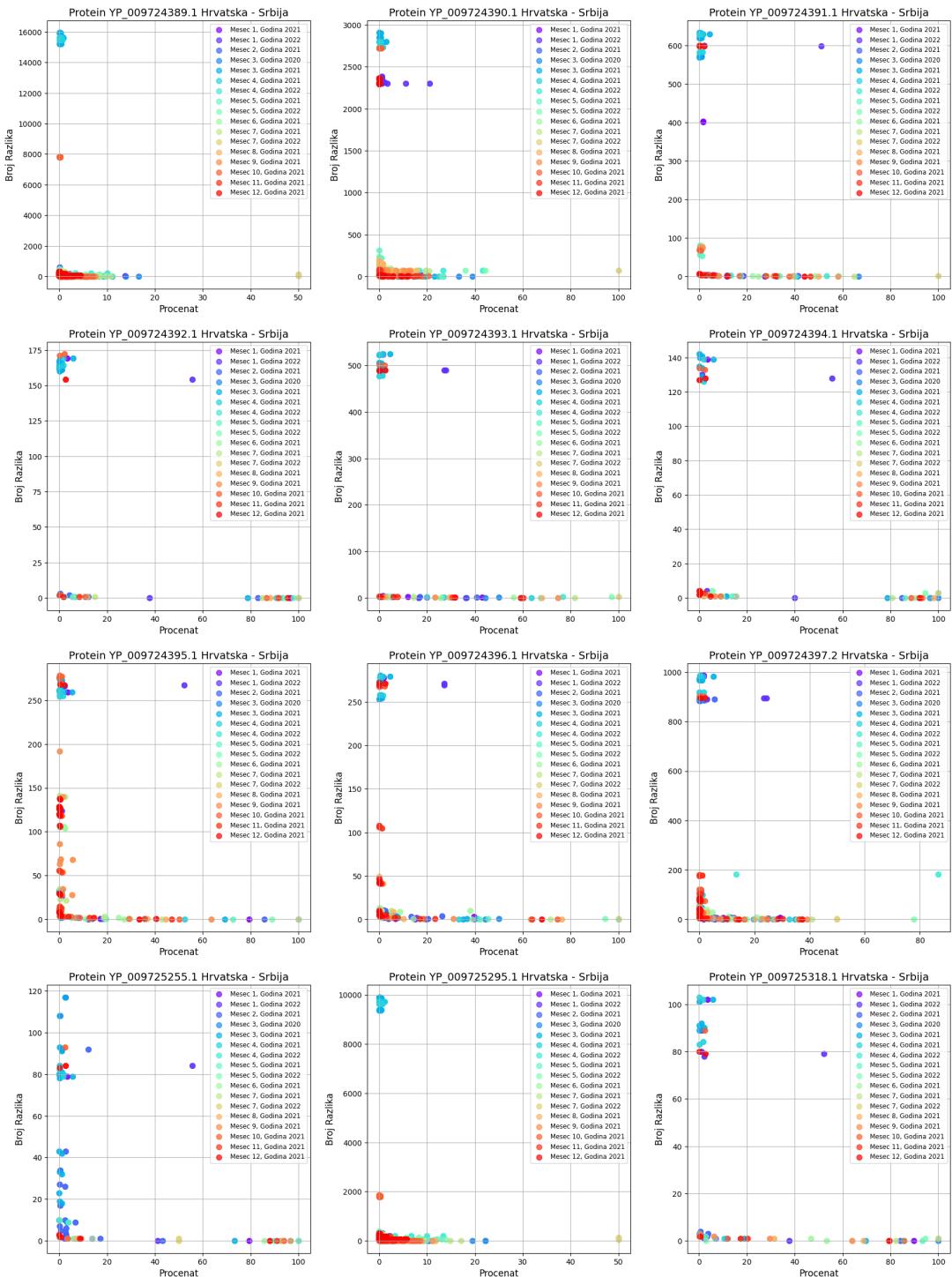
Слика 4.4: Визуелизација процента различитих секвенци за сваки протеин за пар држава: Грчка - Србија



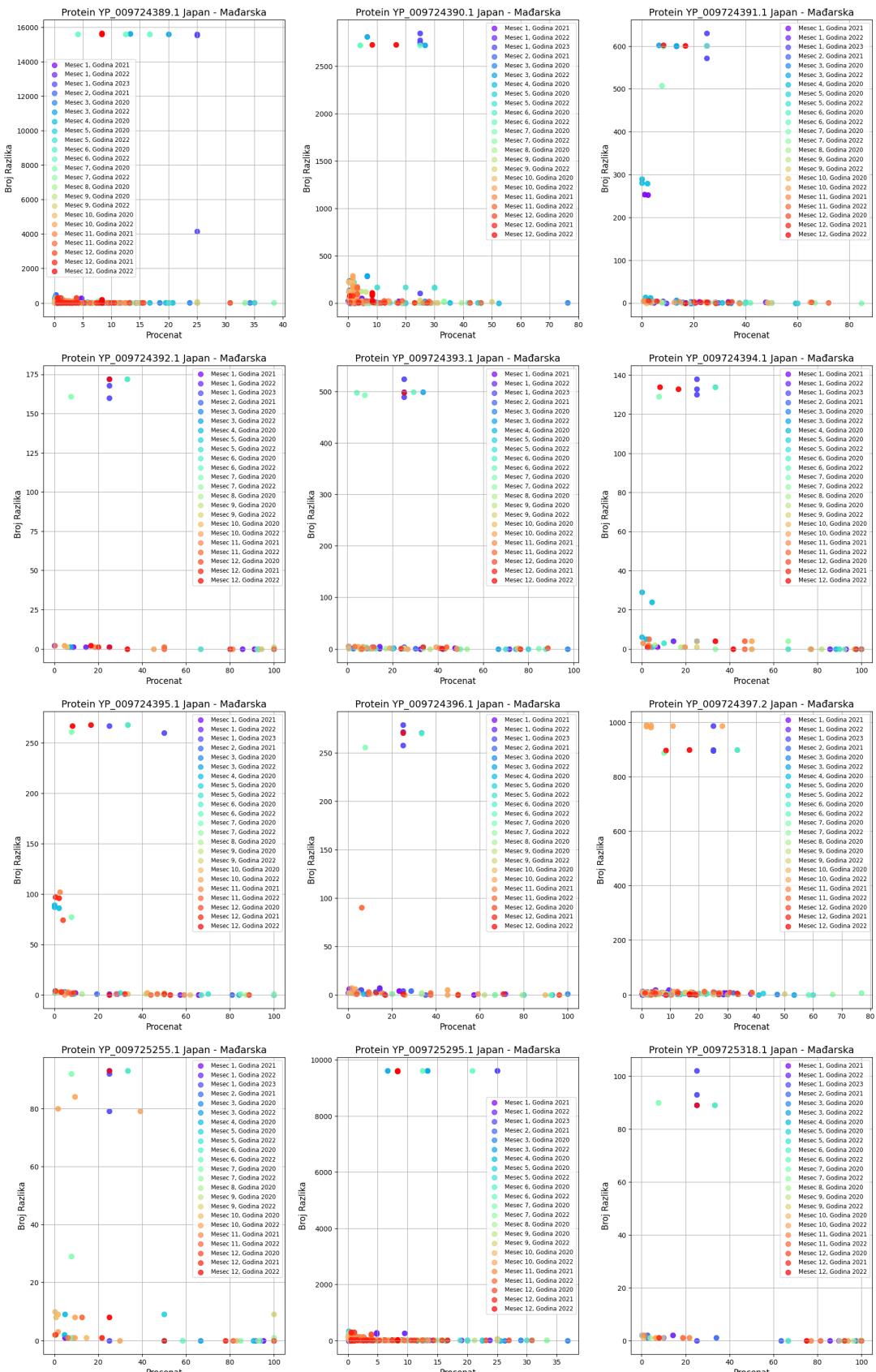
Слика 4.5: Визуелизација процента различитих секвенци за сваки протеин за пар држава: Хрватска - Јапан



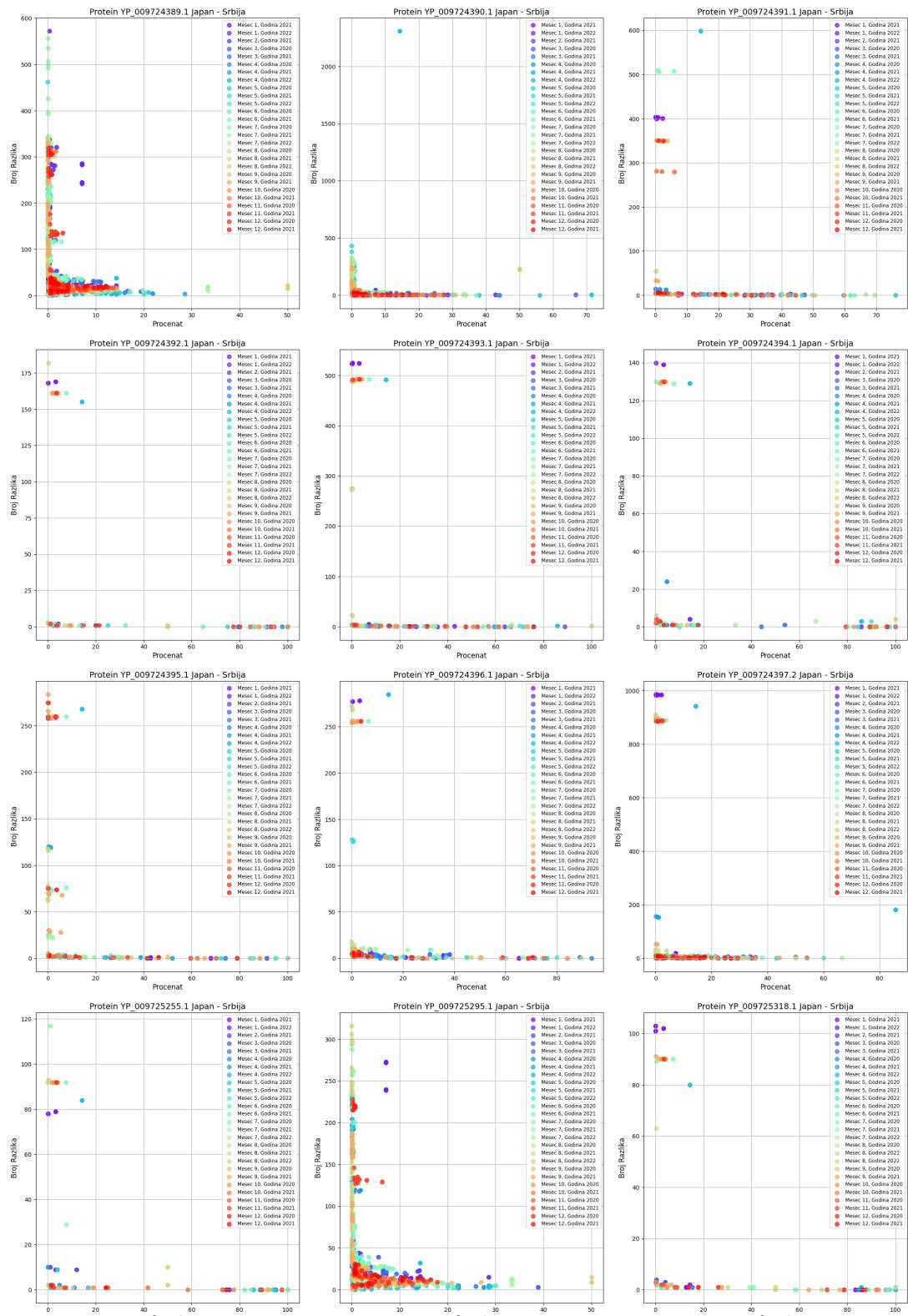
Слика 4.6: Визуелизација процента различитих секвенци за сваки протеин за пар држава: Хрватска - Мађарска



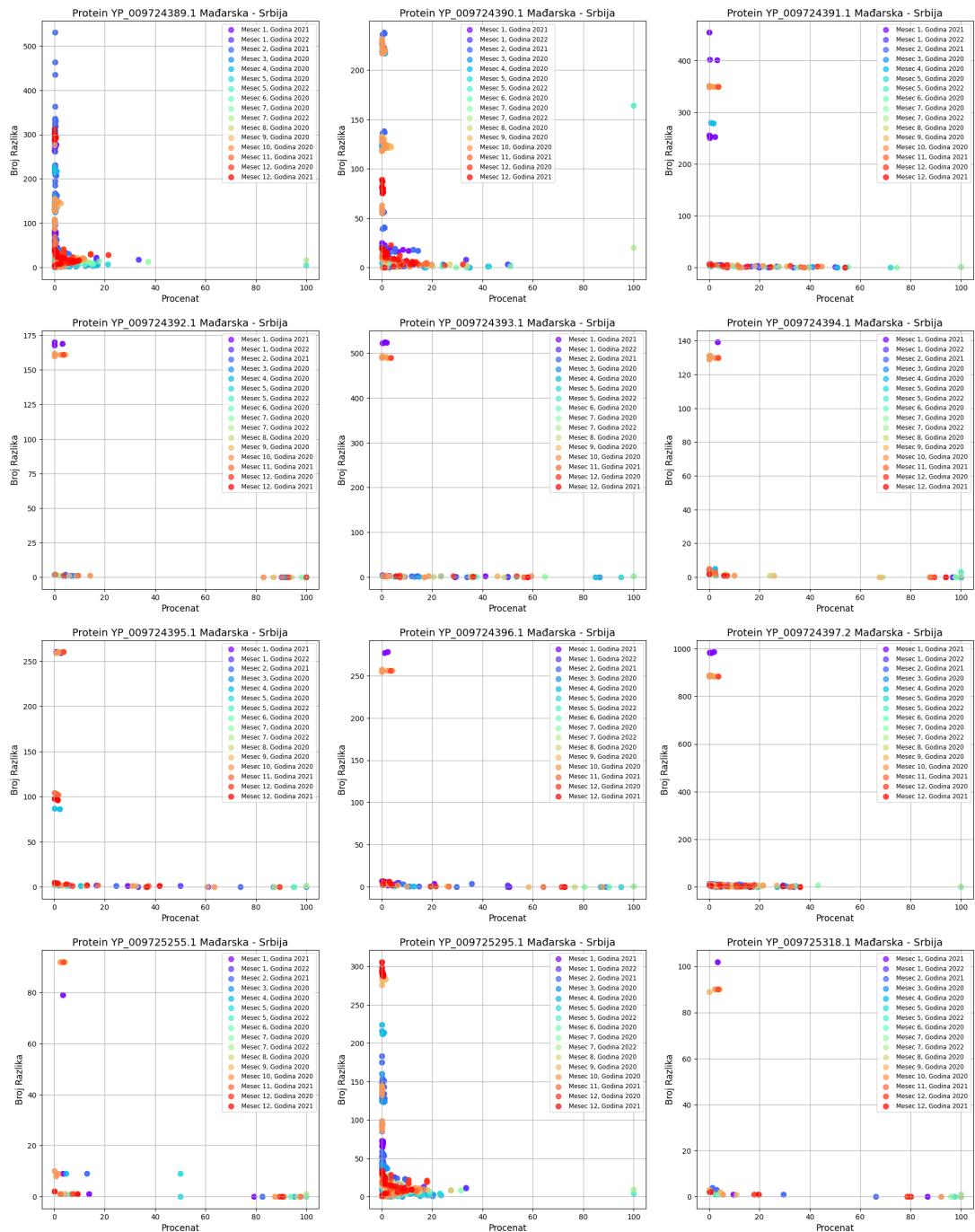
Слика 4.7: Визуелизација процента различитих секвенци за сваки протеин за пар држава: Хрватска - Србија



Слика 4.8: Визуелизација процента различитих секвенци за сваки протеин за пар држава: Јапан - Мађарска



Слика 4.9: Визуелизација процента т различитих секвенци за сваки протеин за пар држава: Јапан - Србија



Слика 4.10: Визуелизација процента различитих секвенци за сваки протеин за пар држава: Мађарска - Србија

# 5 Проценат уникатних секвенци на територији Србије

У овом поглављу бавићемо се рачунањем процента уникатних секвенци (за сваки протеин) које су присутне у укупним узорцима из Србије у односу на укупан број узорака за сваку од остале 4 државе. То значи да смо израчунали проценат секвенци које се јављају у узорцима из Србије, али не и у некој од преосталих држава. Користили смо скрипте `unikatne_table.sql` и `unikatne.sql`. Креиране су комбинације друге државе и протеина. Процедура `CalculateUniquePercentage` коришћена је за израчунавање процента, а резултати су уписаны у табелу `unikatni_procenti_srbija`.

Анализирајући графиконе ([слика 5.1.](#) и [слика 5.2.](#)) можемо донети следеће закључке:

## 1. Доминантни протеини

- За све државе (Грчка, Хрватска, Јапан, Мађарска) важи да је проценат уникатних секвенци протеина YP\_009724389.1 и YP\_009725295.1 висок.
- Протеин YP\_009724389.1 има највећи проценат уникатних секвенци у свим земљама, крећући се од 97.6% (Мађарска) до 100% (Јапан).
- Уникатне секвенце протеина YP\_009725295.1 су такође високо заступљене, са процентима од 94.2% (Мађарска) до 97.4% (Јапан).

## 2. Варијабилност међу земљама

- Протеини YP\_009724392.1, YP\_009724394.1, YP\_009725255.1 и YP\_009725318.1 су највише заступљени у свим земљама, са процентима уникатних секвенци мањим од 5%.
- YP\_009724390.1: Присутност уникатних секвенци се креће између 58.9% (Хрватска) и 74.3% (Јапан), што указује на варијабилност.

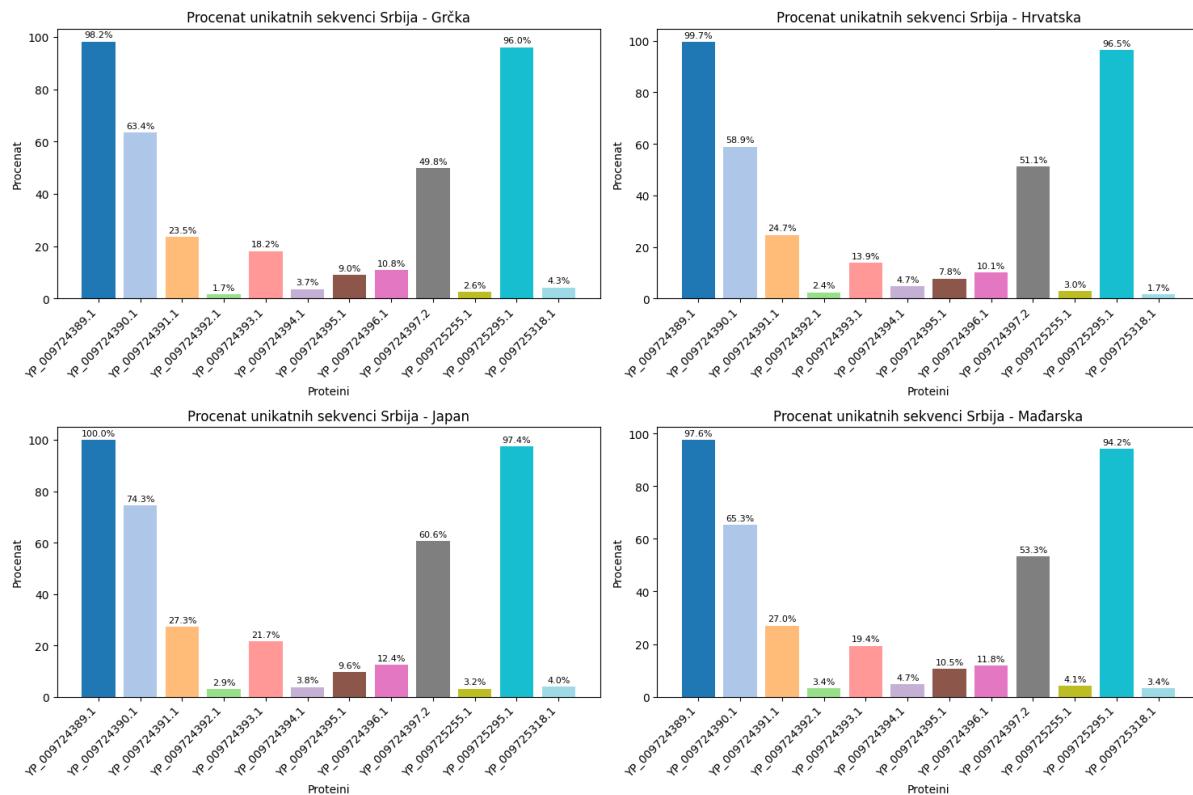
## 3. Сличности између земаља

- Јапан се издваја по потпуном присуству протеина YP\_009724389.1 и високом присуству протеина YP\_009725295.1.
- Драстично одступање вредности истог протеина у различитим државама није забележено.

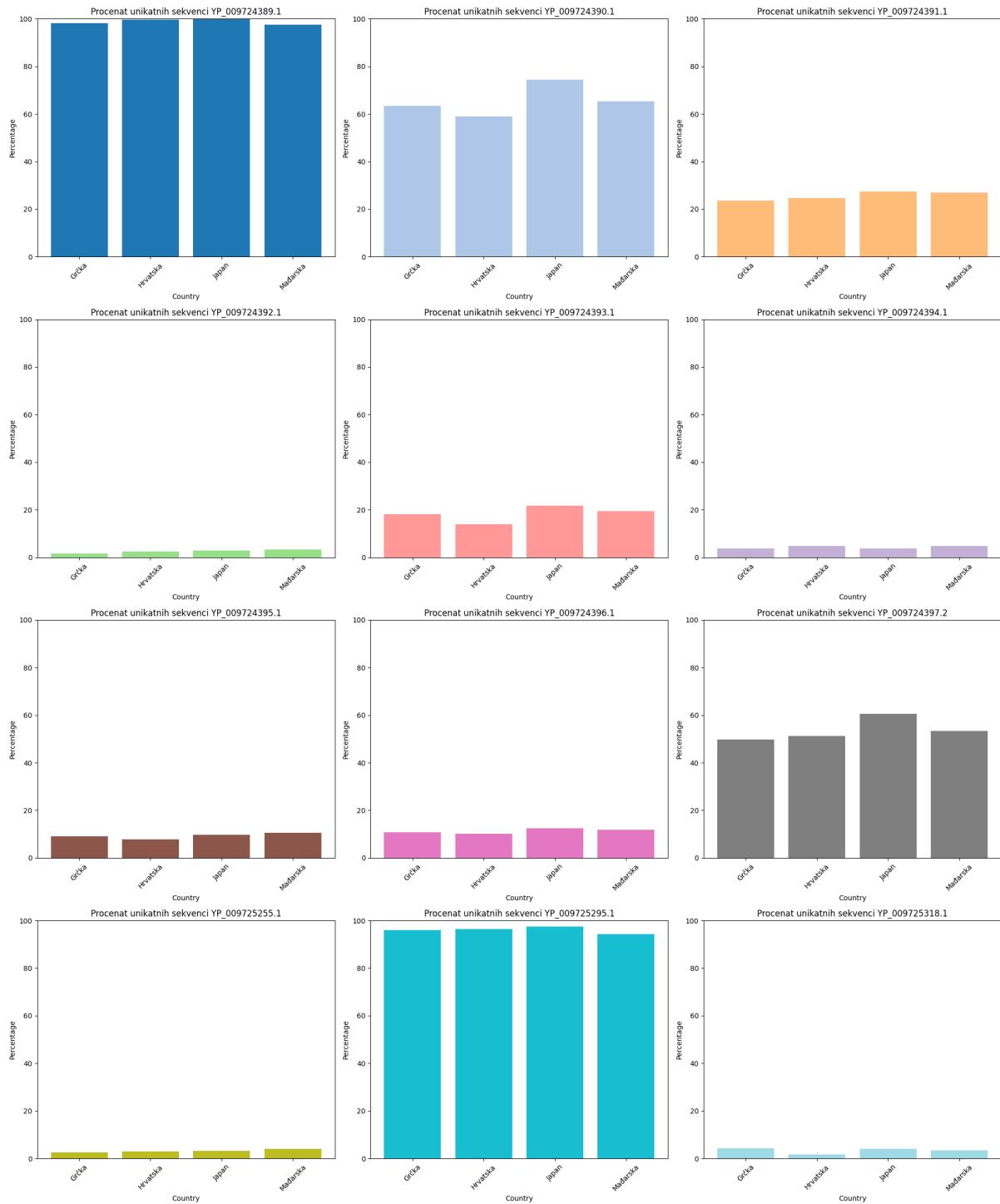
## 4. Закључак о различитостима

- Постоји разлика у процентима одређених протеина између земаља. На пример, протеин YP\_009724390.1 показује распон присуства, што може указивати на географске варијације у секвенцима или различите факторе селекције.

- Доминантни протеини (YP\_009724389.1 и YP\_009725251.1) имају конзистентну присутност у свим земљама, док остали протеини показују мање варијације у процентуалном присуству.



Слика 5.1: Визуелизација процента уникатних секвенци издвојених по државама



Слика 5.2: Визуелизација процента уникатних секвенци издвојених по протеинима

## 6 Најдужи низ нуклеотида идентичан за узорке из Србије и Јапана

У овом поглављу бавимо се одређивањем најдужег низа нуклеотида који је идентичан за узорке из Србије и Јапана као и којим протеинима припадају дати низови за узорке из периода 1.3.2020-1.4.2021. године и периода 1.12.2022-1.1.2023. године.

Креирана су два скупа секвенци, за секвенце из Србије и Јапана за два периода: од 1. марта 2020. до 1. априла 2021. и од 1. децембра 2022. до 1. јануара 2023. године. У овим скуповима тражи се најдужи заједнички подниз секвенци који почиње на истој позицији. Издавање подниза врши се помоћу скрипте `lcs.py`. Секвенце се обрађују редом, бројећи иста слова и чувајући најдужу заједничку ниску. Уколико се било која од ниски разликује на тој позицији или садржи знак '`_`', започиње се нова обрада секвенци од прве следеће позиције. Издвојени поднизови се затим уносе у табелу `lcs_srbija_japan`. Можемо да закључимо да се оба подниза налазе унутар граница за протеин ORF1AB (YP\_009724389.1).

start_date	end_date	leva_granica	desna_granica	lcs
2020-03-01	2021-04-01	19999	20176	TTGATGGTCAGTAGACTTATTAGAAATGCCGTAATGGTGTCTTATTACAGAAGGTAGTGTAAAGGTTAACACCATGTTAACCCATCTGAGGCCCCAACAAAGCTAGTCTTAATGGAGTCACATTAAATTGGAGAACGGCGAAAAACACAGTTAAATTATAAGAAAGTTGATGGTTG TTGAGTAAGAATTACCTGCACCGTCTCTGAGAGCTAACAGCAGATATTGCTTAAGGCAACTTGGCTGCTGAAATTGTTGACACTGTTGAGTG TTACCTGACCCACGACATTGCTAACCTAAGGGCACACTAGAACCGAAATTTCACATTGCTGTAGACTTATGAAATTAGTTGCTAACCTGCTCAA TTGGGTTATGATAATAAGCTTAAAGCACATAAAAGACAATTAGCTCAATGCTTAAATTGTTTATAAGGGTGTACCGTGGTAAGAGAATTCCCTAACAGTAAC TTATCACGCATGATGTTCATCTGCAATTAAACAGGCCAACAAATAGGCGTGGTAAGAGAATTCCCTAACAGTAAC CCTGCTGGAAAGAACGTTCTTACCTCTTAAATTCAAGAGATGCTGTAGGCTCTAAAGGATTTTGGGACTA CCAACCTAACACTGTTGATTCAACAGGGCTCAGAAATTGACTATGTCATATTCACTAACAAACCACTGAACACAGCT CACTCTGTAATGTAACAGATTTAATGCTTACCTACAGAGAACAGTAACTTGGCTATTACAGGAGAACAGTAGGGCATAATTGCTGAT AGAGACCTTTATGACAAGTTGCAATTACAAGCTTGAAGGATGTGCAACTTTACAAGCTGA AAATGTAACAGGACTCTTAAAGATTGTAAGGTAATCACTGGGTTACATCTCACACAGGCCACCTACACC TCAGGCTGACACTAATTCAAAAGCTGAAGGTTTATGTTGAGCTGACTCTGGCATACCTAACAGGACATGACCT AGAAAGACTCATCTATGATGGGTTTAAATTCAAGTTAACGTTAACGTTACCTAACATGTTTATCACCGC GAAGAAGCTATAAGACATGTCGATGGATTGGCTCAGTGTGAGGGGTGTCATGCTACTAGAGAACAGTG TTGGTACCAATTACCTTACAGCTAGGTTAACCTAGTTGCTGTTACCTACAGGTTATGTT ATACACCTTAATAACAGATTTCAGAGTTAGTGTCAAAACCCCGCTGGAGATCAATTAAACCCCTACAC CACTATGTCACAAAGGACTCTTGGAAATGAGTGTCTATAAAAGGTTAACAGTCAAAAGTTAAAGTGAACACTAA ATCTCTGACAGAGCTTGTCTTGGGCACTGGCTTGTAGGTTGACATCTGAGTATTGAAAGTTTTGAAAA TAGGACCTGAGCGCACCTGTTGTCTATGTTGATAGACGTCGACATGCTTTCACACTGCTCAGACACTTATGCC TGTGGGATCATCTTGGGATTGTTGATGGCTCTATAACCTGTTTATGTTGATGTTCAACAAATGGGGTTTACA GGTAACCTACAAAGCAACCATGATCTGTTATGTCAGTCCATGGTAATGTCACATGTTGAGTTGATGCAAT CATGACTAGGGTGTAGCTGTCACAGGTCTTGGTTAACGGTGTGACTGGACTATTGAAATATCTATAATTG GTGATGAACTGAAGATAATGGCTTGTAGAAAAGGTTCAACACATGGTTAAAGCTGCTTATAGCGAC AAATCCCAGTTCTCACGACATTGGTAACCTAAAGCTTAAAGTGTGACCTCAAGCTGATGTTAGAATGGAA GTTCTATGATGTCACAGCTTGTAGTGTGACAAAGCTTAAAGGTTAACACATGGTTAAAGCTGCTTATAGCGAC TGACAAATTACAGATGGTGTAGCTCTACCTTAACTGGCTGGTTGTGATGGTGGCAGTTGTTGATGAAATAAACA TAGATTGACACTAGGTGCTATCTACCTTAACTGGCTGGTTGTGATGGTGGCAGTTGTTGATGAAATAAACA TGCATTCCACACACCGCTTGTAAAGTCTTTGTTAATTAAACAAATTACCATTTCTATTACTCTGAC AGTCATGTGAGTCTCATGGAAAACAAGTAGTGTGAGTATAGATTGACCATAAACTGCTACGTTGAT AACACGTTGCAATTAGTGGTGCTGTAGACATCATGCTAATGAGTACAGATTGTTAT
2022-12-01	2023-01-01	17223	19522	TTGGTACCAATTACCTTACAGCTAGGTTAACCTAGTTGCTGTTACCTACAGGTTATGTT ATACACCTTAATAACAGATTTCAGAGTTAGTGTCAAAACCCCGCTGGAGATCAATTAAACCCCTACAC CACTATGTCACAAAGGACTCTTGGAAATGAGTGTCTATAAAAGGTTAACAGTCAAAAGTTAAAGTGAACACTAA ATCTCTGACAGAGCTTGTCTTGGGCACTGGCTTGTAGGTTGACATCTGAGTATTGAAAGTTTTGAAAA TAGGACCTGAGCGCACCTGTTGTCTATGTTGATAGACGTCGACATGCTTTCACACTGCTCAGACACTTATGCC TGTGGGATCATCTTGGGATTGTTGATGGCTCTATAACCTGTTTATGTTGATGTTCAACAAATGGGGTTTACA GGTAACCTACAAAGCAACCATGATCTGTTATGTCAGTCCATGGTAATGTCACATGTTGAGTTGATGCAAT CATGACTAGGGTGTAGCTGTCACAGGTCTTGGTTAACGGTGTGACTGGACTATTGAAATATCTATAATTG GTGATGAACTGAAGATAATGGCTTGTAGAAAAGGTTCAACACATGGTTAAAGCTGCTTATAGCGAC AAATCCCAGTTCTCACGACATTGGTAACCTAAAGCTTAAAGTGTGACCTCAAGCTGATGTTAGAATGGAA GTTCTATGATGTCACAGCTTGTAGTGTGACAAAGCTTAAAGGTTAACACATGGTTAAAGCTGCTTATAGCGAC TGACAAATTACAGATGGTGTAGCTCTACCTTAACTGGCTGGTTGTGATGGTGGCAGTTGTTGATGAAATAAACA TAGATTGACACTAGGTGCTATCTACCTTAACTGGCTGGTTGTGATGGTGGCAGTTGTTGATGAAATAAACA TGCATTCCACACACCGCTTGTAAAGTCTTTGTTAATTAAACAAATTACCATTTCTATTACTCTGAC AGTCATGTGAGTCTCATGGAAAACAAGTAGTGTGAGTATAGATTGACCATAAACTGCTACGTTGAT AACACGTTGCAATTAGTGGTGCTGTAGACATCATGCTAATGAGTACAGATTGTTAT

Слика 6.1: Визуелизација процента уникатних секвенци издвојених по протеинима

## 7 Закључак

Ово истраживање је пружило детаљну анализу сличности и разлика у секвенцима SARS-CoV-2 вируса прикупљених из пет различитих земаља (Мађарска, Хрватска, Србија, Јапан и Грчка) у периоду од 2020. до 2022. године. Употребом различитих софтверских алата и база података, анализирани су генетски подаци како би се идентификовале кључне сличности и разлике у секвенцима вируса.

Претпроцесирање података је омогућило уклањање дупликата, узорковање секвенци, и њихову припрему за даљу анализу. Поравнање нуклеотидних секвенци у односу на референтни изолат је извршено помоћу алата Clustal Omega, што је омогућило прецизно поређење секвенци и идентификацију кодирајућих региона.

Анализа процента идентичних секвенци показала је да је у неколико случајева проценат идентичности између земаља био изнад нуле, указујући на потенцијалне сличности у преносу вируса у одређеним временским периодима.

Проценат различитих секвенци је варирао међу земљама и месецима, што је омогућило боље разумевање еволуције вируса у различитим географским подручјима.

Проценат уникатних секвенци у Србији указао је на доминантне протеине који су присутни само у Србији у поређењу са осталим земљама, што може бити резултат специфичних фактора или услова у том подручју.

На крају, анализа најдужег низа нуклеотида идентичних за узорке из Србије и Јапана указала је на присуство заједничких секвенци за ORF1AB протеин, што је додатно потврдило значај ове регије у еволуцији SARS-CoV-2 вируса.

Ови налази доприносе бољем разумевању генетске разноликости и еволуције SARS-CoV-2 вируса, пружајући важне информације за будућа истраживања и стратегије за контролу и превенцију COVID-19.

# Референце

- [1] GISAID. <https://gisaid.org/>
- [2] Референтна секвенца NC\_045512.2 [https://www.ncbi.nlm.nih.gov/nuccore/NC\\_045512.2/](https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2/)