

РАД СА БАЗАМА ПОДАТАКА У ПРОГРАМСКОМ ЈЕЗИКУ ПАЈТОН СА ПРИМЕНАМА У БИОИНФОРМАТИЦИ

Студент: Милош Арсић
Ментор: проф. др Весна Маринковић

Универзитет у Београду, Математички факултет

10. септембар 2025.



Увод и мотивација

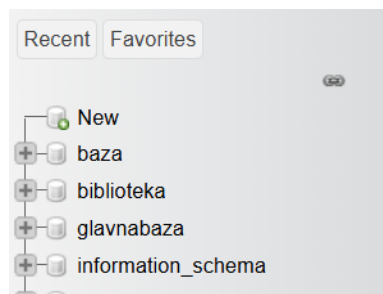
- Са порастом потребе за ефикасним чувањем, претраживањем и руковањем великим количинама података јавила се потреба за развој база података
- Како би се подаци структурирали и елиминисала редувантност и зависност међу њима развијен је релациони модел података
- Циљ рада је илустрација начина на који се превазилазе разлике између релационог модела података и модела података у програмском језику Пајтон
- Специфични циљеви рада обухватају развој апликације у програмском језику Пајтон за рад са базом података у којој су смештени биолошки подаци, израду графичког корисничког интерфејса помоћу библиотеке Tkinter, као и приказ поравнања нуклеотидних секвенци и филогенетске анализе

Систем за управљање базом података

- 1970. године Едгар Франк Код представља релациони модел
- Крајем осамдесетих година настају први комерцијални системи за управљање базом података: Oracle, IBM DB2 и Sybase
- Технике објектно-оријентисане парадигме имају значајну улогу у развоју система за управљање базом података
- Развија се објектно-релациони модел података и системи за управљање засновани на објектно-релационом моделу података међу којима су: Oracle, IBM DB2, Microsoft SQL Server и PostgreSQL
- 1995. године у Шведској настаје MySQL
- 2000. године као библиотека у програмском језику C настаје SQLite
- 2009. године настаје MariaDB

Клијент-сервер архитектура

- Клијент – сервер архитектура представља модел у коме рачунар кога зовемо клијент захтева путем мреже ресурсе од другог рачунара кога зовемо сервер
- Сервер обично има базу података за чување података и покреће програме за обраду захтева
- Програм XAMPP - cross-platform (X), Apache (A), MariaDB (M), PHP (P) i Perl (P)



Databases

Create database

Database name

utf8mb4_general_ci

Create

Database server

- Server: 127.0.0.1 via TCP/IP
- Server type: MariaDB
- Server connection: SSL is not being used
- Server version: 10.4.24-MariaDB - mariadb.org binary distribution
- Protocol version: 10
- User: root@localhost
- Server charset: UTF-8 Unicode (utf8mb4)

Повезивање и комуникација са базом података

- Програм написан у програм језику Пајтон може да приступа бази података коришћењем драјвера **MySQL Connector**
- За повезивање са базом података користи се метода **mysql.connector.connect**

```
bp = mysql.connector.connect(host = hname, user = uname,  
                             password = upass, database = dbname)
```

- Комуникација са базом података се остварује помоћу курсора
- Метода **cursor()** позива се над објектом везе

```
veza = povezivanje_sa_bazom("localhost", "root", "", "biblioteka")  
kursor = veza.cursor()
```


Извршавање и читање резултата упита

- За извршавање SQL упита позивају се методе **execute** и **executemany**

```
kursor = veza.cursor()  
kursor.execute(sql)  
veza.commit()
```

- За читање резултата SQL упита позивају се методе **fetchmany** и **fetchall**

```
kursor = veza.cursor()  
kursor.execute(upit_tabele)  
rezultati = kursor.fetchall()
```

- Потребно је прочитати све редове резултата



Рад са биолошким подацима

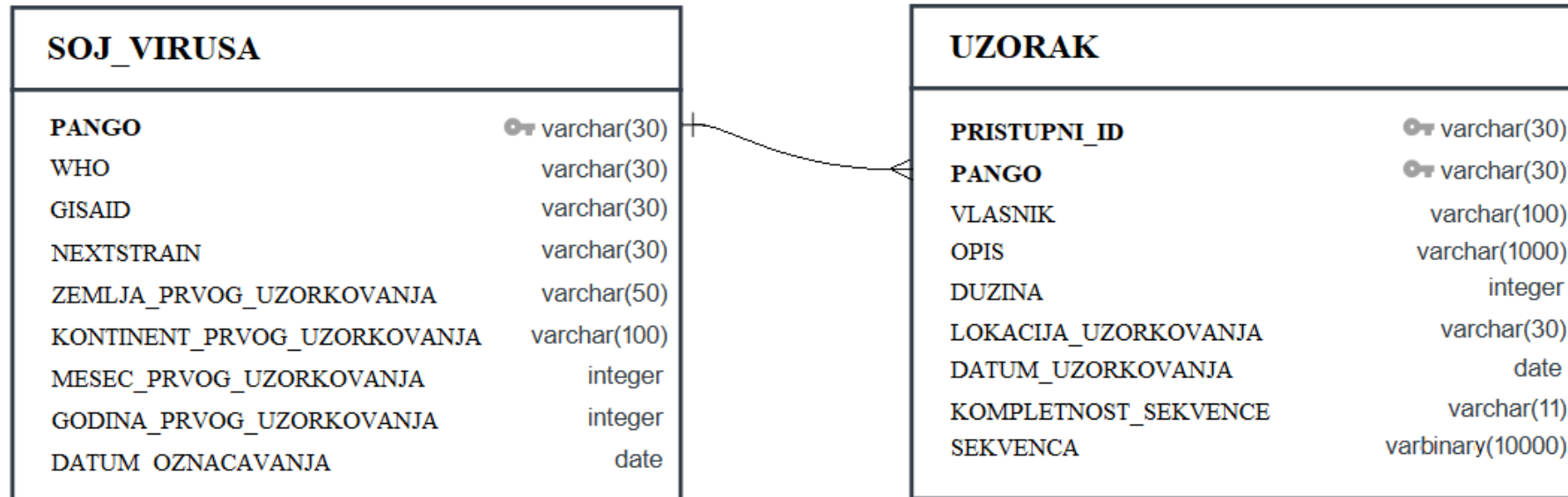
- База података садржи податке о првих једанаест сојева вируса SARS-CoV-2 и о њиховим узорцима
- Подаци о сојевима преузети су за званичне презентације Светске здравствене организације (WHO)
- Сви одабрани узорци потичу из Сједињених Америчких Држава
- Узорци и подаци о узорцима преузети су из базе података Националног центра за биотехнолошке информације (NCBI)

- FASTA FORMAT \longrightarrow STRING $\xrightarrow[\text{и компресија}]{\text{енкодирање}}$ BINARNI FORMAT

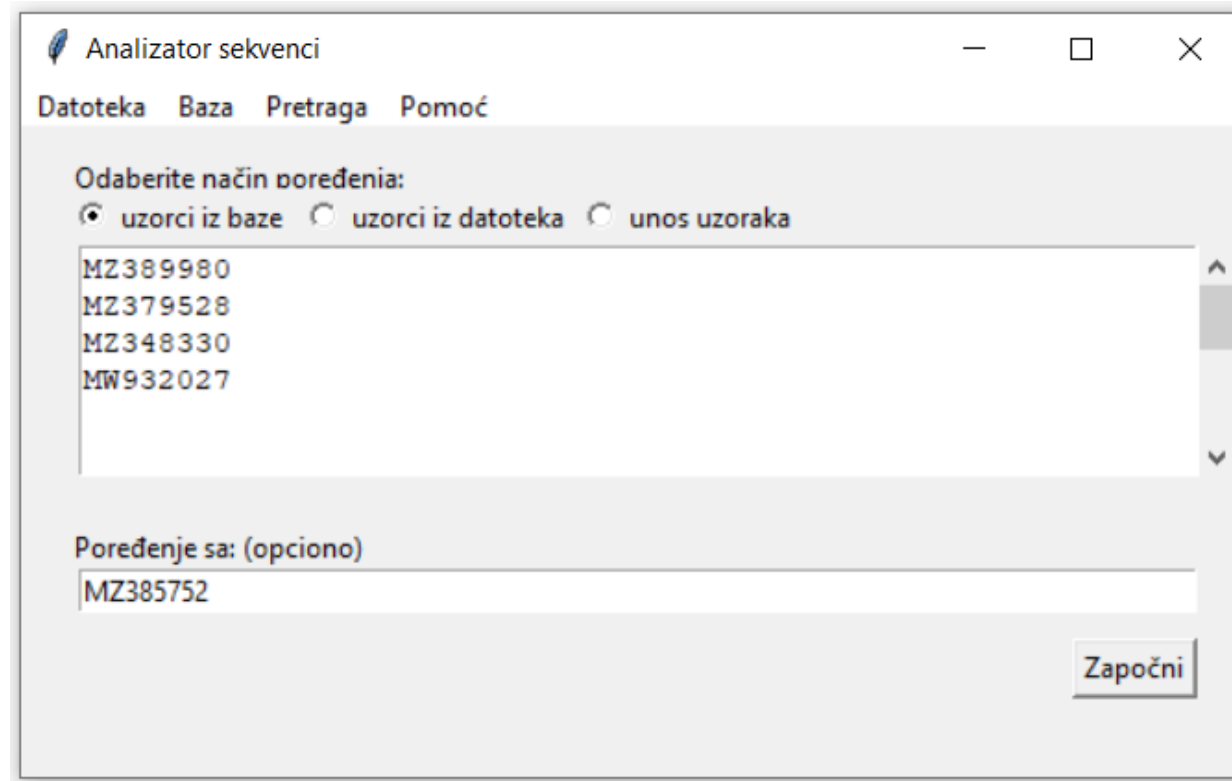
- Просечна уштеда меморије од чак 70,72% за секвенце коришћене у раду.



Структура базе података са биолошким подацима



Почетна страница апликације



Analizator sekvenci

Datoteka Baza Pretraga Pomoć

Odaberite način poređenja:

☒ uzorci iz baze ☐ uzorci iz datoteka ☐ unos uzoraka

MZ389980
MZ379528
MZ348330
MW932027

Poređenje sa: (opciono)

MZ385752

Započni

Поравнање нуклеотидних секвенци

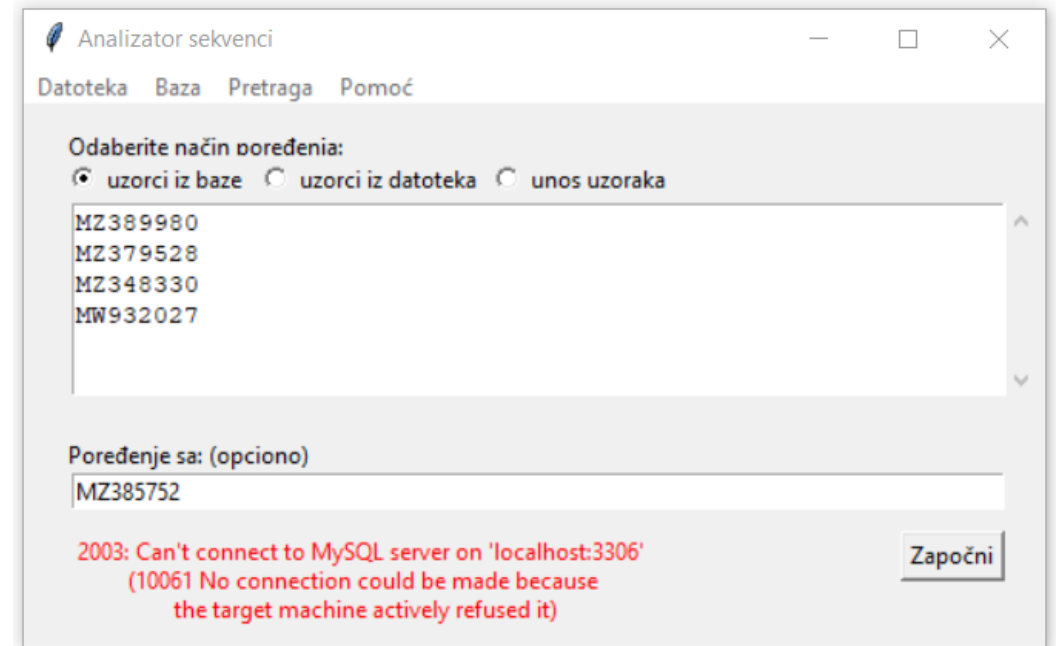
- ДНК се мења услед мутација, које најчешће подразумевају:
 - замене (супституције)
 - брисање (делеције)
 - убацивање (инсерције)
- Проширена нуклеотидна секвенца настаје када се у оригиналну нуклеотидну секвенцу унесу један или више симбола празнине
- Разликујемо глобално и локално поравнање
- Према броју нуклеотидних секвенци које учествују у поравнању, алгоритме делимо на алгоритме за поравнање у паровима и вишеструко поравнање нуклеотидних секвенци
- За вишеструко поравнање нуклеотидних секвенци коришћен је програм Clustal Omega

Преглед функционалности апликације

- Поравнања нуклеотидних секвенци
- Конструкција филогенетског стабла
- Унос и ажурирање података о сојевима вируса и узорцима
- Брисање података
- Претрага:
 - претрага сојева вируса
 - претрага узорака
 - напредна претрага (унос SQL упита)

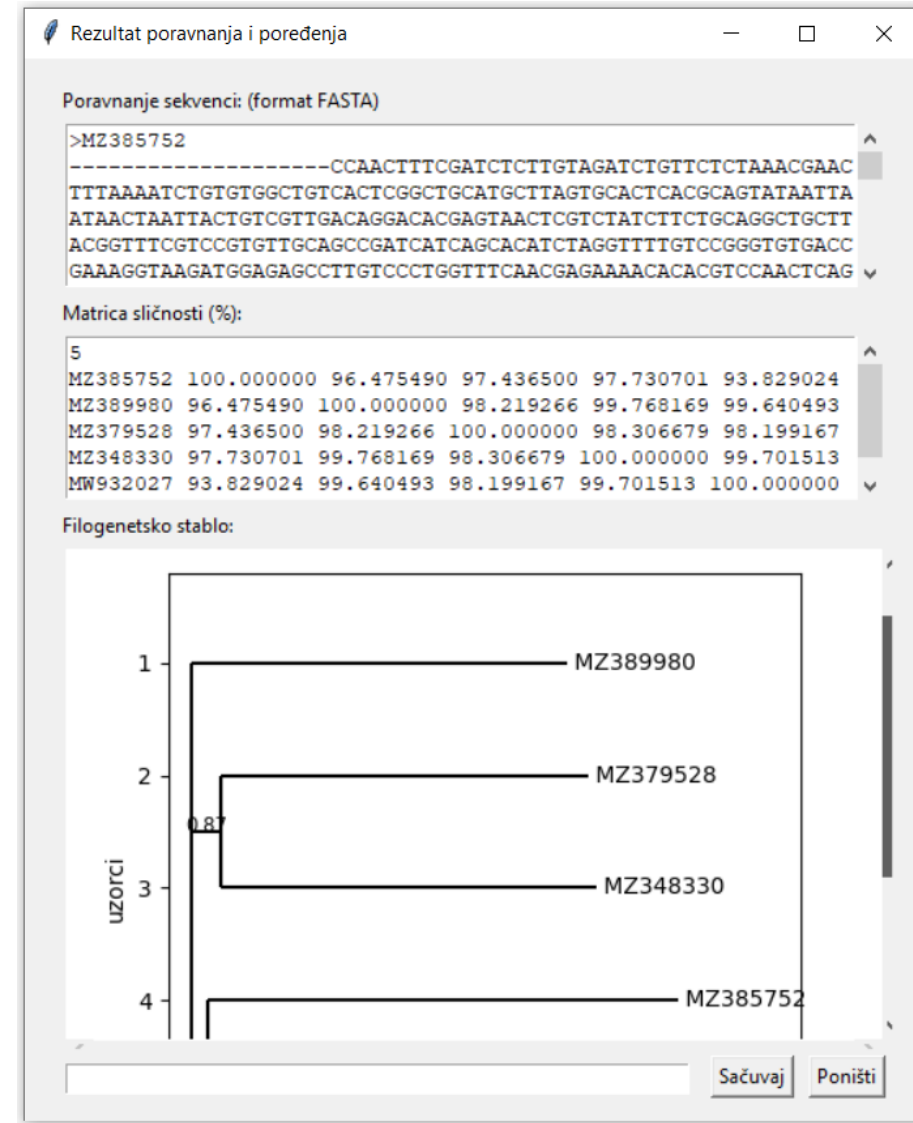
Обрада грешака

- Значајне грешке обрађене током имплементације апликације:
 - грешке приликом успостављања везе са базом података
 - грешке због неправилног уноса података
 - грешке због неправилно наведених путања до улазних и излазних датотека
 - различите SQL грешке
- Најзначајнији обрађени изузеци:
FileNotFoundException, CalledProcessError, mysql.connector.Error, ValueError и Exception



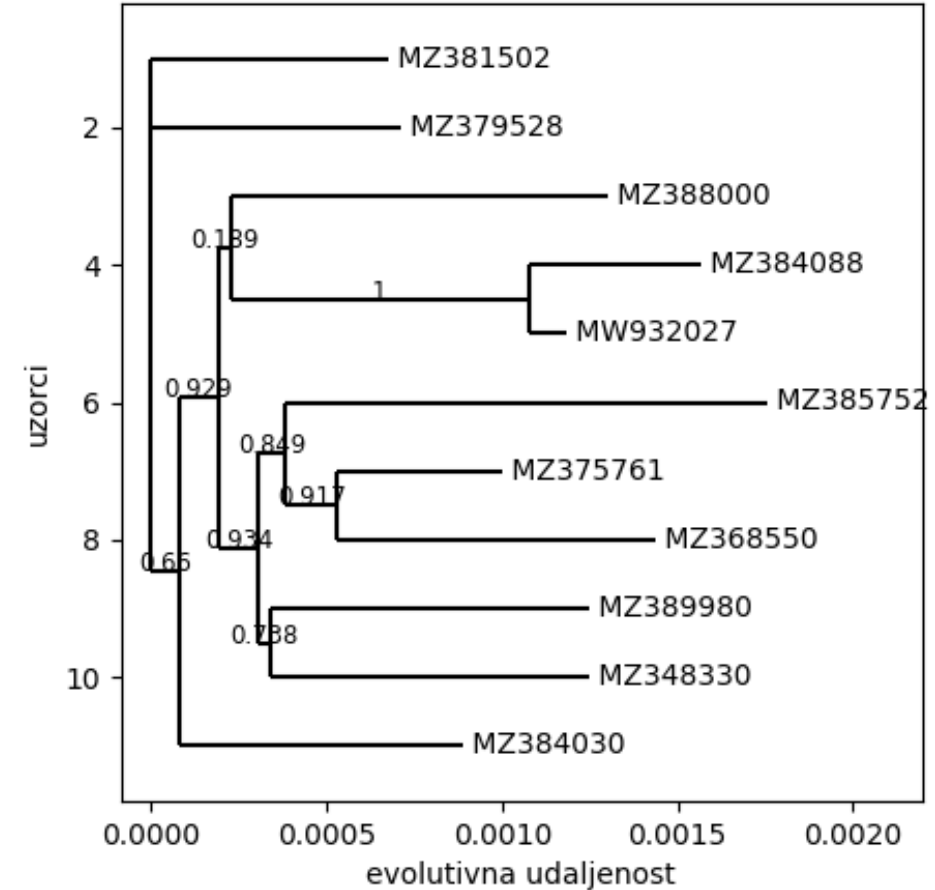
Приказ и анализа добијених резултата

- У засебном прозору апликације приказују се добијени резултати:
 - поравнање секвенци у FASTA формату
 - матрица сличности (у процентима)
 - филогенетско стабло
- Апликација омогућава чување и учитавање већ добијених резултата
- Омогућен је приказ процената сличности истакнутог узорка са преосталима



Приказ и анализа добијених резултата

- На основу добијеног поравнања, коришћењем програма FastTree, генерисано је филогенетско стабло приказано на слици
- Добијено стабло садржи више унутрашњих чворова са различитим нивоима подршке



Закључак

- У раду је развијена апликација за рад са базом података у којој се чувају биолошки подаци
- Омогућено је поравнање нуклеотидних секвенци коришћењем програм Clustal Omega
- На основу добијеног поравнања, коришћењем програма FastTree, конструисано је филогенетско стабло
- У скупу одабраних узорака уочен је истакнути кластер коме припадају нуклеотидне секвенце узорака сојева Alpha, Gamma, Zeta, Theta и Lambda
- Припадност овом кластеру упућује на постојање сличних мутација, њихово заједничко порекло или сличан правац еволутивног развоја ових сојева вируса

Правци за даљи рад и развој апликације

- Могућа унапређења
 - Развој базе података
 - Развој бољег визуелног приказа података у табелама
 - Извоз и генерисање извештаја у формате CSV и PDF
 - Додавање филтера и могућности за сортирање резултата претраге
 - Чување историје претходних SQL упита
 - Увођење провере идентитета корисника и пријаве
 - Развој подршке за конкурентни рад више корисника над базом података

Хвала на пажњи!