

УНИВЕРЗИТЕТ У БЕОГРАДУ  
МАТЕМАТИЧКИ ФАКУЛТЕТ



Емилија Стевановић

# **РАЗВОЈ АПЛИКАЦИЈЕ ЗА АНАЛИЗУ РАСТОЈАЊА АМИНОКИСЕЛИНА У ПРОТЕИНИМА**

мастер рад

Београд, 2026.

**Ментор:**

др Мирјана МАЉКОВИЋ РУЖИЧИЋ  
Универзитет у Београду, Математички факултет

**Чланови комисије:**

др Ненад МИТИЋ  
Универзитет у Београду, Математички факултет

др Саша МАЛКОВ  
Универзитет у Београду, Математички факултет

**Датум одбране:** \_\_\_\_\_



**Наслов мастер рада:** Развој апликације за анализу растојања аминокиселина у протеинима

**Резиме:** Предмет овог рада је анализа просторних односа између аминокиселина у протеинима на основу растојања међу резидуумима. У структурној биоинформатици растојања се најчешће сажимају у матрицу контаката при меном фиксног прага: уместо измерене вредности задржава се само податак о томе да ли контакт постоји. Секундарна структура и класификација резидуума при томе није приказана. У раду је развијена веб апликација која за чување података о растојањима користи графовски систем за управљање базом података *Neo4j* и омогућава њихово истраживање без унапред одабраног прага. Коришћен је скуп података *DistAA*, који обухвата 940 протеинских ланаца из 933 протеина. За сваки пар резидуума тај скуп садржи растојање између  $C_\alpha$  атома, као и минимално и максимално растојање узимајући у обзир тешке атоме, уз придружене *DSSP* кодове секундарне структуре и *SCOP* класе. Аминокиселине су моделоване као чворови типа *AminoAcid*, а растојања као везе типа *DISTANCE*, при чему и чворови и везе носе придружена својства. Тако формирана база података садржи 150 685 чворова и 46 306 160 веза. Апликација је организована у три слоја – кориснички интерфејс заснован на библиотеци *3Dmol.js*, апликативни слој реализован помоћу веб оквира *Django* и слој података заснован на графовском систему за управљање базом података *Neo4j*. За постављање и покретање апликације коришћена је *Docker Compose* конфигурација. У апликацији корисник може да бира протеин и ланац, да ограничи приказ према типу растојања, опсегу вредности, секундарној структури и минималном растојању резидуума у секвенци. Тродимензионални приказ графа и одређени статистички показатељи се освежавају при промени параметара од којих зависе.

**Кључне речи:** протеинске структуре, аминокиселине, растојања између резидуума, графовске базе података, *Neo4j*, *Cypher*, визуализација, *3Dmol.js*, *Django*

# Садржај

|          |                                                                               |           |
|----------|-------------------------------------------------------------------------------|-----------|
| <b>1</b> | <b>Увод</b>                                                                   | <b>1</b>  |
| <b>2</b> | <b>Биолошки појмови</b>                                                       | <b>4</b>  |
| 2.1      | Протеини и аминокиселине . . . . .                                            | 4         |
| 2.2      | Нивои структуре протеина . . . . .                                            | 5         |
| 2.3      | Класификација протеинских структура . . . . .                                 | 8         |
| 2.4      | Координате $C\alpha$ и типови растојања . . . . .                             | 10        |
| <b>3</b> | <b>Технологије</b>                                                            | <b>12</b> |
| 3.1      | NoSQL базе података и графовски модел . . . . .                               | 12        |
| 3.2      | Предности и примене графовских база података . . . . .                        | 13        |
| 3.3      | ГСУБП Neo4j . . . . .                                                         | 14        |
|          | Чворови и везе . . . . .                                                      | 15        |
|          | Cypher упити . . . . .                                                        | 15        |
|          | Индекси и план извршавања . . . . .                                           | 18        |
| 3.4      | Остале коришћене технологије . . . . .                                        | 19        |
|          | Веб оквир Django . . . . .                                                    | 19        |
|          | Технологије клијентског дела . . . . .                                        | 20        |
|          | Docker . . . . .                                                              | 20        |
| <b>4</b> | <b>Опис апликације</b>                                                        | <b>22</b> |
| 4.1      | Архитектура решења . . . . .                                                  | 22        |
| 4.2      | Граф модел и увоз података . . . . .                                          | 24        |
|          | Упити над базом података . . . . .                                            | 26        |
| 4.3      | Серверски део апликације: Django и крајње тачке API-ја . . . . .              | 28        |
| 4.4      | Клијентски део апликације: 3Dmol.js, филтери и кориснички интерфејс . . . . . | 30        |
|          | Панели за анализу . . . . .                                                   | 33        |

|          |                                             |           |
|----------|---------------------------------------------|-----------|
| <b>5</b> | <b>Резултати и анализа</b>                  | <b>38</b> |
| 5.1      | Глобалне статистике базе података . . . . . | 38        |
| 5.2      | Методологија избора протеина . . . . .      | 40        |
| 5.3      | Студије случаја . . . . .                   | 42        |
|          | Протеин 1K3I . . . . .                      | 42        |
|          | Протеин 1A17 . . . . .                      | 48        |
|          | Протеин 1G2Y . . . . .                      | 51        |
|          | Протеин 1FEW . . . . .                      | 55        |
| 5.4      | Дискусија . . . . .                         | 58        |
| <b>6</b> | <b>Закључак</b>                             | <b>63</b> |
|          | <b>Литература</b>                           | <b>65</b> |

# Глава 1

## Увод

Функција протеина не произлази непосредно из редоследа аминокиселина, већ на њу утиче и просторни распоред који ланац заузима након увијања. Аминокиселине међусобно удаљене десетинама позиција у секвенци могу се у простору наћи на свега неколико ангстрема и остварити контакт. Због тога се анализа протеинских структура у великој мери своди на растојања: које аминокиселине су довољно блиске да међусобно интерагују и какав образац те блискости граде.

Растојања се најчешће сажимају у матрицу контаката, која настаје применом одабраног прага на растојања између  $C_\alpha$  атома или између тешких атома. Nath Јha и сарадници [14] користе праг од 6,5 Å за  $C_\alpha - C_\alpha$  растојања, односно 4,5 Å за контакте засноване на тешким атомима.

Такав приказ је сажет и погодан за нумеричку обраду, али унапред одбацује део информација. Избором једног прага сви односи изван њега не улазе у анализу, а вредност растојања своди се на бинарну одлуку о постојању контакта.

Скуп података *DistAA* [11], коришћен у овом раду, чува растојања за све парове резидуума, без примене прага: сваки ред одговара једном пару аминокиселина, а растојање је дато нумерички, у ангстремима. Обухвата 940 протеинских ланаца из 933 протеина, а за сваки пар резидуума бележи три типа растојања: растојање између  $C_\alpha$  атома (*саса*), као и минимално и максимално растојање између тешких атома два резидуума (*minbezh* и *maxbezh*). Уз растојања су придружени и *DSSP* кодови секундарне структуре и *SCOP* класе оба резидуума, чиме је за сваки пар познат и структурни контекст у коме се налази. Подаци су, међутим, организовани у засебне *CSV* датотеке по протеину и

по типу растојања, што отежава добијање одговора на било које питање које обухвата више протеина или повезује растојање са структурним обележјима резидуума.

Овако организовани подаци су по природи графовски: аминокиселине су ентитети, а растојања односи међу њима. У релационом моделу одговор на питање о суседству резидуума захтева спајање табела, док се у графовским базама података суседни резидууми проналазе непосредним праћењем веза, без потребе за спајањем. Уз то, чување ознака и својстава на чворовима и везама омогућава да тип растојања, његова вредност и структурни контекст резидуума стоје на самом елементу графа.

Циљ овог рада је развој веб апликације која растојања између аминокиселина чува у графовској бази података *Neo4j* и омогућава њихово истраживање обједињавањем тродимензионалног приказа и статистичке анализе. Аминокиселине су моделоване као чворови типа *AminoAcid* са својствима која обухватају идентификатор протеина, ланац, позицију у секвенци, ознаку секундарне структуре, *SCOP* класу и координате  $C_{\alpha}$  атома, док су растојања представљена везама типа *DISTANCE* које носе тип и вредност растојања. Овако изграђена база података садржи 150 685 чворова аминокиселина и 46 306 160 веза.

Апликација је организована у три слоја: кориснички интерфејс заснован на библиотеци *3Dmol.js*, апликативни слој реализован помоћу веб оквира *Django* и слој података заснован на графовском систему за управљање базом података (ГСУБП) *Neo4j*. Целокупно решење реализовано је коришћењем технологије *Docker*, при чему су *Django* апликација и ГСУБП *Neo4j* покренути у засебним контејнерима, као сервиси дефинисани *Docker Compose* конфигурацијом. Корисник може да изабере протеин и ланац и да ограничи приказ филтерима – избором аминокиселина, типом растојања, опсегом вредности, типом секундарне структуре и минималним растојањем резидуума у секвенци. Уместо унапред одабраног прага контакта, праг постаје параметар који се мења током рада, а приказ графа у простору и нумерички показатељи се освежавају при свакој промени параметара. Приступ је приказан на глобалној анализи базе података и на четири студије случаја (протеини 1K3I, 1A17, 1G2Y и 1FEW), одабране тако да се разликују по заступљености секундарних структура и просторној распрострањености ланца.

Рад је организован на следећи начин. У другом поглављу су описане биолошке основе: протеини и аминокиселине као градивне јединице, нивои струк-

турне организације, *DSSP* систем и *SCOP* класификација, као и  $C_\alpha$  координате и типови растојања који чине коришћени скуп података. У трећем поглављу описане су нерелационе базе података и графовски модел, а затим ГСУБП *Neo4j*, језик *Cypher* и начин извршавања упита, као и остале коришћене технологије: веб оквир *Django*, технологије клијентског дела и платформа *Docker*. У четвртом поглављу је приказан развој апликације: архитектура решења, граф модел и поступак увоза података, серверски део апликације реализован помоћу оквира *Django* и кориснички интерфејс. У петом поглављу су описане глобалне статистике базе података и студије случаја појединачних протеина, док шесто поглавље резимира резултате и наводи правце могућег даљег развоја.

## Глава 2

# Биолошки појмови

За разумевање података и њиховог значаја потребно је упознати се са основним појмовима структурне биологије. У овом поглављу описани су протеини и аминокиселине као основне градивне јединице, нивои структурне организације протеина, као и алати који се користе за класификацију структура: *DSSP* систем за описивање секундарне структуре и *SCOP* класификација. На крају поглавља представљени су  $C_\alpha$  координате и типови растојања.

### 2.1 Протеини и аминокиселине

Протеини су макромолекули који представљају кључне покретаче процеса у ћелији и имају бројне улоге које омогућавају ћелији да живи, функционише и прилагођава се променама у окружењу. Неке од улога су: обезбеђивање структурне стабилности, каталитичка улога, транспорт супстанци, комуникација између ћелија и околине, пренос генетских информација, омогућавање кретања ћелије и многе друге [1, 3].

Структурно посматрано, протеини су биолошки полимери изграђени од аминокиселина. Аминокиселине су мали молекули који садрже централни атом угљеника, означен са  $C_\alpha$  (алфа-угљеник), за који су везани: амино-група ( $-\text{NH}_3^+$ ), карбоксилна група ( $-\text{COO}^-$ ), атом водоника (H) и бочни ланац (R), по ком се међусобно разликују [16]. Аминокиселине су у протеину повезане пептидним везама које настају између атома угљеника карбоксилне групе једне аминокиселине и атома азота амино-групе друге аминокиселине, при чему се ослобађа молекул воде (реакција кондензације) [16]. Полипептидна окосница (кичма полипептидног ланца) представља понављајући низ

—NH—CHR—CO—, док бочни ланци (R) свакој аминокиселини дају јединствена физичко-хемијска својства [16].

Постоји двадесет стандардних аминокиселина, које се на основу хемијских особина бочних ланаца (R група) сврставају у три основне категорије: аминокиселине са неполарним (хидрофобним) бочним ланцима, аминокиселине са поларним ненаелектрисаним бочним ланцима и аминокиселине са наелектрисаним бочним ланцима, које могу бити позитивно или негативно наелектрисане [1]. Оваква подела одражава начин на који поједине аминокиселине ступају у интеракције са водом и међусобно, што непосредно утиче на структуру и функцију протеина. Преглед свих двадесет аминокиселина, заједно са њиховим трословним и једнословним ознакама и особинама бочног ланца, дат је у табели 1.

## 2.2 Нивои структуре протеина

Број, врста и редослед аминокиселина одређују структуру и функцију протеина. Секвенца аминокиселина у полипептидном ланцу назива се **примарна структура**. Крајеви полипептидног ланца разликују се по структури: крај који садржи слободну амино-групу назива се N-терминус (амино-терминус), а крај са слободном карбоксилном групом C-терминус. Секвенца се по конвенцији приказује од N-терминуса ка C-терминусу, читајући с лева на десно [1].

**Терцијарна структура** представља тродимензионалну конформацију коју полипептидни ланац заузима савијањем у простору, са прецизном оријентацијом свих бочних ланаца [16]. Ова конформација одговара функционалном стању протеина у ком он обавља своју биолошку активност. Терцијарна структура често садржи један или више **домена**, самосталних делова полипептидног ланца који се независно савијају у компактне тродимензионалне јединице. Њихов распоред и међусобна флексибилност одређују функцију протеина, интеракције и активна места [16].

**Секундарна структура** протеина подразумева правилне конформације полипептидне окоснице које се јављају унутар савијеног ланца и стабилизују се водоничним везама између делимично негативног атома кисеоника карбонилне групе и делимично позитивног атома азота пептидне NH-групе. Важно је истаћи да водоничне везе које учествују у формирању секундарне структуре не укључују бочне ланце аминокиселина [18]. Две најчешће и најправилније

Табела 1: Подела аминокиселина према особинама бочног ланца [1]

| Аминокиселина                                  | Трослов-<br>на | Једнослов-<br>на | Особина<br>ланца         | бочног        |
|------------------------------------------------|----------------|------------------|--------------------------|---------------|
| <i>Поларне аминокиселине — наелектрисане</i>   |                |                  |                          |               |
| Аспарагинска киселина                          | Asp            | D                | негативно                | наелектрисана |
| Глутаминска киселина                           | Glu            | E                | негативно                | наелектрисана |
| Аргинин                                        | Arg            | R                | позитивно                | наелектрисана |
| Лизин                                          | Lys            | K                | позитивно                | наелектрисана |
| Хистидин                                       | His            | H                | позитивно                | наелектрисана |
| <i>Поларне аминокиселине — ненаелектрисане</i> |                |                  |                          |               |
| Аспарагин                                      | Asn            | N                | поларна, ненаелектрисана |               |
| Глутамин                                       | Gln            | Q                | поларна, ненаелектрисана |               |
| Серин                                          | Ser            | S                | поларна, ненаелектрисана |               |
| Треонин                                        | Thr            | T                | поларна, ненаелектрисана |               |
| Тирозин                                        | Tyr            | Y                | поларна, ненаелектрисана |               |
| <i>Нейполарне аминокиселине</i>                |                |                  |                          |               |
| Аланин                                         | Ala            | A                | неполарна                |               |
| Глицин                                         | Gly            | G                | неполарна                |               |
| Валин                                          | Val            | V                | неполарна                |               |
| Леуцин                                         | Leu            | L                | неполарна                |               |
| Изолеуцин                                      | Ile            | I                | неполарна                |               |
| Пролин                                         | Pro            | P                | неполарна                |               |
| Фенилаланин                                    | Phe            | F                | неполарна                |               |
| Метионин                                       | Met            | M                | неполарна                |               |
| Триптофан                                      | Trp            | W                | неполарна                |               |
| Цистеин                                        | Cys            | C                | неполарна                |               |

секундарне структуре су  $\alpha$ -хеликс и  $\beta$ -плоча, а поред њих постоје и нерегуларни делови ланца променљивог облика, међу којима се разликују завоји и насумични намотај.

**$\alpha$ -хеликс** је десно-завојна спирална структура која се одржава водоничним везама између сваке четврте аминокиселине у ланцу (између CO-групе остатка  $i$  и NH-групе остатка  $i+4$ ) [18]. Један пун завој спирале обухвата 3,6 аминокиселина [1]. Неки влакнасти протеини су у потпуности сачињени од  $\alpha$ -хеликса. Пример тога је  $\alpha$ -кератин, присутан у коси, ноктима и роговима, а управо је у том протеину  $\alpha$ -хеликс и први пут откривен [1].

**$\beta$ -плоча** (бета-површ) настаје када су два или више сегмената полипептидног ланца положени један поред другог и међусобно повезани водоничним везама [1]. Ланци могу бити *паралелни* (простиру се у истом смеру) или *антипаралелни* (сваки сегмент је у супротном смеру у односу на суседни). Посебан облик који чине антипаралелне бета-нити сложене у затворени цилиндар назива се  $\beta$ -буре. Свака појединачна водонична веза у овим структурама је релативно слаба, али скуп свих веза заједно даје веома круту и стабилну структуру [1].

**Кватернарна структура** означава удруживање два или више полипептидних ланаца у једну функционалну целину. Немају сви протеини кватернарну структуру [18].

Ако се узме у обзир 20 стандардних аминокиселина, број полипептидних ланаца дужине  $n$  у теорији је  $20^n$ . За  $\sim 300$  аминокиселина тај број износи  $20^{300}$ , што представља огроман скуп могућих секвенци. Ипак, само мали део тих молекула могао би да достигне стабилне конформације. Кроз еволуцију су природном селекцијом одабране управо оне секвенце које се правилно савијају и обављају корисну каталитичку или структурну улогу [18]. Управо та зависност функције од прецизне грађе чини протеине изузетно осетљивим на промене: замена само једне аминокиселине другом понекад може озбиљно нарушити конформацију читавог молекула и потпуно укинути његову функцију [18].

Различите болести могу настати као последица промена у структури протеина, било услед мутације у примарној секвенци, било услед денатурације (промена облика изазвана екстерним условима). Класичан пример је српаста анемија, у којој је глутаминска киселина (GLU) замењена валином (VAL). Погрешно савијање  $\beta$ -плоче може довести до оштећења ткива [18].

## 2.3 Класификација протеинских структура

Секундарна структура протеина у овом раду описује се помоћу *DSSP* система (енгл. *Dictionary of Secondary Structure of Proteins*), који су развили Kabsch и Sander [9].

*DSSP* класификује секундарну структуру сваког остатка на основу образаца водоничних веза и геометријских особина које се изводе из кристалографских координата. Аминокиселина уграђена у полипептидни ланац назива се аминокиселински остатак (резидуум), јер се приликом формирања пептидне везе издваја молекул воде. Основа алгоритма су два елементарна обрасца водоничних веза између атома окоснице: *n*-окрет (енгл. *n-turn*), који описује водоничну везу између СО-групе остатка *i* и NH-групе остатка *i+n* (где *n* = 3, 4 или 5), и мост (енгл. *bridge*), који описује водоничну везу између остатака који нису блиски у секвенци. Понављајући 4-окрети дефинишу  $\alpha$ -хеликсе, а понављајући мостови дефинишу  $\beta$ -структуру. Сва остала појављивања ових образаца обухватају  $3_{10}$ -хеликсе,  $\pi$ -хеликсе, појединачне окрете и изоловане  $\beta$ -мостове.

У наставку су наведени *DSSP* кодови који се користе у овом раду [9]:

- H:  $\alpha$ -хеликс;
- G:  $3_{10}$ -хеликс;
- I:  $\pi$ -хеликс;
- E: проширени ланац који учествује у  $\beta$ -плочи;
- B: остатак у изолованом  $\beta$ -мосту;
- T: окрет премошћен водоничном везом (енгл. *H-bonded turn*);
- S: савијање (енгл. *bend*) – геометријски критеријум, без водоничне везе;
- -: аминокиселина није ни у једној од наведених конформација.

Важно је нагласити разлику између кодова T и S. T подразумева постојање водоничне везе, док је S чисто геометријски критеријум заснован на углу између пет узастопних  $C_{\alpha}$  атома. Због тога се присуство кода S не може тумачити као доказ постојања водоничне везе, већ само као показатељ локалне промене правца ланца.

**SCOP** класа се додељује ланцу протеина на основу конформација које садржи, па је у подацима придружена сваком резидууму тог ланца. *SCOP* (енгл. *Structural Classification of Proteins*) је хијерархијски систем за класификацију протеина који су развили Мурзин и сарадници, а који протеине разврстава према њиховој структури и еволутивним односима [13].

У овом систему под идентитетом остатака подразумева се проценат аминокиселина које су исте на одговарајућим позицијама након поравнања две протеинске секвенце. Класификација је организована на следећим хијерархијским нивоима:

- **Породица** (енгл. *family*): протеини са јасном секвенцијском сродношћу, односно идентитетом остатака  $\geq 30\%$ , или протеини са нижим идентитетом, али веома сличном функцијом и структуром, као што су глобини са  $\sim 15\%$  идентитета.
- **Суперпородица** (енгл. *superfamily*): породице са ниским секвенцијским идентитетом, али чија структура и, у многим случајевима, функционална обележја указују на вероватно заједничко еволутивно порекло.
- **Преклоп** (енгл. *fold*): суперпородице и породице чији протеини имају исте главне секундарне структуре у истом просторном распореду и са истим тополошким везама.
- **Класа** (енгл. *class*): преклопи се групишу у класе према доминантним секундарним структурама.

У овом раду коришћене су следеће класе [13]:

- **a**: протеини претежно сачињени од  $\alpha$ -хеликса (енгл. *all-alpha proteins*);
- **b**: протеини претежно сачињени од  $\beta$ -плоча (енгл. *all-beta proteins*);
- **c**: протеини који садрже и  $\alpha$ -хеликсе и  $\beta$ -траке које су међусобно измењане (енгл. *alpha/beta*, означава се  $\alpha/\beta$ );
- **d**: протеини који садрже и  $\alpha$ -хеликсе и  $\beta$ -траке, али су у великој мери просторно раздвојени (енгл. *alpha+beta*, означава се  $\alpha+\beta$ );
- **e**: вишедоменски протеини са доменима различитих преклопа (енгл. *multi-domain proteins*).

## 2.4 Координате $C_\alpha$ и типови растојања

Тродимензионалне координате атома у протеинима одређују се експериментално, најчешће методом рендгенске кристалографије (енгл. *X-ray crystallography*). За сваки атом у структури протеина познате су  $x$ ,  $y$  и  $z$  координате у ангстремима ( $1 \text{ \AA} = 10^{-10} \text{ m}$ ). Угљеников атом  $C_\alpha$  је централни атом аминокиселине и повезан је са свим осталим атомима у остатку, те представља природан избор за описивање просторног положаја аминокиселине у ланцу.

Растојање између  $C_\alpha$  атома два резидуума (саса растојање) често се користи као стандардна апроксимација просторне блискости, и представља основу за формирање матрице контаката протеина применом одговарајућег прага растојања (енгл. *cutoff*). У раду Nath Jha и сарадника [14] коришћен је  $C_\alpha - C_\alpha$  праг од  $6,5 \text{ \AA}$  за дефинисање матрице суседства.

Прецизнија анализа контаката заснива се на растојањима између тешких атома (свих атома осим водоника), уз строжи праг контакта. Атоми водоника се искључују јер њихов положај није поуздано одређен кристалографским методама. У поменутом раду коришћен је праг од  $4,5 \text{ \AA}$  за контакте засноване на тешким атомима [14].

У овом раду коришћена је база података **DistAA** [11], која садржи податке о 940 протеинских ланаца из 933 протеина и обухвата три типа растојања између аминокиселина:

- **саса**: растојање између  $C_\alpha$  атома два резидуума;
- **maxbezh**: максимално растојање између тешких атома два резидуума;
- **minbezh**: минимално растојање између тешких атома два резидуума.

Скуп података садржи засебне *CSV* датотеке за сваки тип растојања и за сваки протеин. Подаци о растојањима дати су у ангстремима, а поред растојања, за сваки пар аминокиселина забележени су и подаци о секундарној структури (*DSSP* кодови) и *SCOP* класи оба резидуума, чиме се омогућава анализа интеракција у функцији структурног контекста. У анализама приказаним у овом раду структурни контекст представљен је секундарном структуром, док је *SCOP* класа задржана у подацима, али није употребљена као критеријум. Уз *DSSP* ознаку бележи се и идентификатор елемента секундарне структуре коме резидуум припада, тако да сви резидууми истог хеликса, односно истог

$\beta$ -ланца, имају исту вредност тог идентификатора. Према том идентификатору растојања се деле на интра-растојања, у оквиру истог елемента секундарне структуре, и интер-растојања, између два различита елемента. Све датотеке деле исту основну структуру колона, при чему датотеке *maxbezh* и *minbezh* додатно садрже и ознаке атома између којих је измерено растојање. Значење колона дато је у табели 2.

Табела 2: Опис колона у датотекама са растојањима (*casa*, *maxbezh*, *minbezh*)

| Колона            | Опис                                                         | Датотеке                           |
|-------------------|--------------------------------------------------------------|------------------------------------|
| <i>protein</i>    | Идентификатор протеина                                       | све                                |
| <i>lanacaa1</i>   | Ознака ланца првог резидуума                                 | све                                |
| <i>aa1</i>        | Тип прве аминокиселине                                       | све                                |
| <i>indeksaa1</i>  | Редни број (позиција) првог резидуума у ланцу                | све                                |
| <i>ss1</i>        | Секундарна структура првог резидуума ( <i>DSSP</i> код)      | све                                |
| <i>idss1</i>      | Идентификатор елемента секундарне структуре првог резидуума  | све                                |
| <i>sclass1</i>    | <i>SCOP</i> класа првог резидуума                            | све                                |
| <i>lanacaa2</i>   | Ознака ланца другог резидуума                                | све                                |
| <i>aa2</i>        | Тип друге аминокиселине                                      | све                                |
| <i>indeksaa2</i>  | Редни број (позиција) другог резидуума у ланцу               | све                                |
| <i>ss2</i>        | Секундарна структура другог резидуума ( <i>DSSP</i> код)     | све                                |
| <i>idss2</i>      | Идентификатор елемента секундарне структуре другог резидуума | све                                |
| <i>sclass2</i>    | <i>SCOP</i> класа другог резидуума                           | све                                |
| <i>aa1atom</i>    | Ознака атома првог резидуума измереног растојања             | <i>maxbezh</i> ,<br><i>minbezh</i> |
| <i>aa2atom</i>    | Ознака атома другог резидуума измереног растојања            | <i>maxbezh</i> ,<br><i>minbezh</i> |
| <i>rastojanje</i> | Измерено растојање између резидуума (у Å)                    | све                                |

Наведена три типа растојања одређена су координатама атома, па се могу израчунати за сваку структуру за коју су те координате познате. У скупу података *DistAA* растојања су већ дата, због чега њихово израчунавање није део овог рада.

## Глава 3

# Технологије

Графовски модел представља природан начин моделовања растојања између аминокиселина у протеинима, па је избор технологија пошао од њега. У наставку су описане основе *NoSQL* база података, затим графовски модел података и његова применљивост на протеинске структуре, и на крају графовски систем за управљање базом података (ГСУБП) *Neo4j*, који је коришћен у овом раду. Уз њега су дати и општи описи преосталих коришћених технологија: веб оквира *Django*, технологија клијентског дела и платформе *Docker*. Начин на који су употребљене у изради апликације, заједно са верзијама, описан је у наредној глави.

### 3.1 NoSQL базе података и графовски модел

Нерелационе базе података не почивају на релационом моделу података, или се барем не придржавају строго тог модела. Карактерише их лако распоређивање података на више сервера и хоризонтална скалабилност, као и способност да поднесу значајне промене шеме без великих оперативних трошкова [19, 21]. Нерелационе базе података називају се и *NoSQL* базе података, што сугерише да не користе *SQL*; с друге стране, неке се служе језиком налик на *SQL*, па се назив ових база података тумачи и као „*not only SQL*“ [21].

Приликом пројектовања дистрибуираних система често се разматрају три важна својства: конзистентност (енгл. *Consistency*), расположивост (енгл. *Availability*) и толеранција на мрежне партиције (енгл. *Partition Tolerance*). Према CAP теореме, дистрибуирани систем не може истовремено гарантовати сва три својства у случају настанка мрежне партиције. Пошто се у

практичним дистрибуираним системима могућност прекида комуникације између чворова не може искључити, систем у таквим условима мора направити компромис између конзистентности и расположивости. Поједини *NoSQL* системи предност дају расположивости и ослањају се на слабије моделе конзистентности, укључујући *BASE* принципе (енгл. *Basically Available, Soft State, Eventually Consistent*), прихватајући привремену неконзистентност података ради веће расположивости и једноставнијег рада у дистрибуираном окружењу [21].

Пословне веб апликације традиционално почивају на релационим базама података. Међутим, карактеристике података су се измениле: порасла је количина података и брзина њихових промена (енгл. *velocity*) [19]. У релационим базама података велики број операција спајања (енгл. *JOIN*) може значајно да утиче на перформансе упита, док честе измене шеме представљају додатно оптерећење у процесу одржавања система [20]. *NoSQL* базе података користе се када је прихватљиво жртвовати одређен ниво конзистентности и аутоматску проверу интегритета и када је потребно дистрибуирање података, лако додавање чворова, висока расположивост и флексибилна структура података [21].

Постоји више породица *NoSQL* база података, међу којима су: складишта парова кључ-вредност (енгл. *key-value stores*), која функционишу као велике дистрибуиране хеш-мапе; документно оријентисане базе података (енгл. *document databases*), у којима су подаци организовани у независне документе; колонска складишта (енгл. *column-family*), заснована на табелама са флексибилним скупом колона; и **графовске базе података**, које су предмет овог поглавља.

## 3.2 Предности и примене графовских база података

Граф  $G = (V, E)$  дефинисан је као уређени пар, где је  $V$  (енгл. *vertex*) непразан, коначан скуп чворова, а  $E$  (енгл. *edge*) скуп грана које спајају парове чворова. Графови имају широку примену у информатици, математици, хемији, лингвистици и другим наукама. У контексту графовских база података, чворови представљају ентитете из стварности, а гране њихове односе.

Једноставан графовски модел погодан је за представљање структуре повезаности између ентитета, али често није довољан за опис сложених система из реалног света. У многим применама потребно је, поред саме структуре повезаности, чувати и додатне информације о чворовима и везама. Због тога савремене графовске базе података најчешће користе модел означеног графа са својствима (енгл. *Labeled Property Graph – LPG*), који служи за придруживање ознака и произвољног броја својстава како чворовима, тако и везама. Ознаке се користе за класификацију елемената графа, док се својства представљају као парови кључ-вредност и служе за чување додатних информација о њима [2].

*LPG* модел, поред интуитивности, има и следеће карактеристике [2]:

- садржи чворове и гране;
- чворови садрже својства у облику парова (кључ, вредност);
- чворови могу бити означени са једном или више ознака;
- везе су именоване и усмерене;
- везе могу садржати својства.

Оваква структура модела погодна је за представљање сложених система у којима су односи између ентитета подједнако важни као и сами ентитети. Захваљујући могућности придруживања својстава и чворовима и везама, *LPG* модел се користи за природно моделовање домена у којима су потребни додатни атрибути над везама, што га чини погодним за анализу просторних односа у протеинским структурама [2].

### 3.3 ГСУБП Neo4j

*Neo4j* је графовски систем за управљање базом података отвореног кода, имплементиран у програмском језику *Java*, и сматра се референтним софтвером у овој области [8]. Међу главним карактеристикама ГСУБП истичу се: подршка за трансакције са пуним *ACID* својствима, интуитиван графички интерфејс и хоризонтална скалабилност читања у *Enterprise* верзији [8, 15].

Важно је напоменути да хоризонтална скалабилност система *Neo4j* није подједнако изражена код операција читања и уписа. У кластерском окружењу, ради очувања трансакционих гаранција и конзистентности реплика, уписи у једну базу података координишу се између примарних копија базе. У сваком тренутку једна од примарних копија има улогу писца и преко ње се обрађују трансакције које мењају податке. Због тога се не може једноставно расподелити на више сервера оптерећење настало операцијама уписа у базу података ради повећања пропусности, као што је то могуће код операција читања. У том погледу *Neo4j* нема предност карактеристичну за поједине *NoSQL* системе код којих се уписи хоризонтално расподељују, већ је његов начин репликације и обраде уписа ближи приступу који се примењује код дистрибуираних релационих СУБП. Додавањем секундарних копија пре свега се повећава пропусност читања, док хоризонтално скалирање уписа захтева партиционисање података [15].

### Чворови и везе

Основне градивне јединице *Neo4j* базе података су чворови (енгл. *nodes*) и везе (енгл. *relationships*), који заједно чине *LPG* модел [19]. Чворови садрже својства у облику парова (кључ, вредност) и могу имати једну или више ознака (енгл. *labels*) које одређују њихов тип. Везе су увек **усмерене** и **именоване**, и такође могу садржати својства [19]. Ознаке чворова одређују њихов тип у моделу података и користе се за издвајање чворова током извршавања упита.

### Cypher упити

Упити над *Neo4j* базом података пишу се у језику ***Cypher*** [19], који је развила компанија *Neo Technology*. *Cypher* је декларативан упитни језик чија је синтакса заснована на визуелном описивању образаца у графу помоћу *ASCII* симбола, што га чини читљивим и интуитивним. Захваљујући томе, сложени упити који би у релационој бази података захтевали вишеструке операције *JOIN* могу се у упитном језику *Cypher* изразити компактно [8].

*Cypher* омогућава проналажење података дефинисањем образаца које чворови и везе у графу треба да задовоље. Приликом извршавања упита, *Cypher* планер настоји да пронађе полазне тачке које највише ограничавају простор претраге, након чега се проверавају суседни чворови и везе који одговара-

ју задатом обрасцу. У том процесу значајну улогу имају индекси, којима се ефикасније проналазе почетни елементи и смањује простор претраге током извршавања упита.

Основу *Cypher* упита најчешће чине клаузуле **MATCH**, **WHERE** и **RETURN**, којима се проналазе, издвајају и приказују подаци из графа. За измену графа користе се клаузуле **CREATE**, **MERGE**, **SET** и **DELETE**: клаузула **CREATE** увек прави нове чворове или везе, док **MERGE** прво проверава да ли задати образац постоји и прави га само ако не постоји. Клаузулом **SET** мењају се већ постојећи елементи графа: постављају се нове или ажурирају постојеће вредности својстава чворова и веза, а могу се додавати и ознаке чворовима. Клаузула **WITH** служи за пренос међурезултата између делова упита и писање сложенијих *Cypher* упита.

Поред наведених клаузула, у сложенијим упитима често се користе и **UNWIND** и **CALL**. Клаузула **UNWIND** разлаже листу на појединачне редове и користи се за засебну обраду сваког елемента, док клаузула **CALL** служи за позивање процедура и извршавање подупита, укључујући и подупите који се извршавају у засебним трансакцијама.

Општи облик упита за читање података из графа приказан је у листингу 1.

Листинг 1: Издвајање података из графа

```
MATCH (n:Label1)-[r:REL_TYPE]->(m:Label2)
WHERE n.property = 'value' AND m.number > 0
RETURN n.property AS a, r.value AS value, m.property AS b
ORDER BY value DESC
LIMIT 10
```

Приликом састављања упита клаузуле се наводе у складу са њиховом улогом. Клаузулом **MATCH** задаје се образац који се тражи у графу, при чему образац може обухватити чворове, везе и њихова својства. Обрасци могу бити различите сложености, од појединачног чвора до више чворова повезаних одређеним типовима веза. Навођењем ознака чворова, типова и смерова веза и вредности својстава додатно се одређује који елементи графа могу одговарати задатом обрасцу. На тај начин се описује структура дела графа који треба пронаћи [15]. Клаузулом **WHERE** том обрасцу задају се додатни услови којима се он ограничава [15]. Клаузула **RETURN** одређује податке који се укључују у резултат, након чега се клаузулом **ORDER BY** резултат може уредити, а

клаузулом LIMIT ограничити [15]. У самом обрасцу чворови се пишу у облику, а везе у угластим заградама, при чему стрелица одређује смер везе. Уколико се стрелица изостави, образац се подудара са везом без обзира на њен смер.

У листингу 1, клаузула MATCH проналази два чвора са ознакама Label1 и Label2, повезана везом типа REL\_TYPE. Стрелица у обрасцу усмерена је од чвора n ка чвору m, па се у резултат укључују само везе тог смера. Да је веза записана без стрелице, у облику (n:Label1)-[r:REL\_TYPE]-(m:Label2), образац би обухватио везе тог типа без обзира на њихово усмерење. Клаузула WHERE задржава само парове код којих својство n.property има вредност 'value', а m.number је веће од нуле. Као резултат се враћају својства n.property, r.value и m.property, под називима a, value и b. Резултати су сортирани опадајуће према value и ограничени на десет редова.

У листингу 2 најпре се проналазе или праве два чвора са вредностима својства id 1 и 2. Затим се између њих проналази или прави веза типа REL\_TYPE. Клаузулом SET првом чвору се поставља својство property, а вези својство value.

Листинг 2: Прављење и повезивање чворова уз измену својстава

```
MERGE (n:Label {id: 1})
SET n.property = 'value'
MERGE (m:Label {id: 2})
MERGE (n)-[r:REL_TYPE]->(m)
SET r.value = 10
```

Издвојени упити над базом података која се користи у овом раду приказани су у главама 4 и 5.

Поред приказаних клаузула, *Cypher* подржава и клаузулу OPTIONAL MATCH, која је аналогна операцији левог спољашњег спајања, односно операцији LEFT JOIN у релационим базама података. За разлику од клаузуле MATCH, која враћа само редове за које је задати образац пронађен, OPTIONAL MATCH не изоставља ред из резултата када образац не постоји, већ за недостајуће вредности враћа null. Ова клаузула није коришћена у овом раду, али је значајна за упите у којима одређена веза или чвор не морају постојати.

## Индекси и план извршавања

Индекс представља копију чвора, својства или везе у виду посебне структуре података (каталога), која пресликава вредности својстава на чворове и везе у бази података. Индекси се трајно чувају на диску. Приликом извршавања упита држе се у оперативној меморији, чиме се избегава читање са диска. Систем за управљање базом података брине се о индексима аутоматски, тако што их ажурира, пуни и брише.

ГСУБП разликује две врсте индекса: индексе за побољшање перформанси претраге (енгл. *search performance index*) и семантичке индексе (енгл. *semantic index*) [8]. У оквиру индекса за побољшање перформанси претраге постоји неколико типова [15]:

- **Индекс опсега** (енгл. *range index*): подржава предикате поређења, претрагу по опсегу, претрагу по префиксу и проверу припадности листи; прави се подразумевано ако тип индекса није наведен;
- **Текстуални индекс** (енгл. *text index*): оптимизован за предикате над стринговима;
- **Индекс тачака** (енгл. *point index*): намењен за просторне податке који су типа тачке  $(x, y, z)$ ;
- **Индекс ознака и типова веза** (енгл. *token lookup index*): два таква индекса аутоматски се праве приликом иницијализације базе података: копија свих чворова и копија свих веза; овај индекс нема информацију о вредностима својстава.

Када је на располагању више индекса, *Cypher* планер бира индекс који омогућава најефикасније извршавање задатог предиката. На тај начин индекси се користе за ефикасно проналажење почетних чворова од којих започиње претрага образаца задатих клаузулом MATCH. На избор индекса може се утицати и експлицитним навођењем наговештаја (енгл. *index hint*) USING INDEX [8].

Препорука је увек навести назив индекса приликом прављења, јер у супротном добија насумично генерисано име. Употреба клаузуле IF NOT EXISTS спречава грешку ако индекс са истим именом већ постоји, а прављење индекса захтева INDEX привилегију. Примери индекса су (листинг 3):

Листинг 3: Наредбе за прављење индекса коришћених у раду

```
CREATE INDEX aa_index IF NOT EXISTS
FOR (a:AminoAcid) ON (a.protein, a.chain, a.index);

CREATE INDEX aa_ss IF NOT EXISTS
FOR (a:AminoAcid) ON (a.ss);
```

Првом наредбом прави се композитни *range* индекс над својствима **protein**, **chain** и **index** чворова типа **AminoAcid**, што омогућава ефикаснију претрагу аминокиселина према протеину, ланцу и позицији у секвенци. Другом наредбом прави се *range* индекс над својством **ss**, чиме се убрзава издвајање аминокиселина према типу секундарне структуре, што је један од филтера доступних у корисничком интерфејсу.

Иста комбинација својстава као у првом индексу одређује идентитет аминокиселине при увозу података, где се наредбом **MERGE** проверава да ли чвор већ постоји. Без индекса, свака таква провера захтевала би преглед свих чворова типа **AminoAcid**, због чега је увоз података из *CSV* датотека временски знатно захтевнији. Уз индекс, провера постојања своди се на претрагу по индексу и увоз је осетно бржи.

За анализу плана извршавања упита *Neo4j* нуди опције **EXPLAIN** и **PROFILE**. **EXPLAIN** приказује план извршавања, при чему се сам упит не извршава, и показује које операције ће бити извршене и који индекси ће бити искоришћени. **PROFILE** извршава упит и уз резултате враћа стварно измерене метрике за сваки корак плана, чиме се идентификују уска грла [8].

### 3.4 Остале коришћене технологије

Поред графовске базе података, у изради решења коришћени су веб оквир *Django* за серверски слој, стандардне технологије клијентског дела (*HTML*, *CSS* и *JavaScript*) и платформа *Docker*, у којој се делови система покрећу као међусобно повезани сервиси.

#### Веб оквир Django

*Django* је веб оквир отвореног кода, писан у програмском језику *Python*, који пружа готове компоненте за задатке заједничке већини веб апликација:

рутирање захтева, рад са базом података, обраду образаца и генерисање одговора [6, 12]. Обрада захтева организована је тако да се *URL* адреса пресликава у поглед (енгл. *view*) – функцију или класу која прима захтев и враћа одговор, било као *HTML* страницу добијену из шаблона, било као податке у формату *JSON*. *Django* не садржи уграђену подршку за графовске базе података, па се комуникација са ГСУБП *Neo4j* остварује званичним драјвером за програмски језик *Python*. Драјвер успоставља везу протоколом *Bolt*, отвара сесију, шаље *Cypher* упит са именованим параметрима и враћа записе резултата. За израду програмског интерфејса апликације (енгл. *application programming interface*, *API*), скупа крајњих тачака *API*-ја (енгл. *endpoint*) преко којих други програми приступају њеним функционалностима, користи се проширење *Django REST Framework*, које преузима серијализацију података и обраду захтева по крајњим тачкама [12].

### Технологије клијентског дела

Кориснички део веб апликације имплементиран је употребом језика *HTML* и *CSS* и програмског језика *JavaScript*. *HTML* описује структуру странице, а *CSS*-ом је дефинисан њен изглед. *JavaScript* се извршава у веб прегледачу и садржај странице се приказује без поновног освежавања. Подаци које клијентски део добија од серверског најчешће су у формату *JSON*.

За тродимензионални приказ молекула у веб прегледачу користи се библиотека *3Dmol.js*, објектно-оријентисана библиотека заснована на технологији *WebGL* (енгл. *Web Graphics Library*) за исцртавање тродимензионалне графике [10, 17]. Поред уобичајених приказа молекула, библиотека допушта и програмско прављење и исцртавање произвољних геометријских облика, као што су сфере и цилиндри [17]. Управо та могућност, уз једноставну интеграцију у веб апликацију и приказивање тродимензионалних структура непосредно у веб прегледачу, била је разлог за њен избор у овом раду, будући да се граф модел формира динамички, на основу података из базе.

### Docker

*Docker* је платформа која омогућава покретање апликација и њихових зависности у изолованим контејнерима. Контејнер обједињује апликацију и окружење у коме се она извршава, па се исто решење покреће на исти начин

без обзира на подешавања појединачне машине [4]. Настаје из слике (енгл. *image*), записа у коме су све зависности већ инсталиране, тако да свако покретање полази од истог стања. За разлику од виртуелне машине, слика дели језгро оперативног система са домаћином и стога је знатно лакша и мање захтевна по ресурсима [4]. Решења састављена од више међусобно повезаних делова описују се у *Docker Compose* конфигурацији, у којој сваки део чини засебан сервис са сопственом сликом, променљивима окружења и трајним просторима за складиштење, а сви се покрећу и заустављају заједно [7].

## Глава 4

# Опис апликације

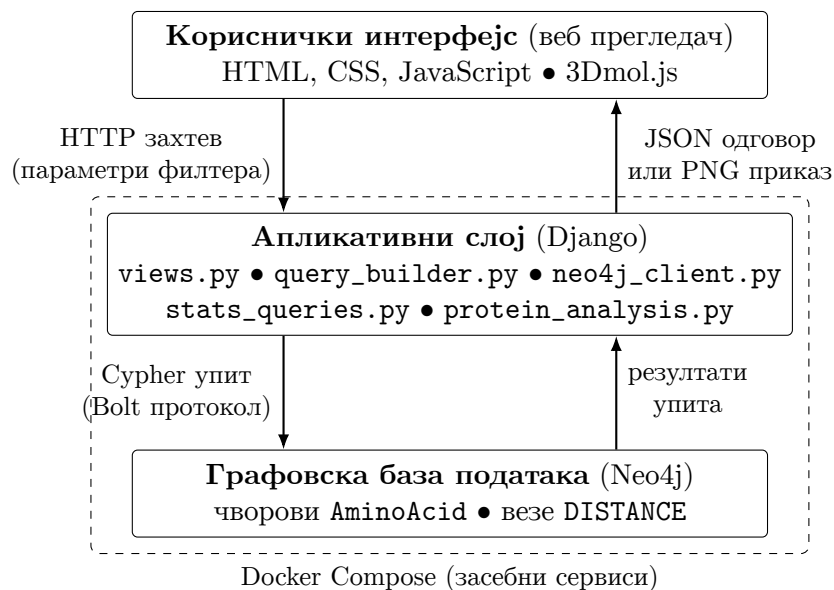
У претходним поглављима приказани су подаци о растојањима између аминокиселина, графовски модел погодан за њихово чување и технологије којима је решење изграђено. У наставку је описано како су те технологије употребљене у апликацији. Најпре је представљена архитектура решења, затим граф модел и поступак увоза података у графовску базу података, потом апликативни слој реализован помоћу веб оквира *Django*, и на крају кориснички интерфејс заснован на библиотеци *3Dmol.js*.

### 4.1 Архитектура решења

Апликација је развијена тако да има три слоја: кориснички интерфејс, апликативни слој и слој података. Кориснички интерфејс реализован је употребом *HTML*, *CSS* и *JavaScript* технологија, док је апликативни слој имплементиран применом оквира *Django*. За складиштење и анализу података о протеинским структурама коришћен је поменути ГСУБП *Neo4j*. Целокупно решење реализовано је коришћењем технологије *Docker*, при чему су *Django* апликација и ГСУБП *Neo4j* покренути у засебним контејнерима као сервиси дефинисани *Docker Compose* конфигурацијом [7].

Корисник интерагује са системом путем веб прегледача, где је могуће одабрати протеин и подесити различите параметре. Захтеви се шаљу *Django* серверу посредством *HTTP* позива. Резултати упита се затим обрађују и враћају клијенту у облику *JSON* одговора или генерисаних графичких приказа. Комуникација између слојева приказана је на слици 1.

Кориснички интерфејс задужен је за интеракцију са корисником, односно



Слика 1: Архитектура решења и комуникација између слојева

за прикупљање параметара анализе и приказ добијених резултата. Он не садржи логику анализе нити зависи од начина на који су подаци организовани у бази.

Апликативни слој прихвата унете параметре, проверава њихову исправност и на основу њих формира *Cypher* упите, које прослеђује слоју података на извршавање. Враћене резултате прилагођава облику погодном за приказ у корисничком интерфејсу. Када је потребан графички приказ, из тих резултата генерише слику.

Слој података обухвата графовску базу у којој се чувају подаци о аминокиселинама и везама међу њима. У овом слоју се извршавају упити, а њихов резултат враћа се апликативном слоју. На тај начин су приказ, обрада захтева и приступ подацима међусобно раздвојени, па измене у организацији базе не морају непосредно да утичу на кориснички интерфејс, као што ни измене начина приказа не захтевају промену самих упита.

Употреба технологије *Docker* у овом раду има више разлога. База података и апликативни слој захтевају различита окружења, а *Docker Compose* омогућава да се покрећу као засебни, међусобно повезани сервиси. Покретање *Django* апликације може се условити доступношћу сервиса *Neo4j*, чиме се избегава покушај успостављања везе пре него што база података буде спремна за рад. Подаци се чувају у именованом простору за складиштење (енгл.

*named volume*), па увезена база података остаје сачувана између покретања, што је значајно јер је сам увоз временски захтеван. Фиксиране су и верзије базе података, програмског језика и библиотека, чиме је омогућено да се резултати приказани у раду репродукују без ручног подешавања окружења.

Ради поновљивости резултата наводе се верзије коришћених технологија. Као база података коришћен је *Neo4j* 2025.09.0, издање *Enterprise*, а приступ бази података остварен је званичним *Python* драјвером *neo4j* верзије 6.0.2. Апликативни слој реализован је у програмском језику *Python* 3.12, помоћу веб оквира *Django* 5.2.7, уз *Django REST Framework* 3.16.1 за изградњу *HTTP API*-ја, док се графички прикази генеришу помоћу библиотека *Matplotlib* 3.10.8 и *seaborn* 0.13.2, уз *NumPy* и *pandas* за обраду података. Апликација и база података покрећу се као засебни сервиси *Docker Compose* конфигурације, при чему се *Django* апликација извршава преко *WSGI* сервера *gunicorn*.

## 4.2 Граф модел и увоз података

Графовске базе података погодне су за домене у којима односи између ентитета представљају значајан део самог модела. Подаци о протеинским структурама моделовани су као граф у којем чворови представљају аминокиселине, док везе представљају просторне односе између њих. Оваква организација података природно одговара проблему анализе протеинских структура, јер је могуће непосредно представљање интеракција између аминокиселина и проналажење њихових суседстава на основу различитих критеријума и међусобних односа. Просторни односи су тиме непосредно представљени у структури базе података и могу се претраживати и визуелизовати без њиховог накнадног реконструисања из више табела.

Основни елемент модела је чвор типа *AminoAcid*. Сваки чвор описује једну аминокиселину и садржи следећа својства: идентификатор протеина, ознаку ланца и позицију у секвенци, као и додатне информације као што су назив аминокиселине, ознака секундарне структуре и координате  $C\alpha$  атома. Позиција у секвенци задржана је онако како је дата у полазним подацима, па се вредности индекса не морају поклапати са редним бројем резидуума у ланцу. Између парова аминокиселина формиране су везе типа *DISTANCE*, које чувају вредност растојања и његов тип. Свака веза усмерена је од аминокиселине са мањом ка аминокиселини са већом позицијом у секвенци. У бази података су

моделована три типа растојања: растојање између  $C\alpha$  атома (*saca*), минимално растојање између тешких атома (*minbezh*) и максимално растојање између тешких атома (*maxbezh*). Код минималних и максималних растојања додатно се чувају називи атома који формирају одговарајуће растојање. Преглед својстава чворова и веза граф модела, заједно са одговарајућим типовима података, дат је у табели 3. Својство *sclass* чува *SCOP* класу резидуума и преузето је из полазног скупа података ради потпуности модела, али га ниједна од анализа приказаних у глави 5 не користи као критеријум.

Табела 3: Својства чворова и веза граф модела

| Елемент   | Својство       | Тип     |
|-----------|----------------|---------|
| AminoAcid | protein        | String  |
|           | chain          | String  |
|           | index          | Integer |
|           | name           | String  |
|           | ss             | String  |
|           | ss_id          | String  |
|           | sclass         | String  |
|           | ca_coordinates | Point   |
| DISTANCE  | type           | String  |
|           | value          | Float   |
|           | atom1          | String  |
|           | atom2          | String  |

Својство *ss\_id* издваја појединачне елементе секундарне структуре: све резидууми истог непрекинутог сегмента добијају исту вредност, док суседни сегменти добијају различите. Ознака *ss* казује само тип структуре, па се тек уз *ss\_id* разликују интра-растојања, између резидуума истог елемента (на пример, два остатка унутар истог  $\alpha$ -хеликса), од интер-растојања, између резидуума различитих елемената (на пример, остатка једног хеликса и остатка другог хеликса или  $\beta$ -плоче).

Подаци су у базу података увезени из *CSV* датотека употребом *Cypher* наредбе *LOAD CSV*. Током увоза чворови и везе направљени су коришћењем наредбе *MERGE*, чиме је обезбеђено да поновно покретање поступка увоза не доведе до дуплирања података. Идентитет аминокиселине дефинисан је комбинацијом идентификатора протеина, ланца и позиције у секвенци, чиме се јединствено одређује сваки чвор у оквиру базе података. Због великог обима података, увоз је извршаван поступно, у више трансакција фиксне величи-

не, чиме је смањено меморијско оптерећење и омогућена стабилнија обрада података.

База, формирана на овај начин, доступна је јавно и може се преузети и користити у даљим истраживањима.<sup>1</sup> Поред растојања, за сваки резидуум садржи и ознаку секундарне структуре и *SCOP* класу.

### Упити над базом података

Растојања између аминокиселина у изабраном протеину добијају се упитом који проналази парове чворова са ознаком `AminoAcid`, међусобно повезаних везом типа `DISTANCE` задате врсте. Својство `protein` оба чвора мора одговарати изабраном протеину, а опциони филтер по ланцу ограничава анализу на одређени ланац. За сваки пронађени пар враћају се називи аминокиселина, њихове позиције у секвенци, као и вредност растојања. Резултати су сортирани према позицијама аминокиселина у секвенци (листинг 4).

Листинг 4: Издвајање одређеног типа растојања између аминокиселина изабраног протеина

```
MATCH (a1:AminoAcid)-[d:DISTANCE{type:$type}]->(a2:AminoAcid)
WHERE a1.protein = $protein AND a2.protein = $protein
      AND ($chain IS NULL OR a1.chain = $chain)
RETURN a1.name AS aa1, a1.index AS idx1,
        a2.name AS aa2, a2.index AS idx2,
        d.value AS distance
ORDER BY a1.index, a2.index ASC
```

Статистика растојања између елемената секундарне структуре добија се агрегирањем растојања по паровима ознака секундарне структуре. Укључују се само аминокиселине изабраног протеина за које је својство `ss` дефинисано. Сваки такав пар аминокиселина сачуван је у бази података само једном, везом усмереном од аминокиселине са мањом ка аминокиселини са већом позицијом у секвенци, па у агрегацију улази тачно једном. Како би се парови ознака секундарне структуре који се разликују само по редоследу посматрали као иста комбинација, израз `CASE WHEN` користи се за њихово свођење на исти облик. За сваки добијени пар израчунавају се просечна, минимална и мак-

---

<sup>1</sup><https://github.com/emilijastevanovic/protein-distance-graph-db>

симална вредност растојања, стандардна девијација, као и медијана помоћу агрегатних функција `avg`, `min`, `max`, `stdev` и `percentileCont` (листинг 5).

Листинг 5: Израчунавање статистике растојања по паровима елемената секундарне структуре

```
MATCH (a1:AminoAcid)-[d:DISTANCE{type:$type}]->(a2:AminoAcid)
WHERE a1.protein = $protein AND a2.protein = $protein
      AND a1.ss IS NOT NULL AND a2.ss IS NOT NULL
WITH CASE WHEN a1.ss <= a2.ss
          THEN a1.ss + ' ' + a2.ss
          ELSE a2.ss + ' ' + a1.ss
      END AS ss_pair,
      d.value AS dist
RETURN ss_pair,
       count(dist) AS count,
       round(avg(dist), 3) AS mean,
       round(min(dist), 3) AS min,
       round(max(dist), 3) AS max,
       round(stdev(dist), 3) AS std,
       round(percentileCont(dist, 0.5), 3) AS median
ORDER BY count DESC
```

За анализу састава протеина користи се упит који приказује расподелу аминокиселина у изабраном протеину. Чворови се најпре издвајају према идентификатору протеина и, опционо, према ланцу, а затим се групишу по називу аминокиселине. За сваку аминокиселину рачуна се број појављивања и њен процентуални удео у односу на укупан број аминокиселина. Функција `collect` користи се за привремено формирање листе агрегираних резултата, док `UNWIND` ту листу разлаже назад у редове ради израчунавања и приказа удела (листинг 6).

Листинг 6: Бројчана и процентуална расподела аминокиселина у изабраном протеину

```
MATCH (aa:AminoAcid)
WHERE aa.protein = $protein
      AND ($chain IS NULL OR aa.chain = $chain)
WITH aa.name AS name, count(*) AS cnt
```

```
WITH collect({name: name, cnt: cnt}) AS rows,  
      sum(cnt) AS total  
UNWIND rows AS row  
RETURN row.name AS name, row.cnt AS cnt,  
       round(toFloat(row.cnt) * 100 / total, 1) AS pct,  
       total  
ORDER BY row.cnt DESC
```

### 4.3 Серверски део апликације: Django и крајње тачке API-ја

Серверски део (енгл. *backend*) реализован је помоћу оквира *Django*. Његова основна улога је прихватање захтева са корисничког интерфејса, формирање *Cypher* упита и њихово прослеђивање ГСУБП *Neo4j*, који их извршава и враћа резултат, као и припрема тог резултата у облику који је погодан за приказ на клијентском делу апликације (енгл. *frontend*).

Комуникација између клијентског и серверског слоја одвија се посредством *HTTP API*-ја. Главне функционалности система доступне су кроз више крајњих тачака:

- `/api/graph3d/`: дохватање чворова и веза у циљу тродимензионалне визуализације графа;
- `/api/stats/`: израчунавање статистичких показатеља над изабраним протеином или ланцем;
- `/api/heatmap/distance/`: генерисање матрице растојања између аминокиселина у облику *PNG* слике;
- `/api/heatmap/segments/`: генерисање хистограма дужина сегмената секундарне структуре;
- `/api/global_stats/`: приказ глобалних статистика над целокупном базом података.

Крајња тачка `/api/graph3d/` представља основу интерактивне визуализације и враћа податке о аминокиселинама и њиховим међусобним растојањи-

ма. Поред тога, крајња тачка `/api/stats/` омогућава израчунавање различитих статистичких показатеља над изабраним протеином, ланцем и типом растојања. Подржане анализе обухватају одређивање састава аминокиселина, приказ секвенце аминокиселина и секундарних структура, статистику растојања, анализу интеракција између парова секундарних структура, као и идентификацију издвојених (изолованих) аминокиселина на основу њихових просечних растојања од осталих резидуума. За графичке приказе користе се посебне крајње тачке које генеришу *PNG* слике на основу резултата добијених из ГСУБП-а *Neo4j*.

Логика за формирање *Cypher* упита, комуникацију са ГСУБП-ом *Neo4j* и израчунавање статистика издвојена је у посебне модуле, чиме су ове одговорности одвојене од обраде *HTTP* захтева. Модул `query_builder.py` задужен је за динамичко формирање параметризованих *Cypher* упита и додавање филтера према изабраном протеину, ланцу, аминокиселинама, секундарној структури и опсегу растојања. Комуникација са ГСУБП-ом *Neo4j* централизована је у модулу `neo4j_client.py`, који управља драјвером и извршавањем упита. Драјвер се прави једном, при првом упиту, а параметри везе преузимају се из променљивих окружења. Сваки упит извршава се у сесији отвореној у режиму за читање, а резултат се одмах преузима у листу записа, након чега се сесија затвара. Вредности које корисник задаје не уграђују се у текст упита, већ се драјверу предају као именовани параметри. Уколико ГСУБП *Neo4j* није доступан, апликација не прекида рад, већ враћа одговор о грешци са статусом 503. Статистички упити и припрема података за анализу распоређени су у модуле `stats_queries.py` и `protein_analysis.py`, који обухватају израчунавање састава аминокиселина, расподеле секундарних структура, статистика растојања и парова секундарних структура, као и припрему и генерисање графичких приказа. Модул `views.py` задржава улогу *HTTP* слоја: преузима параметре захтева, позива одговарајуће функције из наведених модула и формира одговор у облику *JSON* података или слике. Оваква подела одговорности поједностављује одржавање постојећег кода и додавање нових типова анализа.

У кешу апликативног слоја чувају се резултати чије израчунавање траје најдуже. Глобалне статистике базе података рачунају се над свим структурама у бази података, па се њихов резултат чува шест сати, а при покретању апликације припрема се унапред, у позадинској нити. Слике матрица растоја-

ња и хистограма дужина сегмената генеришу се на серверу и чувају 24 сата, при чему се кључ кеша формира од параметара упита сажетих у *SHA-256* вредност. Остали упити се не кеширају, будући да се извршавају над једним протеином и да њихово трајање не оправдава додатну сложеност.

Коришћен је *Django* кеш у меморији процеса. То значи да се садржај губи при поновном покретању апликације и да се не дели између радних процеса. Пошто *Docker Compose* конфигурација покреће један радни процес, та ограничења у тренутној поставци не долазе до изражаја.

### 4.4 Клијентски део апликације: 3Dmol.js, филтери и кориснички интерфејс

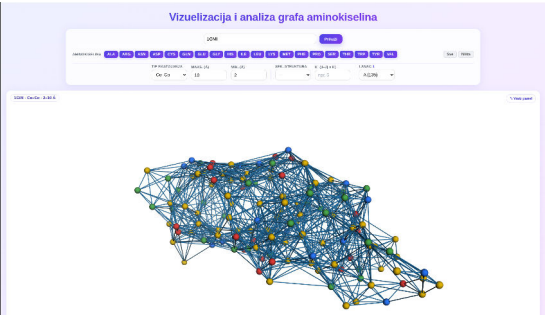
Главни прозор апликације приказан је на слици 2. У горњем делу налазе се поља и филтери за избор протеина, ланца и аминокиселина, као и типа и опсега растојања, секундарне структуре и минималног секвенцијалног растојања. Испод дела у ком се задају параметри анализе налазе се две целине. Са леве стране је тродимензионални приказ протеина, у ком су аминокиселине представљене сферама на положајима  $C\alpha$  атома, а парови који задовољавају задате услове повезани везама. Са десне стране се налази панел са статистичким приказима. На врху је картица са основним подацима о целој бази, испод чега су картице Преглед, Чвор, Секвенца, Растојања, *SS* парови, Сегменти и Екстреми, у оквиру којих су приказане табеле, топлотне мапе и хистограми. Део приказаних показатеља односи се на цео протеин, док статистике изабраног опсега прате постављене филтере и описују исти скуп парова који је приказан у графу. Приказ графа и панел са статистикама могу се проширити преко целе ширине прозора, чиме се изабрани део интерфејса добија у већем формату, што приказују слике 3 и 4.

Фокус корисничког интерфејса је тродимензионални приказ графа аминокиселина и растојања међу њима у изабраном протеину, који је реализован уз помоћ библиотеке *3Dmol.js*.

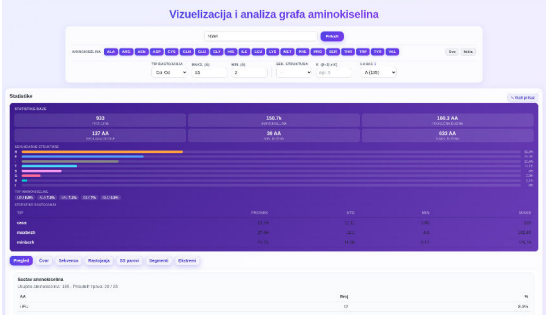
У оквиру развијене апликације библиотека *3Dmol.js* коришћена је за приказ графа добијеног из *Neo4j* базе података. Аминокиселине су представљене као сфере чији положај у простору одговара стварним координатама њихових  $C\alpha$  атома, добијеним кристалографским методама и сачуваним у бази података као својство *ca\_coordinates*. Приказ стога није шематски распоред



Слика 2: Главни прозор апликације са тродимензионалним приказом графа аминокиселина, филтерима и панелом за анализу података.



Слика 3: Проширени приказ графа



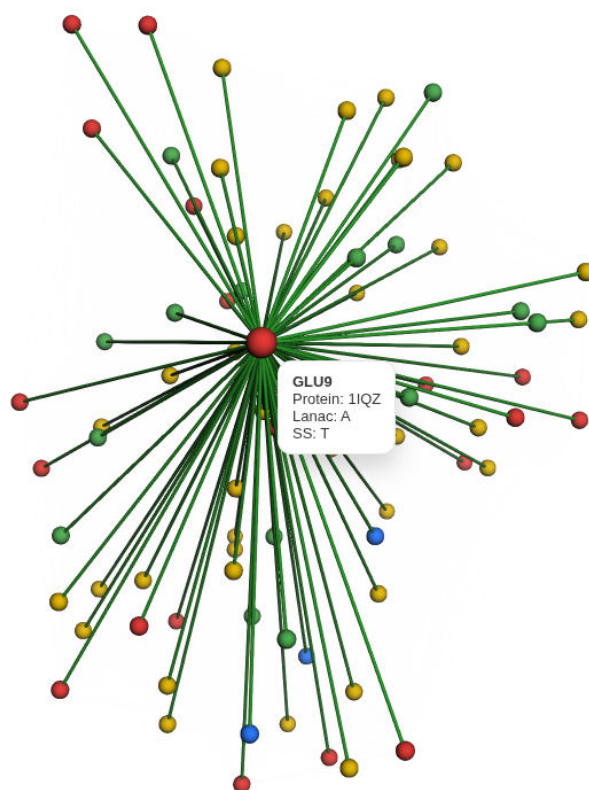
Слика 4: Проширени панел са статистицама

чворова какав дају уобичајени алгоритми за цртање графова, већ просторно веран приказ структуре, у коме међусобни односи чворова на екрану одговарају стварним просторним односима у протеину. Растојања између аминокиселинских чворова приказана су цилиндрима који представљају неусмерене гране у графу. Библиотека се користи и за интеракцију са елементима графа, тако да корисник може да ротира и помера сцену, мења ниво увећања и добије додатне информације о изабраним аминокиселинама и њиховим интеракцијама. Избором чвора приказују се додатне информације о аминокиселини, а њени суседи се визуелно издвајају. Чворови су обојени према особини бочног ланца: неполарне аминокиселине су жуте, поларне ненаелектрисане зелене, базне плаве, киселе црвене, а сиви су остаци који нису препознати као стандардне аминокиселине. Слично, избором везе корисник добија детаље као што су тип и вредност растојања.

Избор чвора приказан је на слици 5. Уз одабрану аминокиселину исписују се њен назив и редни број, идентификатор протеина, ознака ланца и ознака секундарне структуре, док се њени суседи и везе ка њима издвајају бојом, чиме се непосредно уочава окружење изабраног резидуума. Избором везе, приказаним на слици 6, добијају се пар аминокиселина које она повезује, тип растојања и његова измерена вредност.

Интерфејс подржава више филтера који ограничавају приказан подскуп података. Корисник може да зада идентификатор протеина, ланац, тип растојања, минималну и максималну вредност растојања, тип секундарне структуре, као и минимално растојање аминокиселина у секвенци. У једном приказу бира се само један тип секундарне структуре. Уколико услови над вредностима растојања нису задати, филтер секундарне структуре примењује се на чворове, односно на приказане аминокиселине. Уколико јесу, у приказу остају само аминокиселине изабраног типа секундарне структуре и парови међу њима који задовољавају остале услове. Избор протеина је неопходан. Променом параметара аутоматски се покреће нови захтев ка серверском делу апликације и освежавају се приказани подаци.

Поред набројаних филтера, свака од двадесет аминокиселина може се засебно укључити или искључити из приказа, што омогућава издвајање подграфа који чине само одабрани типови резидуума. Ред са филтерима приказан је на слици 7, где су аминокиселине ASN и PHE искључене, а задат је опсег растојања од 11 до 17 Å за тип *maxbez*. Дугмад *Све* и *Ништа* служе за брзо



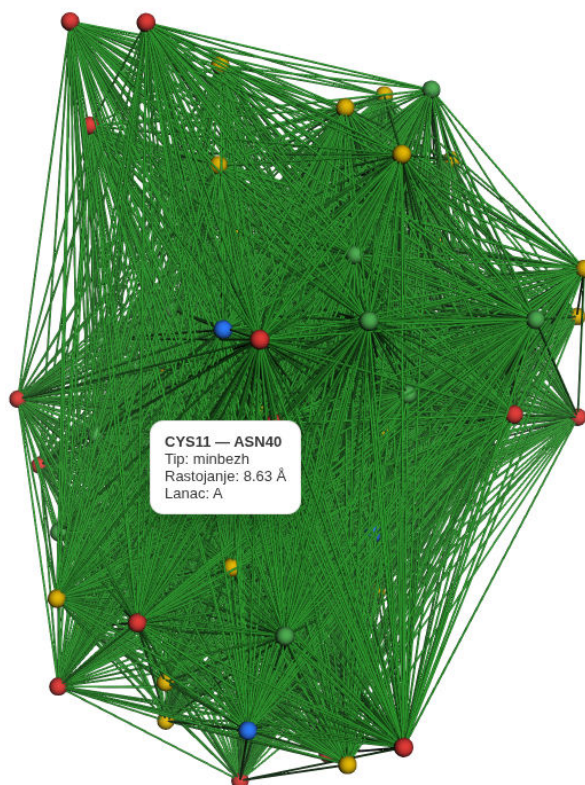
Слика 5: Приказ података о изабраној аминокиселини са истакнутим суседима

укључивање, односно искључивање свих типова одједном.

### Панели за анализу

Поред тродимензионалног приказа, интерфејс садржи панеле са статистицама, секвенцом аминокиселина и анализама секундарне структуре, тако да корисник може непосредно да упореди просторни приказ са нумеричким резултатима.

Панели су организовани у картице, при чему свака картица одговара једној врсти анализе изабраног протеина. Картица *Преглед* даје састав аминокиселина са бројем појављивања и процентуалним уделом сваког типа (слика 8), док картица *Секвенца* приказује примарну секвенцу са ознакама резидуума (слика 9). Приказ секвенце и тока секундарне структуре односи се на цео протеин, независно од изабраног ланца.



Слика 6: Приказ података о изабраној вези између две аминокиселине

AMINOKISELINA **ALA** **ARG** **ASN** **ASP** **CYS** **GLN** **GLU** **GLY** **HIS** **ILE** **LEU** **LYS** **MET** **PHE** **PRO** **SER** **THR** **TRP** **TYR** **VAL** Sve Nista

TIP RASTOJANJA MAKS. (Å) MIN. (Å) SEK. STRUKTURA K (|i-j| ≥ K) LANAC

Maks. bez H 17 11 — npr. 5 —

Слика 7: Филтери корисничког интерфејса: избор аминокиселина, тип и опсег растојања, секундарна структура, најмање растојање у секвенци и ланац



Слика 8: Картица *Пре-глед* Слика 9: Картица *Се-квенца* Слика 10: Картица *Чвор*

Картица *Чвор* садржи анализу окружења појединачне аминокиселине: поред основних података о резидууму, приказују се статистике растојања до њених суседа (најмање, просечно, медијана, стандардна девијација, највеће растојање и број контаката), као и састав суседства по типовима аминокиселина и по типовима секундарне структуре (слика 10). Картица *SS њарови* даје статистике растојања по паровима секундарних структура (слика 11), а картица *Експреми* просечна растојања по типу секундарне структуре и листу просторно најизолованијих резидуума (слика 12). Управо ове две картице коришћене су при изради студија случаја у поглављу 5.

Показатељи се не рачунају сви над истим скупом података. Картица *Чвор* описује изабрану аминокиселину и њено окружење унутар подскупа издвојеног филтерима, док се остале картице рачунају над целим изабраним ланцем, односно над целим протеином ако ланац није изабран, уз поштовање изабраног типа растојања. Панел са глобалним статистикама односи се на целу базу података и не зависи од задатих филтера.

Топлотне мапе растојања генеришу се на серверу, прате задате филтере

Pregled Čvor Sekvenca Rastojanja **SS parovi**

Segmenti Ekstremi

**Statistike SS parova**  
 Ceo protein: granice, k i SS se ne primenjuju.  
 Tip: caca - Parova: 15

| SS par | Broj | Prosek (Å) | Medijana (Å) | Min (Å) | Maks. (Å) | Std   |
|--------|------|------------|--------------|---------|-----------|-------|
| H H    | 1431 | 13.98      | 13.96        | 3.78    | 29.07     | 5.33  |
| - H    | 864  | 17.56      | 17.21        | 3.81    | 36.36     | 6.38  |
| H T    | 756  | 18.45      | 17.63        | 3.82    | 36.25     | 7.13  |
| - T    | 224  | 18.66      | 19.09        | 3.82    | 34.75     | 8.86  |
| G H    | 162  | 19.33      | 19.57        | 6.58    | 30.41     | 5.24  |
| H S    | 162  | 17.45      | 17.45        | 5.09    | 29.30     | 5.80  |
| - -    | 120  | 19.87      | 20.35        | 3.79    | 35.37     | 8.52  |
| T T    | 91   | 16.58      | 15.04        | 3.77    | 33.96     | 9.25  |
| - G    | 48   | 22.98      | 28.39        | 3.83    | 34.83     | 9.98  |
| - S    | 48   | 19.71      | 22.69        | 3.80    | 33.70     | 10.22 |
| G T    | 42   | 25.27      | 28.71        | 9.11    | 37.27     | 10.03 |
| S T    | 42   | 20.80      | 22.12        | 3.82    | 35.43     | 11.02 |
| G S    | 9    | 14.97      | 8.25         | 3.81    | 33.63     | 13.02 |
| G G    | 3    | 4.37       | 3.82         | 3.81    | 5.49      | 0.96  |
| S S    | 3    | 21.66      | 30.00        | 3.83    | 31.14     | 15.45 |

**SS parovi u izabranom opsegu**  
 Ca-Ca rastojanja u opsegu 2–10 Å.

| SS par | Broj | %    | Prosek (Å) | Medijana (Å) | Min (Å) | Maks. (Å) | Std   |
|--------|------|------|------------|--------------|---------|-----------|-------|
| H H    | 309  | 47.0 | 6.768      | 6.172        | 3.783   | 9.996     | 2.072 |

Слика 11: Картица *SS парови*

Pregled Čvor Sekvenca Rastojanja **SS parovi**

Segmenti **Ekstremi**

**Prosečno rastojanje po SS tipu**  
 Tip: caca - SS tipova: 5

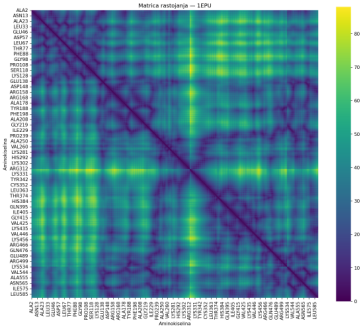
| SS | N rezidua | Prosek (Å) | Medijana (Å) | p90 (Å) | Maks. (Å) |
|----|-----------|------------|--------------|---------|-----------|
| G  | 3         | 20.440     | 20.998       | 21.485  | 21.607    |
| T  | 14        | 18.525     | 17.971       | 21.448  | 22.025    |
| S  | 3         | 18.397     | 18.225       | 18.722  | 18.846    |
| -  | 16        | 18.378     | 18.169       | 20.289  | 21.487    |
| H  | 54        | 15.623     | 15.331       | 17.963  | 20.351    |

**Najmanji prosek rastojanja**  
 Računato do svih ostalih aminokiselina.

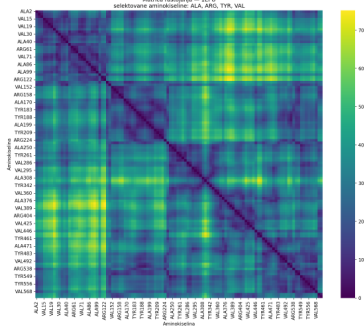
| AA  | Indeks | Lanac | SS | Prosečno rastojanje (Å) |
|-----|--------|-------|----|-------------------------|
| ILE | 38     | B     | H  | 12.869                  |
| LEU | 68     | B     | H  | 13.144                  |
| ALA | 41     | B     | H  | 13.273                  |
| SER | 16     | B     | H  | 13.374                  |
| SER | 65     | B     | H  | 13.428                  |
| LEU | 15     | B     | H  | 13.451                  |
| ILE | 20     | B     | T  | 13.570                  |
| HIS | 42     | B     | H  | 13.719                  |
| LEU | 12     | B     | H  | 13.824                  |
| GLN | 37     | B     | H  | 13.852                  |

Слика 12: Картица *Екстреми*

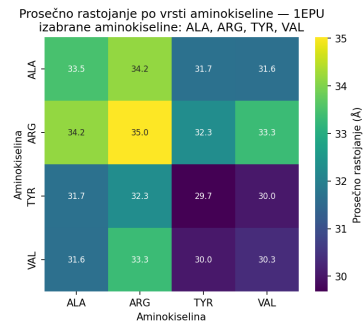
и налазе се у картици *Растојања*, што приказују слике 13–15. Приказ за цео протеин обухвата све парове резидуума, избор појединих врста аминокиселина сужава мапу на њихове међусобне парове, док се додатним сажимањем по врстама добија преглед просечних растојања између типова аминокиселина. Тако се исти скуп података посматра на три нивоа детаљности.



Слика 13: Топлотна мапа за цео протеин 1ERU



Слика 14: Топлотна мапа протеина 1ERU за изабране аминокиселине



Слика 15: Просечна растојања по врстама аминокиселина за протеин 1ERU

## Глава 5

# Резултати и анализа

У овом поглављу приказани су резултати добијени применом развијене апликације. Најпре су дате глобалне статистике целокупне базе података, затим је описан поступак којим су из базе података издвојени протеини за детаљну анализу, након чега следе студије случаја четири одабрана протеина. Поглавље се завршава упоредним освртом на добијене резултате.

### 5.1 Глобалне статистике базе података

База података садржи податке о 933 протеинске структуре са укупно 150 685 чворова аминокиселина. Просечна дужина ланца износи 160,3 аминокиселине, док је медијана 137 аминокиселина. Најкраћи ланац у бази података садржи 30 аминокиселина, а најдужи 633 аминокиселине. У бази података чувају се сви парови резидуума једног протеина без примене прага. Укупан број веза за све протеине је 46 306 160, приближно 15,4 милиона за сваки од три типа растојања. Резидууми различитих ланаца нису повезани, због чега број веза расте квадратно у односу на дужину ланца.

Међу секундарним структурама најзаступљенији су  $\alpha$ -хеликси (32,3%), затим  $\beta$ -листови (24,4%), док нерегуларни сегменти означени симболом „-“ чине 19,4% свих резидуума. Преостале секундарне структуре су мање заступљене, при чему завоји (T) чине 11,1%, сегменти типа S 8,0%, а  $3_{10}$ -хеликси (G) 3,8%. Структуре типа B јављају се у 1,1% случајева, док структуре типа I нису присутне у анализираном скупу података. Најчешће аминокиселине у бази података су леуцин (LEU, 8,8%), аланин (ALA, 7,9%), валин (VAL, 7,1%), глицин (GLY, 7,0%) и глутаминска киселина (GLU, 6,9%).

Глобална анализа растојања показује разлике у просечним вредностима између посматраних типова растојања. Најмање просечне вредности добијене су за минимална растојања између тешких атома (*minbezh*), која у просеку износе 20,75 Å. Просечна вредност *Ca–Ca* растојања (*saca*) износи 23,79 Å, док су највеће вредности добијене за максимална растојања између тешких атома (*maxbezh*), са просеком од 27,49 Å. Стандардне девијације су веома сличне за сва три типа растојања и износе приближно 11 Å. Наведене вредности приказане су у табели 4.

Табела 4: Глобалне статистике растојања у бази података

| Тип            | Просек (Å) | Стд. девијација (Å) | Мин. (Å) | Макс. (Å) |
|----------------|------------|---------------------|----------|-----------|
| <i>saca</i>    | 23,79      | 11,11               | 2,68     | 128,00    |
| <i>minbezh</i> | 20,75      | 11,08               | 0,12     | 125,18    |
| <i>maxbezh</i> | 27,49      | 11,10               | 4,60     | 132,45    |

Показатељи наведени у овом одељку добијени су упитима над целом базом података. У наставку су описана два упита: први даје општи увид у садржај базе, а други издваја пет најзаступљенијих аминокиселина.

Први упит израчунава укупан број протеина и аминокиселина, као и просечну и медијанску дужину ланца. Аминокиселине се најпре групишу према протеину и ланцу, при чему се за сваку групу одређује њена дужина. Кључна реч *DISTINCT* користи се како би се сваки протеин рачунао само једанпут (листинг 7).

Листинг 7: Основне статистике базе података

```
MATCH (a:AminoAcid)
WITH a.protein AS p, a.chain AS c, count(a) AS len
RETURN count(DISTINCT p) AS proteins,
       sum(len) AS amino_acids,
       round(avg(len), 1) AS avg_len,
       round(percentileCont(len, 0.5), 1) AS median_len,
       min(len) AS min_len,
       max(len) AS max_len
```

Други упит, за проналажење пет најзаступљенијих аминокиселина у бази података, користи клаузулу *LIMIT* за ограничавање броја враћених резултата (листинг 8):

Листинг 8: Пет најзаступљенијих аминокиселина у бази података

```
MATCH (a:AminoAcid)
WITH a.name AS name, count(a) AS n
ORDER BY n DESC
LIMIT 5
RETURN name, n
```

## 5.2 Методологија избора протеина

Из базе података која обухвата 933 протеинске структуре било је потребно издвојити мали број репрезентативних примера за детаљну анализу и демонстрацију могућег коришћења апликације. Протеини су бирани тако да се међусобно разликују по обележјима која непосредно утичу на просторни распоред аминокиселина: по дужини ланца, по доминантној секундарној структури и по облику расподеле растојања. Примери нису одабрани унапред, већ је сваки од њих добијен упитом над самом базом података, тако да је развијена апликација послужила и као средство за одабир предмета анализе.

Број веза по протеину није коришћен као критеријум, иако би у општем случају могао да послужи као мера густине графа. Пошто база података садржи растојања између сваког пара аминокиселина, број веза за ланац од  $n$  резидуума увек износи  $\frac{n(n-1)}{2}$ , независно од просторне организације структуре, па је та величина само посредна мера дужине ланца.

### Дужина ланца

Први критеријум је број аминокиселина у протеину. Чворови се групишу по идентификатору протеина, а број резидуума добија се пребројавањем чворова у свакој групи.

Сортирањем у растућем поретку издвајају се најкраћи ланци у бази података, а изменом поретка у DESC и најдужи. На овај начин је одабран протеин **1G2Y**, који са свега 30 аминокиселина представља један од најкраћих ланаца у бази података и служи као пример структуре мале просторне распрострањености.

## Доминантна секундарна структура

Други критеријум је удео појединачног типа секундарне структуре у укупном броју резидуума. Резидууми са траженом *DSSP* ознаком најпре се пребројавају по протеину, а затим се тај број дели укупним бројем резидуума истог протеина. За  $\alpha$ -хеликсе користи се ознака Н.

Издвојен је протеин **1FEW**, код кога чак 93,1% резидуума припада типу Н, што га чини изразито хеликалним примером. Заменом ознаке Н ознаком Е исти упит даје протеине са доминантном  $\beta$ -структуром, међу којима је одабран протеин **1K3I** са израженим уделом резидуума типа Е.

## Асиметрија расподеле растојања

Трећи критеријум не почива на структурним обележјима, већ на облику расподеле измерених растојања. Уколико се средња вредност и медијана растојања у једном протеину значајно разликују, расподела је асиметрична, што указује на постојање мањег броја веома удаљених парова аминокиселина који померају аритметичку средину. Као мера је коришћена релативна разлика средње вредности и медијане, изражена у процентима.

Услов  $n > 10000$  искључује кратке ланце, код којих је број парова аминокиселина премали да би добијене статистике биле поуздане. Због тога се протеини попут 1G2Y, са свега неколико стотина парова, не појављују међу резултатима овог упита, па је критеријум мале дужине ланца морао бити примењен засебно. Први резултати описаног упита приказани су у табели 5.

Табела 5: Протеини са највећом релативном разликом средње вредности и медијане  $C\alpha-C\alpha$  растојања

| Протеин | Просек (Å) | Медијана (Å) | Разлика (%) |
|---------|------------|--------------|-------------|
| 1A17    | 29,98      | 24,55        | 18,13       |
| 1I2H    | 25,04      | 20,91        | 16,48       |
| 1WWB    | 24,07      | 20,43        | 15,12       |

Као пример протеина са изразитом асиметријом расподеле одабран је **1A17**, који се налази на врху листе са релативном разликом од 18,13%.

## Преглед одабраних примера

Примењени критеријуми и одабрани протеини сажети су у табели 6. Овако одабрани примери покривају међусобно различите структурне ситуације: најкраћи ланац, две супротне доминантне секундарне структуре и изразито асиметричну расподелу растојања.

Табела 6: Критеријуми примењени при избору протеина за студије случаја

| Протеин | Критеријум одабира                                                                         |
|---------|--------------------------------------------------------------------------------------------|
| 1G2Y    | најкраћи ланац у бази података (30 аминокиселина)                                          |
| 1K3I    | изражен удео резидуума типа Е (доминантна $\beta$ -структура)                              |
| 1FEW    | изразито доминантна $\alpha$ -хеликална структура (93,1% резидуума типа Н)                 |
| 1A17    | највећа релативна разлика средње вредности и медијане $C\alpha-C\alpha$ растојања (18,13%) |

## 5.3 Студије случаја

Свака студија случаја обухвата исти скуп анализа: састав аминокиселина, расподелу секундарних структура, статистике за сва три типа растојања, матрицу растојања и идентификацију просторно издвојених резидуума. Јединствен поступак примењен на све одабране протеине омогућава њихово међусобно поређење, које је дато у закључном делу поглавља.

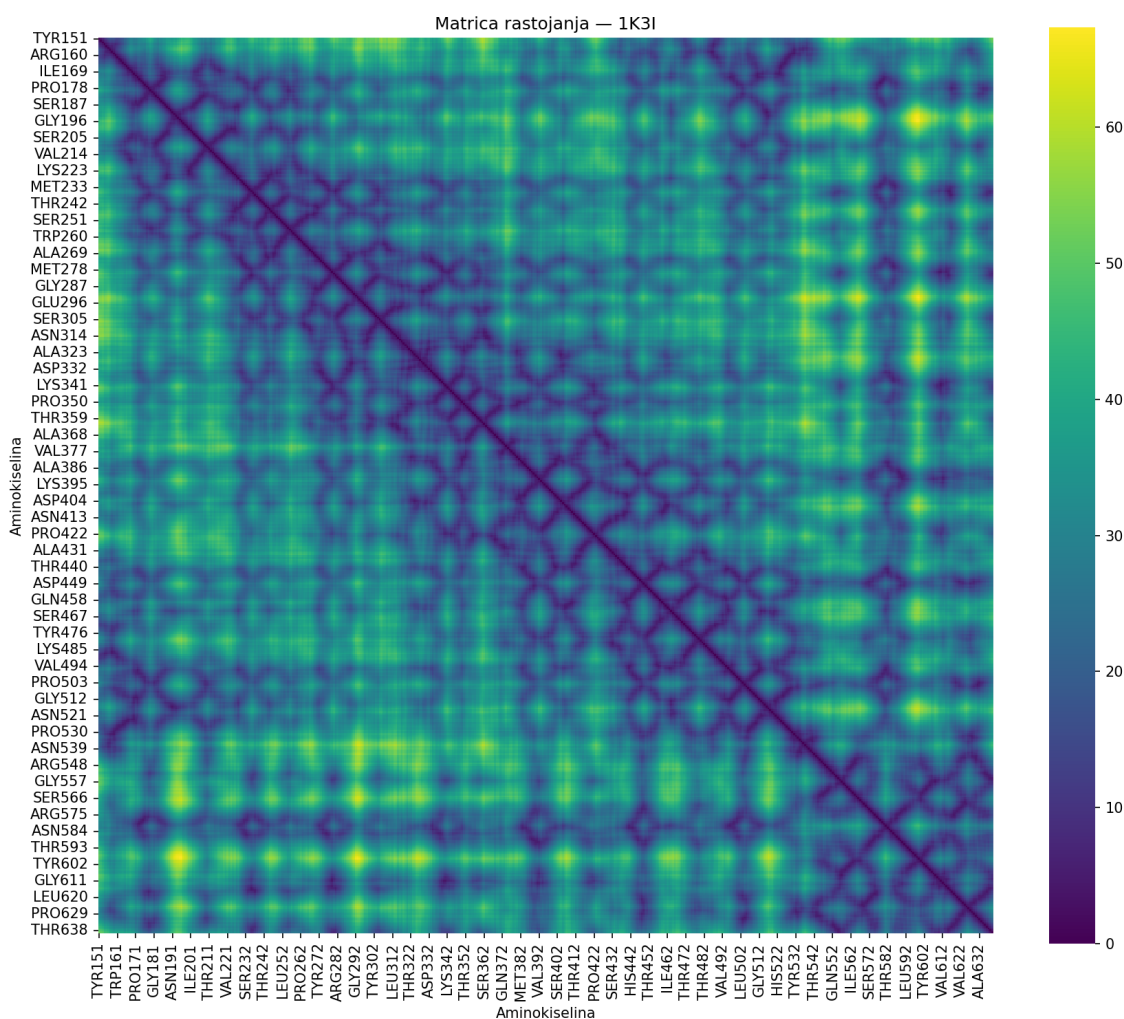
### Протеин 1K3I

Састав аминокиселина показује присуство свих двадесет стандардних типова аминокиселина. Најзаступљенији су глицин (GLY, 11,5%), серин (SER, 11,3%) и треонин (THR, 9,8%), док су најређе заступљени цистеин (CYS, 0,6%), хистидин (HIS, 1,4%) и глутаминска киселина (GLU, 1,4%).

Расподела секундарних структура указује на доминацију елемената са ознаком Е, који чине 38,7% свих резидуума. Поред њих значајно су заступљени региони без додељене секундарне структуре (-, 29,5%), као и завоји (Т, 13,7%) и кривине (S, 11,9%). У посматраном протеину нису идентифико-

вани  $\alpha$ -хеликси (H), док висок удео резидуума типа E потврђује доминацију  $\beta$ -структурних елемената.

На слици 16 приказана је матрица растојања за протеин 1K3I. Матрица показује очекивану симетрију у односу на главну дијагоналу, док се дуж дијагонале уочавају карактеристични блокови сличних вредности растојања. Блокови сличних вредности одговарају групама аминокиселина које су међусобно просторно ближе него остатку структуре и издвајају локално организоване регионе протеина.



Слика 16: Матрица  $C\alpha$ - $C\alpha$  растојања за протеин 1K3I

За тип растојања саса средња вредност износи  $27,63 \text{ \AA}$ , а медијана  $27,42 \text{ \AA}$ , што указује на релативно симетричну расподелу растојања. За растојања типа

$\text{minbezh}$  добијен је мањи просек од 24,73 Å, док је код растојања типа  $\text{maxbezh}$  просек 31,12 Å. Добијени резултати су у складу са дефиницијама коришћених метрика, будући да  $\text{minbezh}$  представља минимално растојање између тешких атома два резидуума, а  $\text{maxbezh}$  максимално.

Основне статистике за три типа растојања приказане су у табели 7.

Табела 7: Поређење статистика растојања за протеин 1K3I

| Метрика             | $\text{minbezh}$ | $\text{caca}$ | $\text{maxbezh}$ |
|---------------------|------------------|---------------|------------------|
| Просек (Å)          | 24,73            | 27,63         | 31,12            |
| Медијана (Å)        | 24,50            | 27,42         | 30,91            |
| Стд. девијација (Å) | 10,64            | 10,63         | 10,63            |
| Минимум (Å)         | 1,32             | 2,93          | 4,91             |
| Максимум (Å)        | 65,32            | 67,32         | 68,91            |

Бројност парова секундарних структура и припадајуће статистике растојања приказане су на слици 17. Најзаступљеније су интеракције типа - – Е и Е – Е, што одговара високом уделу резидуума са ознаком Е у овом протеину. За парове Е – Е забележена су и нижа просечна растојања у односу на већину осталих комбинација секундарних структура.

Као најизолованији резидууми идентификовани су GLY599, GLY598 и ASN597. Исти или веома сличан скуп резидуума појављује се међу најизолованијим за сва три типа растојања. То указује да идентификовани издвојени регион није последица избора конкретне метрике растојања, већ представља стабилну структурну карактеристику посматраног протеина.

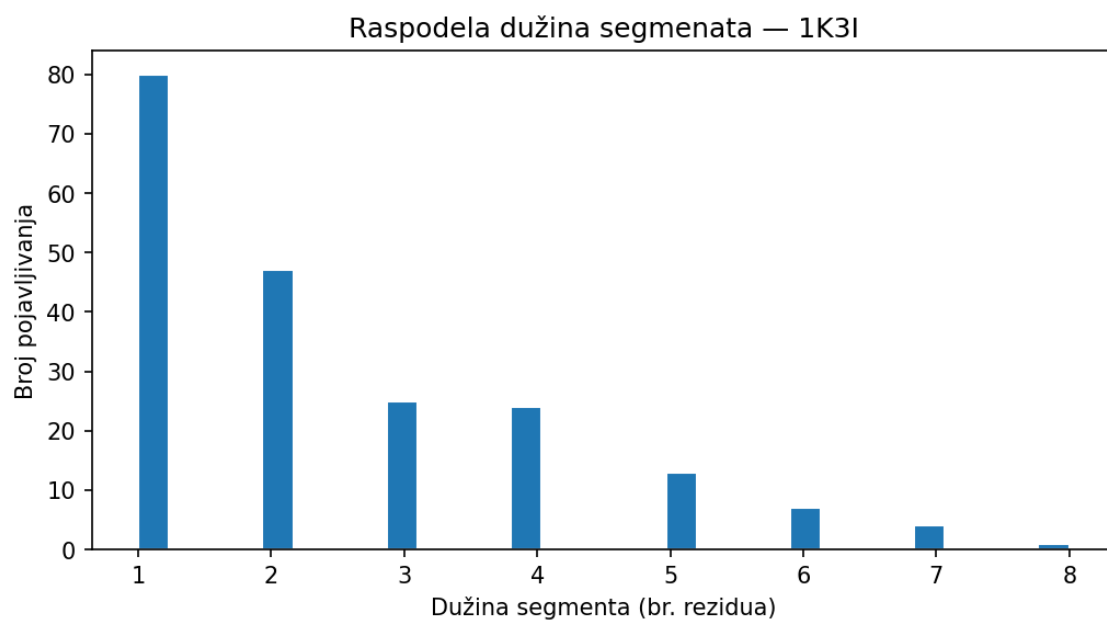
Расподела дужина сегмената секундарне структуре приказана је на слици 18, при чему се дужине сегмената посматрају без разликовања типова секундарне структуре. Најбројнији су сегменти дужине један и два резидуума, а број појављивања опада са порастом дужине, при чему ниједан сегмент не прелази осам резидуума. Одсуство дугих сегмената последица је тога што у протеину нема  $\alpha$ -хеликса, будући да хеликални региони могу обухватити знатно већи број узастопних резидуума. Расподела дужина сегмената по типу секундарне структуре приказана је на слици 19.

Добијени резултати показују да је применом развијене апликације могуће истовремено анализирати просторне односе између аминокиселина, структуру контаката и карактеристике појединачних региона протеина. Протеин 1K3I представља пример структуре код које су доминантни  $\beta$ -структурни елемен-

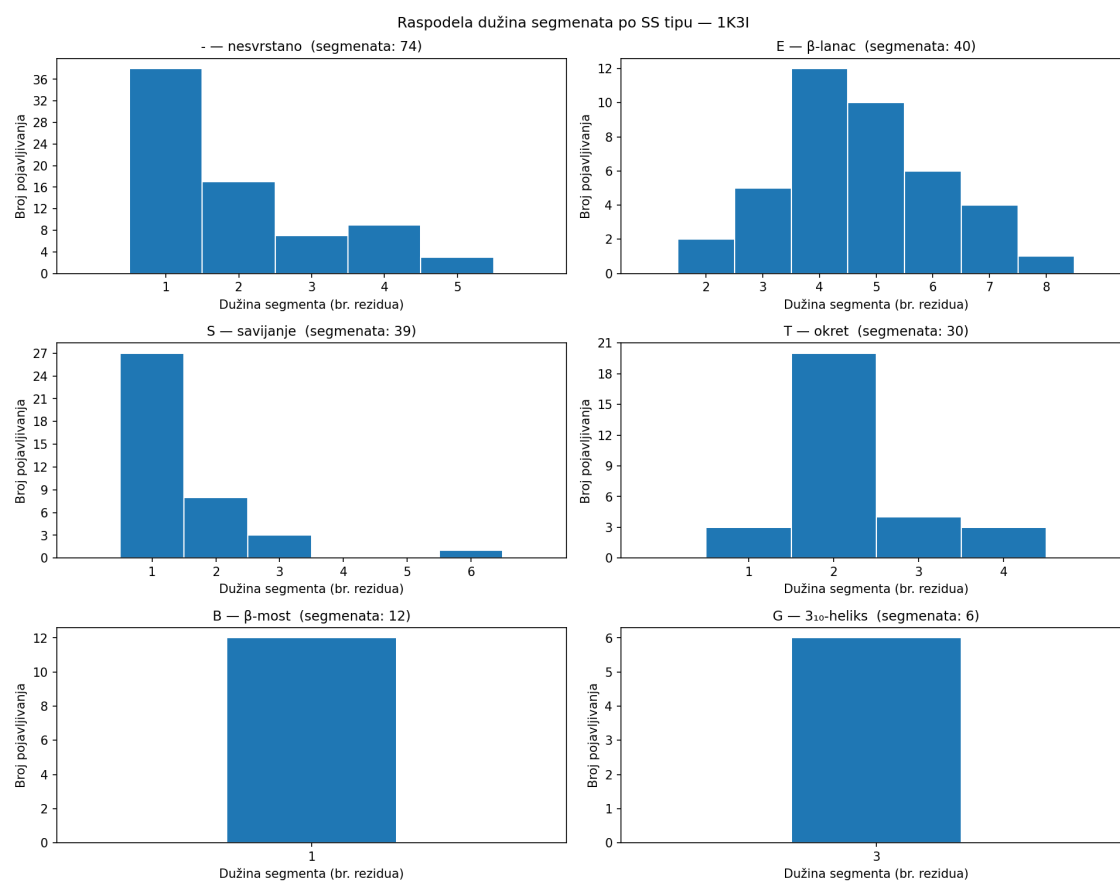
ти праћени јасно уочљивим просторним обрасцима у матрици растојања и стабилно идентификованим изолованим регионима.

| Statistike SS parova                           |       |            |              |         |           |       |
|------------------------------------------------|-------|------------|--------------|---------|-----------|-------|
| Ceo protein: granice, k i SS se ne primenjuju. |       |            |              |         |           |       |
| Tip: saca · Parova: 21                         |       |            |              |         |           |       |
| SS par                                         | Broj  | Prosek (Å) | Medijana (Å) | Min (Å) | Maks. (Å) | Std   |
| - E                                            | 27216 | 27.38      | 27.16        | 3.75    | 67.28     | 10.26 |
| E E                                            | 17766 | 24.71      | 24.84        | 2.93    | 55.29     | 9.92  |
| E T                                            | 12663 | 27.78      | 27.05        | 3.79    | 66.76     | 10.41 |
| E S                                            | 10962 | 28.24      | 27.84        | 3.79    | 67.32     | 10.48 |
| - -                                            | 10296 | 27.66      | 27.72        | 3.58    | 61.37     | 10.71 |
| - T                                            | 9648  | 29.52      | 29.55        | 3.77    | 66.65     | 10.99 |
| - S                                            | 8352  | 28.33      | 28.19        | 3.00    | 62.26     | 11.31 |
| S T                                            | 3886  | 30.08      | 29.84        | 3.69    | 66.72     | 11.50 |
| E G                                            | 3402  | 27.61      | 27.32        | 3.80    | 60.43     | 9.59  |
| - G                                            | 2592  | 29.24      | 29.34        | 3.79    | 55.70     | 10.32 |
| B E                                            | 2268  | 27.57      | 26.72        | 4.26    | 54.72     | 9.47  |
| T T                                            | 2211  | 30.29      | 30.16        | 3.79    | 65.94     | 11.78 |
| - B                                            | 1728  | 26.09      | 26.27        | 3.79    | 55.45     | 10.57 |
| S S                                            | 1653  | 29.08      | 28.51        | 3.63    | 63.07     | 11.81 |
| G T                                            | 1206  | 29.62      | 30.17        | 3.79    | 61.45     | 11.25 |
| G S                                            | 1044  | 29.56      | 29.79        | 3.82    | 57.06     | 10.84 |
| B T                                            | 804   | 29.55      | 30.34        | 3.80    | 55.39     | 10.37 |
| B S                                            | 696   | 27.51      | 28.21        | 3.79    | 57.97     | 11.60 |
| B G                                            | 216   | 29.25      | 28.23        | 3.83    | 50.43     | 9.95  |
| G G                                            | 153   | 28.90      | 32.85        | 3.79    | 45.65     | 12.44 |
| B B                                            | 66    | 21.86      | 25.06        | 4.27    | 39.29     | 11.62 |

Слика 17: Бројност и статистике растојања по паровима секундарних структура за протеин 1K3I (тип саса)



Слика 18: Расподела дужина сегмената секундарне структуре за протеин 1K3I



Слика 19: Расподела дужина сегмената по типу секундарне структуре за протеин 1K3I

## Протеин 1A17

Протеин 1A17 садржи 159 аминокиселина, при чему је присутно 19 од 20 стандардних типова аминокиселина. Најзаступљеније су аминокиселине аланин (ALA, 13,2%), лизин (LYS, 11,3%), тирозин (TYR, 8,8%) и глутаминска киселина (GLU, 8,2%), док триптофан (TRP) није присутан у посматраној структури. Анализа секундарне структуре, приказана на слици 20, показује изражену доминацију  $\alpha$ -хеликса, који обухватају 76,7% свих резидуума, док су преостале аминокиселине распоређене у завојима, неправилним сегментима и краћим структурним елементима. Секундарна структура овог протеина је, дакле, изразито хеликална.

| Raspodela sekundarne strukture    |      |       |
|-----------------------------------|------|-------|
| Ukupno rezidua sa SS oznakom: 159 |      |       |
| SS                                | Broj | %     |
| H                                 | 122  | 76.7% |
| -                                 | 20   | 12.6% |
| T                                 | 14   | 8.8%  |
| S                                 | 3    | 1.9%  |

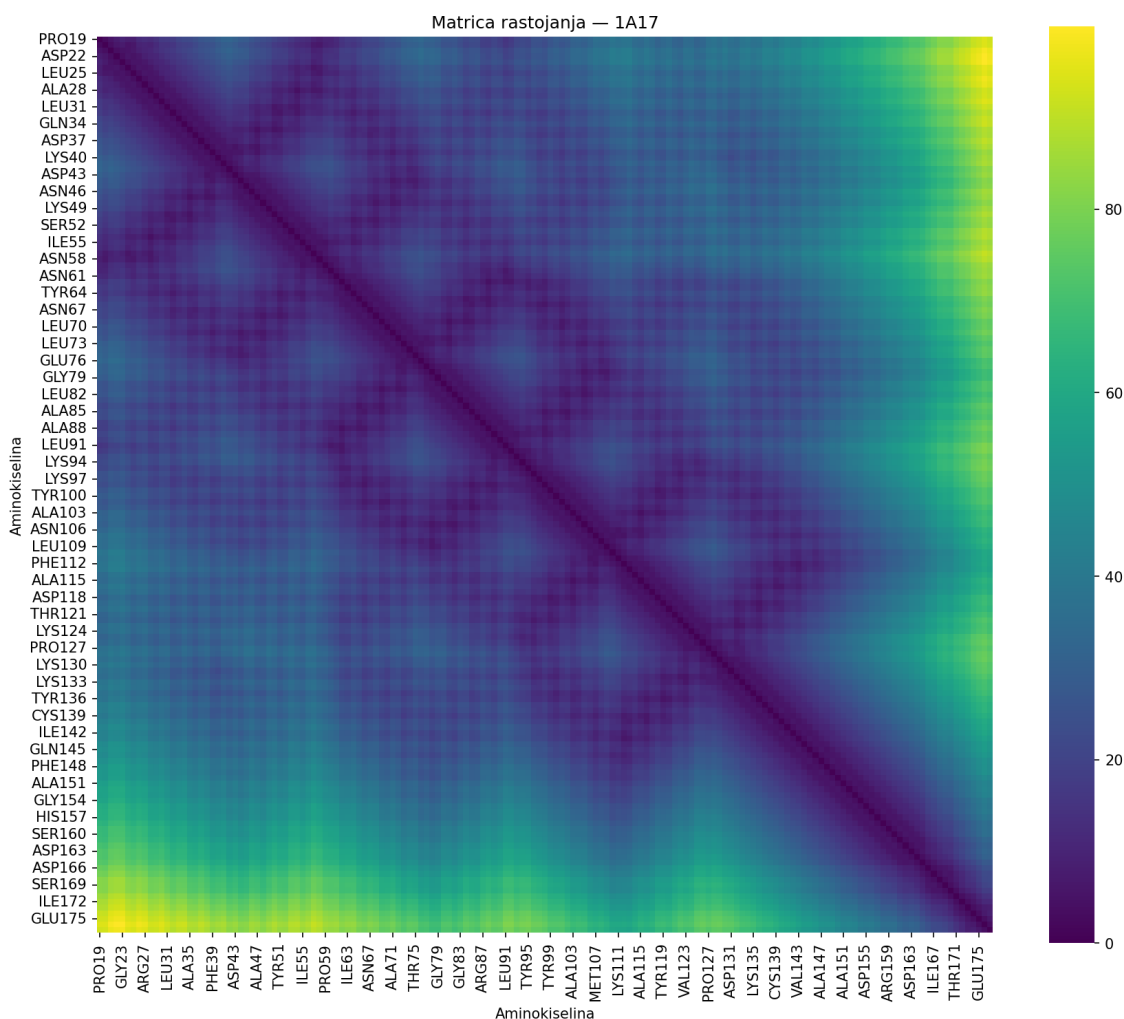
Слика 20: Расподела резидуума по типовима секундарне структуре за протеин 1A17

Основне статистике за три типа растојања приказане су у табели 8. За сва три типа растојања уочава се значајна разлика између средње вредности и медијане. Код  $C\alpha$ – $C\alpha$  растојања средња вредност износи 29,98 Å, док је медијана 24,55 Å. Сличан однос присутан је и код минималних и максималних растојања. Разлика између средње вредности и медијане одговара асиметричној расподели у којој мањи број веома великих растојања повећава аритметичку средину. Додатно, максималне вредности достижу приближно 100 Å, што је знатно више него код већине анализираних протеина.

Табела 8: Основне статистике растојања за протеин 1A17

| Тип растојања | Просек (Å) | Медијана (Å) | Максимум (Å) |
|---------------|------------|--------------|--------------|
| саса          | 29,98      | 24,55        | 99,93        |
| minbezh       | 26,84      | 21,44        | 97,25        |
| maxbezh       | 34,03      | 28,63        | 104,99       |

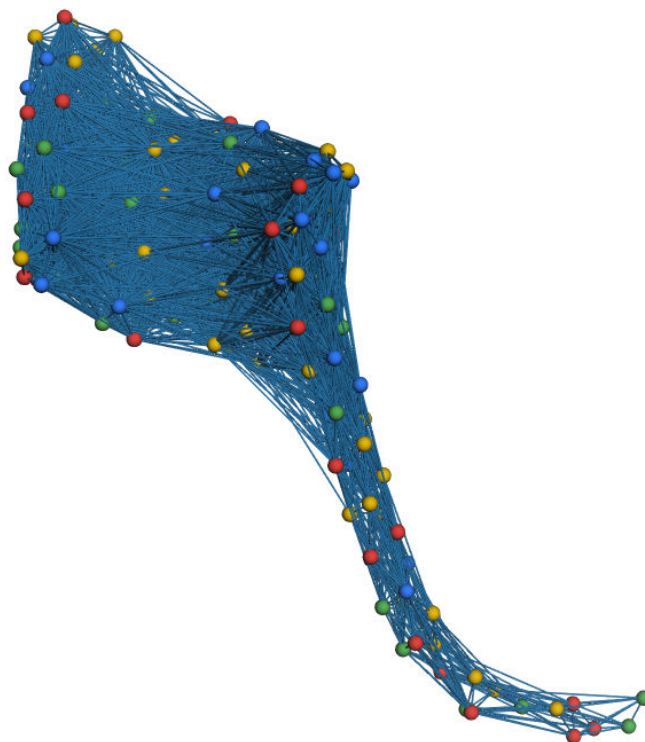
На слици 21 приказана је матрица растојања за тип саса. За разлику од протеина 1K3I, код кога су уочљиви компактнији блокови мањих растојања, код протеина 1A17 приметан је постепени пораст вредности растојања ка крају секвенце. Светлије области у доњем десном делу матрице указују на присуство парова аминокиселина који су просторно веома удаљени. Овај образац је у складу са великом разликом између средње вредности и медијане и указује на постојање издвојеног дела структуре.



Слика 21: Матрица  $C\alpha$ – $C\alpha$  растојања за протеин 1A17

Тродимензионални приказ графа, дат на слици 22, додатно потврђује ову појаву. На визуализацији се уочавају компактно језгро и издужен завршни део ланца који се од њега пружа у простору, због чега су резидууми на ње-

говом крају удаљени од већине осталих. Анализа најизолованијих резидуума показује да се већина њих налази у области индекса 168–177, што указује да завршни део секвенце формира структурни регион који је удаљен од остатка протеина.

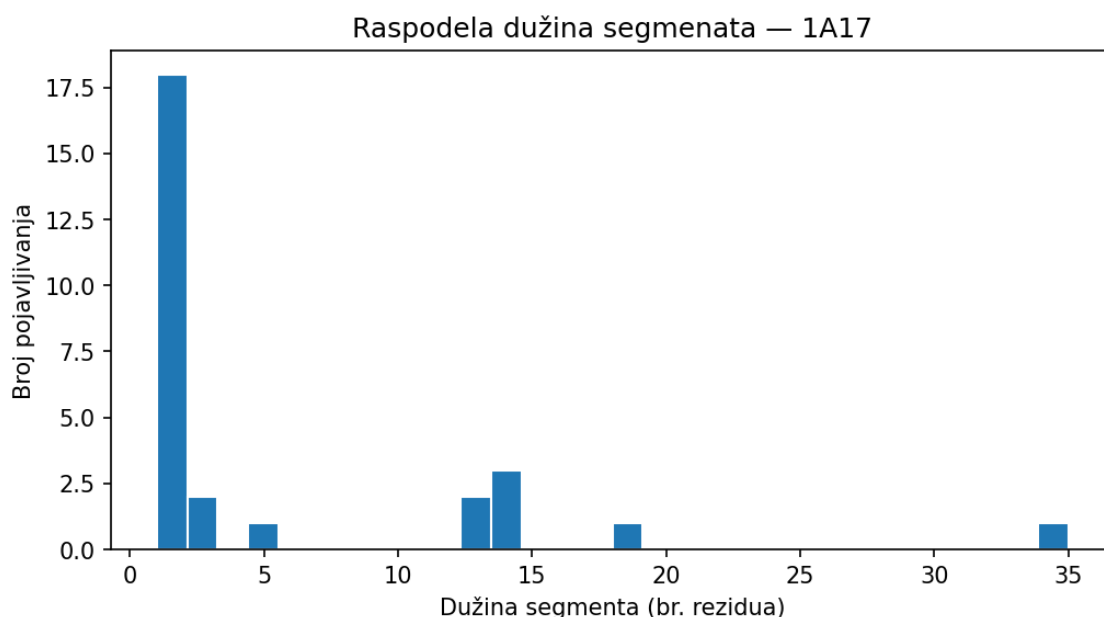


Слика 22: Тродимензионални приказ графа за протеин 1A17

Анализа растојања између различитих типова секундарних структура показује да су најмање просечне вредности добијене за парове хеликс--хеликс (H--H), док су највеће вредности карактеристичне за парове који укључују аминокиселине ван регуларних секундарних структура. Ово је очекиван резултат, јер доминантни хеликални сегменти формирају компактнији део структуре, док су неправилни и терминални делови протеина просторније распршени.

Расподела дужина сегмената секундарне структуре, приказана на слици 23, показује две јасно раздвојене групе. Прву чини велики број врло кратких сегмената, дужине један до пет резидуума, а другу мањи број знатно

дужих сегмената, међу којима се издвајају они од приближно тринаест и четрнаест резидуума, као и појединачни сегменти дужине око деветнаест и тридесет пет резидуума. Оваква подела одговара структури састављеној од неколико дугих хеликалних региона повезаних кратким неправилним сегментима. Дуги хеликални сегменти обухватају већи део ланца у простору и доприносе појави великих растојања која повећавају аритметичку средину у односу на медијану.



Слика 23: Расподела дужина сегмената секундарне структуре за протеин 1A17

Код протеина 1A17 уочена је изражена асиметрија расподеле растојања упркос доминантно хеликалној структури. Статистички показатељи, матрица растојања и тродимензионални приказ заједно издвајају регион који је просторно удаљен од остатка структуре и значајно утиче на глобалне вредности растојања.

## Протеин 1G2Y

Протеин 1G2Y представља најмањи протеин међу анализираним примерима и садржи свега 30 аминокиселина. У структури је присутно 11 од 20 стандардних аминокиселина, док аргинин (ARG), аспарагин (ASN), аспарагинска киселина (ASP), цистеин (CYS), хистидин (HIS), фенилаланин (PHE),

пролин (PRO), триптофан (TRP) и тирозин (TYR) нису заступљени. Најчешће аминокиселине су леуцин (LEU, 26,7%), серин (SER, 13,3%), аланин (ALA, 13,3%), глутамин (GLN, 10,0%) и глутаминска киселина (GLU, 10,0%).

Анализа секундарне структуре показује доминацију  $\alpha$ -хеликса који обухватају 73,3% свих резидуума. Преостале аминокиселине припадају нерегуларним сегментима и завојима. За разлику од протеина 1K3I, код кога доминирају  $\beta$ -листови, и протеина 1A17 који садржи већи број хеликалних сегмената различитих дужина, структура протеина 1G2Y је релативно једноставна и компактна.

Основне статистике растојања приказане су у табели 9. За сва три типа растојања средња вредност и медијана су веома блиске, што указује на приближно симетричну расподелу без изражених екстремних вредности. Код C $\alpha$ -C $\alpha$  растојања средња вредност износи 13,50 Å, док је медијана 13,27 Å. Максимално измерено растојање износи 27,28 Å, што је знатно мање у односу на протеине 1K3I и 1A17.

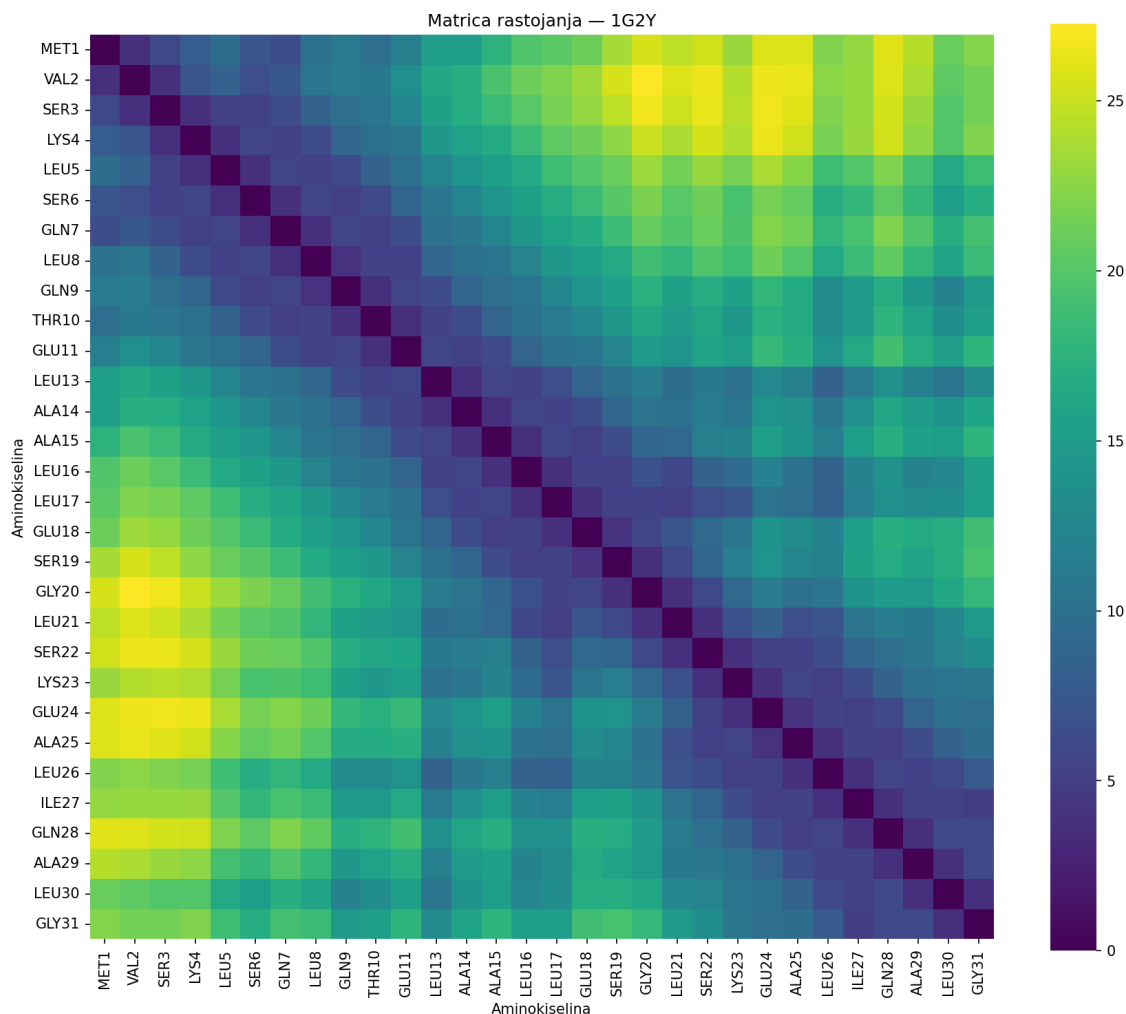
Табела 9: Основне статистике растојања за протеин 1G2Y

| Тип растојања | Просек (Å) | Медијана (Å) | Максимум (Å) |
|---------------|------------|--------------|--------------|
| саса          | 13,50      | 13,27        | 27,28        |
| minbezh       | 10,66      | 10,15        | 25,31        |
| maxbezh       | 17,42      | 16,66        | 36,68        |

Матрица растојања приказана на слици 24 има правилну структуру са јасно израженом дијагоналом и без великих издвојених региона. За разлику од протеина 1A17, код кога су уочене области веома великих растојања, овде се вредности постепено мењају дуж секвенце. Равномерна промена вредности и одсуство издвојених области одговарају компактној просторној организацији без јасно одвојених удаљених региона.

Анализа парова секундарних структура показује да најмања просечна растојања имају парови хеликс – хеликс (H – H), што је очекивано с обзиром на доминацију хеликалних сегмената. Слично томе, анализа најизолованијих резидуума не указује на постојање изразито издвојених делова структуре. Највећа просечна растојања имају резидууми VAL2, MET1 и SER3, али су добијене вредности значајно мање од одговарајућих вредности код протеина 1A17.

Расподела дужина сегмената секундарне структуре приказана је на сли-

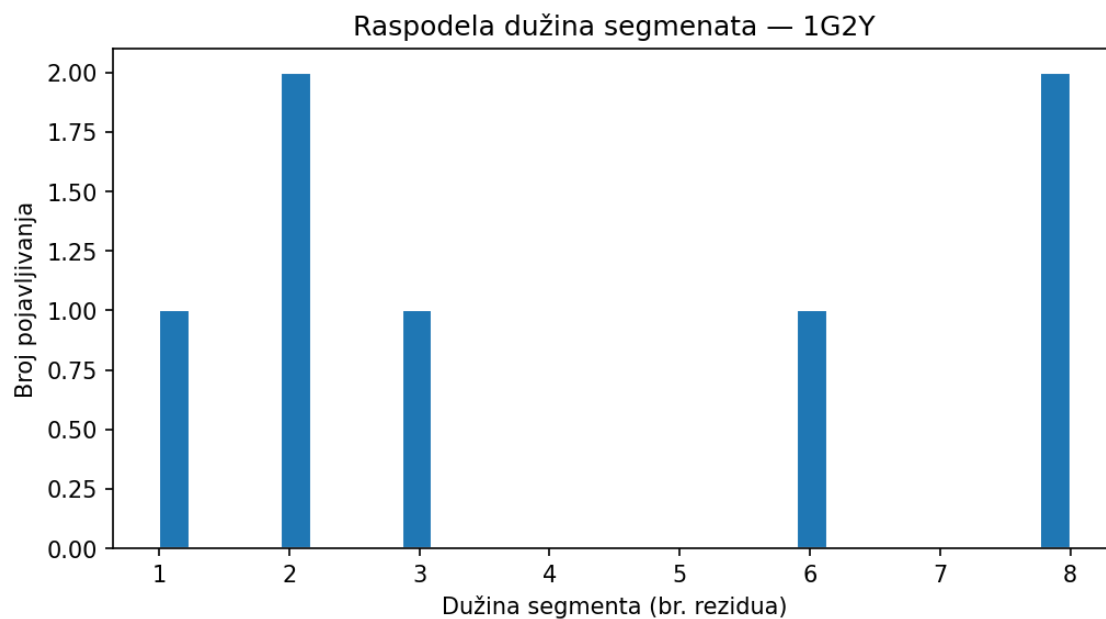


Слика 24: Матрица  $C\alpha$ – $C\alpha$  растојања за протеин 1G2Y

ци 25. Целокупан ланац од 30 аминокиселина разложен је на свега седам сегмената, од којих су најдужи два сегмента од по осам резидуума, праћени сегментом од шест резидуума, док су преостали сегменти дужине један до три резидуума. Мали број сегмената одговара једноставној структурној организацији кратког ланца, у којој не постоји довољно материјала за формирање просторно раздвојених региона.

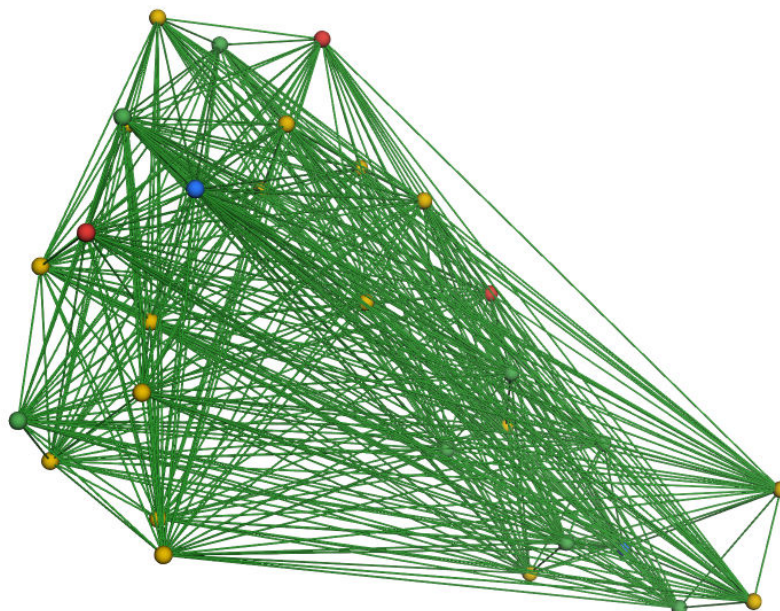
Тродимензионални приказ графа, дат на слици 26, показује збијену структуру у којој су резидууми густо повезани, без просторно издвојених делова какви су уочени код већих протеина.

Резултати за 1G2Y разликују се од претходних примера: кратка секвен-



Слика 25: Расподела дужина сегмената секундарне структуре за протеин 1G2Y

ца праћена је малим максималним растојањима и приближно симетричном расподелом вредности. У структури се зато не издвајају просторно удаљени региони попут оних уочених код протеина 1A17.



Слика 26: Тродимензионални приказ графа за протеин 1G2Y

## Протеин 1FEW

Протеин 1FEW садржи 173 аминокиселине и представља пример протеина у коме доминира  $\alpha$ -хеликална секундарна структура. Од укупно 173 резидуума, чак 161 припада типу H, што чини 93,1% свих аминокиселина. Поред хеликалних сегмената присутан је мали број завоја (T), као и неколико резидуума без дефинисане секундарне структуре. У аминокиселинском саставу најзаступљеније су глутаминска киселина (GLU), аланин (ALA), леуцин (LEU) и треонин (THR), док цистеин (CYS) и пролин (PRO) нису присутни.

Основне статистике растојања приказане су у табели 10. За растојања типа саса добијена је средња вредност од 32,51 Å и медијана од 28,46 Å. Сличан однос примећује се и код типова *minbezh* и *maxbezh*, где је средња вредност осетно већа од медијане. Ово указује на асиметричну расподелу растојања и присуство мањег броја веома удаљених парова аминокиселина који повећавају

аритметичку средину.

Табела 10: Основне статистике растојања за протеин 1FEW

| Тип растојања | Просек (Å) | Медијана (Å) | Максимум (Å) |
|---------------|------------|--------------|--------------|
| саса          | 32,51      | 28,46        | 102,50       |
| minbezh       | 28,98      | 24,74        | 99,34        |
| maxbezh       | 36,56      | 32,52        | 106,85       |

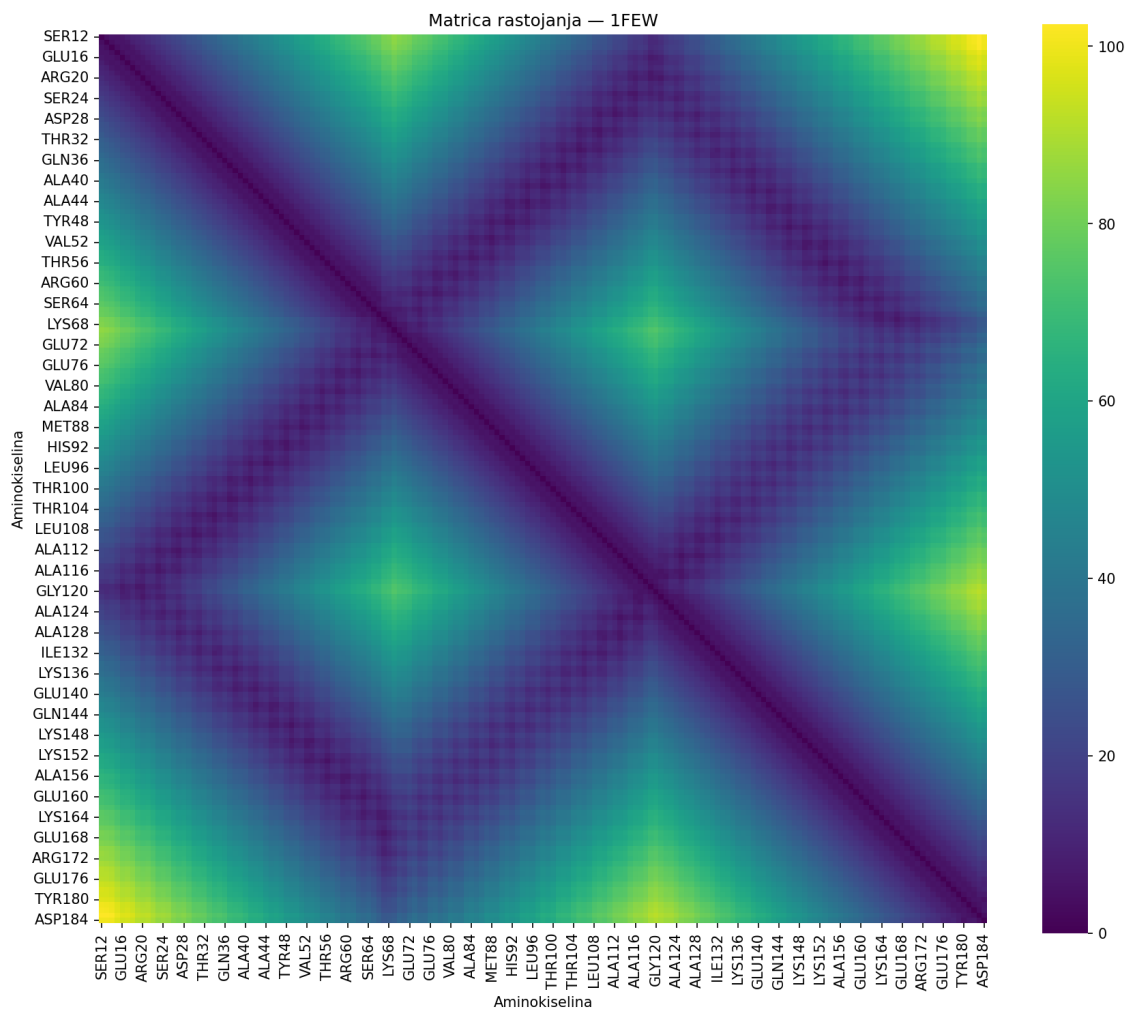
На слици 27 приказана је матрица растојања за тип *саса*. Уочава се изражена симетрија око главне дијагонале, као и постепено повећање растојања између аминокиселина које су удаљене у секвенци. Највеће вредности јављају се у угловима матрице и одговарају интеракцијама између резидуума са супротних крајева протеинског ланца. За разлику од протеина 1K3I, код кога се уочавају сложени блокови и локални обрасци повезани са  $\beta$ -листовима, код протеина 1FEW доминира гладак и правилан образац карактеристичан за издужене  $\alpha$ -хеликалне структуре.

Због изразите доминације  $\alpha$ -хеликса, највећи број посматраних парова припада типу Н – Н. Просечно растојање за ове парове износи 30,79 Å за *саса* растојања, док су интеракције које укључују резидууме ван дефинисане секундарне структуре у просеку знатно удаљеније. Посебно велике вредности уочене су код парова типа --Н и --Т, што упућује на положај таквих резидуума у периферним деловима структуре. Ознаке прате *DSSP* конвенцију уведену у одељку 2.3: Н означава  $\alpha$ -хеликс, Т завој, а - резидуум без додељене секундарне структуре.

Најизолованије аминокиселине налазе се на крају секвенце. Највеће просечно растојање има резидуум GLU183, након кога следе ASP184 и ARG182. Ове аминокиселине достижу просечна растојања већа од 50 Å за тип *maxbezh*, што указује на њихов положај у структурно издвојеним сегментима протеина.

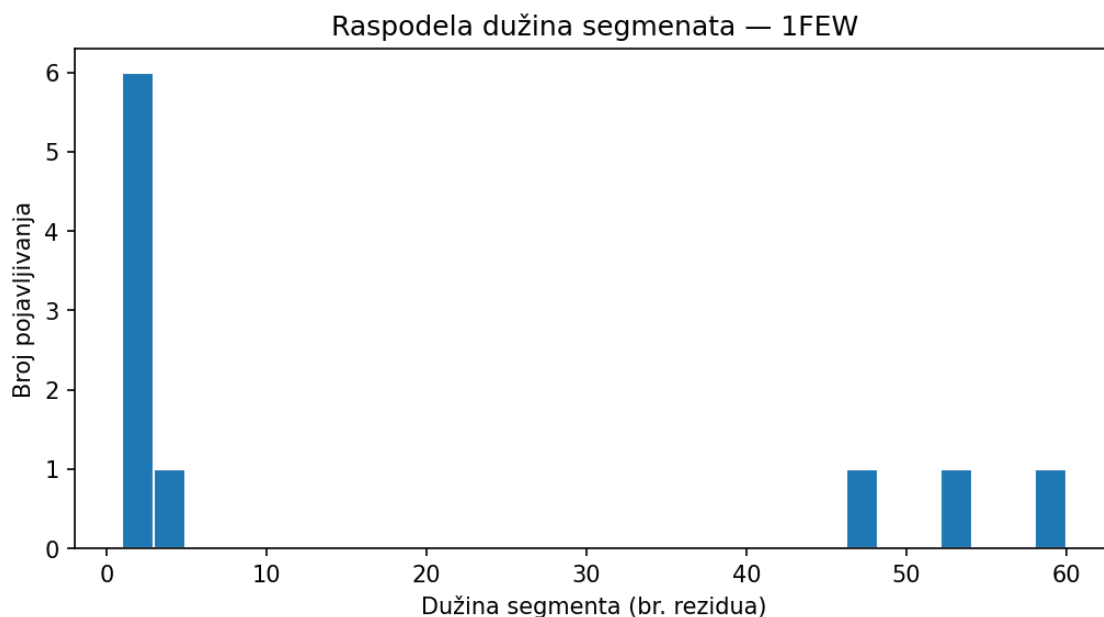
Расподела дужина сегмената секундарне структуре приказана на слици 28 додатно потврђује доминацију дугих  $\alpha$ -хеликалних региона. Уочава се неколико веома дугих хеликалних сегмената дужине приближно 50–60 резидуума, што објашњава велики број међусобно удаљених парова аминокиселина и повећане вредности максималних растојања.

Протеин 1FEW показује да висока заступљеност  $\alpha$ -хеликса не подразумева нужно малу просторну распрострањеност структуре. Дуги хеликални сегменти у овом случају доводе до великих растојања између појединих делова



Слика 27: Матрица  $C\alpha-C\alpha$  растојања за протеин 1FEW

ланца, што се одражава и на глобалне статистике растојања.



Слика 28: Расподела дужина сегмената секундарне структуре за протеин 1FEW

## 5.4 Дискусија

Појединачне студије случаја описале су четири протеина одабрана по различитим критеријумима. Постављени заједно, њихови резултати омогућавају поређење које превазилази појединачне примере. Основни показатељи сажети су у табели 11.

Табела 11: Упоредни преглед анализираних протеина (растојања типа *саса*); релативна разлика представља разлику између средње вредности и медијане растојања, изражену у процентима средње вредности

| Протеин | Резидууми | Доминантна <i>SS</i> | Просек (Å) | Медијана (Å) | Релативна разлика средње вредности и медијане (%) | Најдужи сегмент |
|---------|-----------|----------------------|------------|--------------|---------------------------------------------------|-----------------|
| 1A17    | 159       | H 76,7%              | 29,98      | 24,55        | 18,1                                              | 35              |
| 1FEW    | 173       | H 93,1%              | 32,51      | 28,46        | 12,5                                              | 60              |
| 1G2Y    | 30        | H 73,3%              | 13,50      | 13,27        | 1,7                                               | 8               |
| 1K3I    | 488       | E 38,7%, H 0%        | 27,63      | 27,42        | 0,8                                               | 8               |

Најизразитији образац јесте веза између дужине сегмената секундарне структуре и облика расподеле растојања. Протеини код којих најдужи сегмент обухвата 35, односно 60 резидуума (1A17 и 1FEW) показују релативну

разлику средње вредности и медијане од 18,1% и 12,5%, док протеини чији најдужи сегмент не прелази осам резидуума (1G2Y и 1K3I) имају разлику мању од 2%.

Објашњење следи из просторне природе дугих сегмената. Хеликални сегмент од педесетак резидуума протеже се кроз знатан део структуре, због чега резидууми са његових крајева нужно улазе у парове са великим међусобним растојањем. Такви парови су малобројни у односу на укупан број парова, па слабо померају медијану, али подижу аритметичку средину. Код структура грађених од кратких сегмената таквих парова нема, па су средња вредност и медијана готово изједначене.

Однос, међутим, није строго монотон. Протеин 1FEW има знатно дуже сегменте од протеина 1A17 (60 према 35 резидуума), али мању асиметрију (12,5% према 18,1%). Дужина најдужега сегмента стога није довољна као једина мера; код протеина 1A17 асиметрији додатно доприноси и просторно издвојен регион у области индекса 168–177, који је уочен и на матрици растојања и у тродимензионалном приказу на слици 22.

Други образац тиче се односа између дужине ланца и просторне распрострањености структуре: број аминокиселина не одређује непосредно колико је структура распрострањена. Протеин 1K3I са 488 резидуума готово је три пута дужи од протеина 1FEW са 173 резидуума, али му је просечно  $C_{\alpha}$ – $C_{\alpha}$  растојање мање (27,63 наспрам 32,51 Å), а највеће измерено растојање знатно мање (67,32 наспрам 102,50 Å). Дуг ланац изграђен од кратких  $\beta$ -структурних сегмената пакује се збијено, док се краћи ланац изграђен од дугих хеликалних сегмената у простору знатно више разилази.

Овај налаз потврђује запажање изнето уз протеин 1FEW: висока заступљеност  $\alpha$ -хеликса не подразумева компактну структуру. Штавише, поређење показује и да сам удео секундарне структуре слабо описује просторни распоред. Протеини 1A17 и 1G2Y имају готово једнак удео  $\alpha$ -хеликса (76,7% и 73,3%), а њихова највећа измерена растојања разликују се више од три и по пута (99,93 наспрам 27,28 Å). Тек посматрани заједно, удео секундарне структуре, дужина сегмената и дужина ланца дају потпунију слику о структури.

Анализа најизолованијих резидуума показала је две правилности. Прво, издвојени резидууми не јављају се појединачно, већ у узастопним низовима: код протеина 1K3I то су резидууми ASN597, GLY598 и GLY599. Област издвојених резидуума код протеина 1A17 обухвата индексе 168–177, док се код

протеина 1FEW као издвојени јављају резидууми ARG182, GLU183 и ASP184. Овакав налаз у складу је са геометријом полипептидног ланца, будући да су просторни положаји резидуума суседних у секвенци ограничени геометријом полипептидне окоснице и пептидне везе [1]. Због тога се може очекивати да просторна издвојеност једног резидуума буде праћена и издвојеношћу његових непосредних суседа у секвенци, што може објаснити зашто се као издвојен јавља сегмент ланца, а не само појединачан резидуум.

Друго, код протеина 1K3I исти или готово исти скуп резидуума добијен је за сва три типа растојања. Пошто *casac*, *minbezh* и *maxbezh* мере различите величине, поклапање резултата указује да издвојеност није последица избора метрике, већ својство саме структуре. Код протеина 1G2Y такве регије нису уочене, што одговара кратком ланцу у коме нема довољно резидуума да би се формирали просторно одвојени делови.

Што се самих типова растојања тиче, у свим анализираним протеинима задржан је исти поредак просечних вредности:  $\text{minbezh} < \text{casac} < \text{maxbezh}$ , што је очекивано с обзиром на дефиниције ових величина. Разлике међу њима релативно су уједначене и износе неколико ангстрема, док су стандардне девијације готово изједначене (приближно 11 Å на нивоу целе базе података). Из тога следи да сва три типа растојања носе сличну слику о глобалној организацији структуре. Избор типа растојања постаје значајан тек када се уводи праг контакта, јер исти праг примењен на различите типове издваја различите скупове парова.

Развијено решење може се упоредити са постојећим апликацијама сличне намене, као што су *DistAA* [11], база растојања између аминокиселина са веб апликацијом за њихов статистички преглед, и *RING* [5], веб сервер који за задату структуру гради мрежу интеракција између резидуума. Сва три решења полазе од анализе односа између резидуума у тродимензионалној структури протеина, али се разликују по томе који се односи чувају и како се касније тумаче. *DistAA* је првенствено усмерен на статистичку анализу растојања између аминокиселина, док *RING* издваја парове резидуума који испуњавају услове за конкретне физичко-хемијске интеракције. У овом раду задржавају се подаци о растојањима за све расположиве парове резидуума, а критеријуми анализе, као што су тип и праг растојања, задају се приликом постављања упита. На тај начин исти протеин може да се посматра по различитим критеријумима, без поновног освежавања података, а последице промене критеријума

јума могу се пратити и кроз статистичке показатеље и кроз графовски приказ структуре. Са *DistAA* решење дели употребу истих типова растојања и статистички приступ, док му је са *RING*-ом заједничко представљање протеина у облику графа резидуума. Ипак, веза у овом моделу не означава потврђену интеракцију, већ растојање између резидуума, због чега се добијени резултати не могу тумачити као директан доказ хемијске или физичке интеракције.

Добијене закључке треба посматрати уз неколико ограничења. Детаљно су анализирана четири протеина, што је мали узорак у односу на 933 структуре колико их база података садржи; уочени обрасци стога представљају запажања над одабраним примерима, а не статистички проверене тврдње о целој бази података. Даље, као што је већ напоменуто, *SCOP* класа резидуума заступљена је у моделу података, али није искоришћена ни у једној од спроведених анализа, а пошто је база јавно доступна, та могућност остаје отворена за даља истраживања. Најзад, мера издвојености резидуума заснована је на просечном растојању до осталих резидуума и представља релативан показатељ, будући да праг контакта није примењен.

Из добијених резултата произлази и неколико праваца даљег рада. Уочени однос између дужине сегмената и асиметрије расподеле растојања погодан је за проверу на целој бази података. Упитом који за сваки протеин израчунава дужину најдужег сегмента и релативну разлику средње вредности и медијане могуће је испитати да ли образац уочен на четири примера важи и на 933 структуре, као и да ли постоји праг дужине сегмента изнад кога асиметрија постаје изразита.

Скуп статистика могуће је проширити. Пошто су подаци већ представљени графом, природну допуну чине графовске мере попут степена чвора, централности и коефицијента груписања, којима би се идентификовале аминокиселине са посебним положајем у мрежи контаката. Уз примену прага контакта, добијене мере могле би се упоредити са постојећим приступима заснованим на матрицама контаката.

Посебан правац представља укључивање доменског знања које база података већ садржи или му се може придружити. *SCOP* класа омогућава поређење просечних растојања и образаца паковања између структурних класа протеина, чиме би се проверило да ли се уочене разлике између  $\alpha$ - и  $\beta$ -структура задржавају на нивоу класа. Придруживањем функционалних података, као што су активна и везивна места, било би могуће испитати да ли се такви реги-

они издвајају и по просторним показатељима који су у овом раду коришћени.

У базу података могле би се увести и структуре које нису обухваћене скупом података *DistAA*, под условом да су доступне координате свих атома и ознаке секундарне структуре резидуума, а сва три типа растојања је могуће израчунати на основу координата.

Апликацију се може проширити истовременим приказом више протеина ради поређења, извозом добијених резултата у различите понуђене формате погодне за даљу обраду, као и памћењем задатих комбинација филтера, чиме би се анализа могла лакше поновити.

## Глава 6

### Закључак

Растојања између аминокиселина најчешће се сажимају у матрицу контаката применом унапред одабраног прага.

У раду је најпре формирана графовска база података у систему *Neo4j*, у коју су пренета растојања између аминокиселина, где су аминокиселине чворови, а растојања везе са придруженим типом и вредношћу. Уз сваки резидуум сачувани су и ознака секундарне структуре и *SCOP* класа, тако да структурни контекст остаје доступан упиту. Над тако формираном базом података развијена је веб апликација која омогућава интерактивно истраживање тих података. Праг контакта тако не мора да буде одлука донета пре анализе, већ параметар који се мења током ње, заједно са типом растојања, изабраним аминокиселинама, типом секундарне структуре, најмањим растојањем у секвенци и ланцем.

Свака таква измена одражава се истовремено на два приказа исте структуре: на тродимензионални, у коме положаји сфера одговарају координатама  $C_\alpha$  атома, а везе растојањима која испуњавају задате услове, и на статистички, који обухвата састав аминокиселина, расподелу секундарних структура и дужина њихових непрекинутих сегмената, топлотну мапу растојања, статистике по комбинацијама секундарних структура, просечна растојања и просторно издвојене резидууме. Анализа стога није низ засебних израчунавања над унапред припремљеним матрицама, него рад над истим подацима у коме се критеријуми мењају, а статистике и граф добијају упоредо. Поновно израчунавање захтевају само они показатељи који зависе од измењеног критеријума.

Анализа при томе није везана за један ниво посматрања. Уз показатеље

целе базе података, израчунате при покретању система, стоје показатељи изабраног протеина, а избором појединачне аминокиселине у тродимензионалном приказу добијају се њена својства и суседи са којима образује приказана растојања, чиме се са нивоа базе података прелази на ниво једног резидуума без промене података над којима се ради. Топлотне мапе, хистограми дужина сегмената и остали изведени прикази чувају се као слике и преузимају за могућу даљу употребу.

Протеини за студије случаја издвојени су упитима над самом базом података, према дужини ланца, доминантној секундарној структури и асиметрији расподеле растојања. Студије случаја послужиле су као примери употребе апликације. У њима је показано да асиметрија расподеле растојања прати присуство дугих сегмената секундарне структуре, да дужина ланца не одређује просторну распрострањеност структуре и да просторно издвојени резидууми чине узастопне низове у периферним деловима структуре, при чему се исти скуп добија за сва три типа растојања. Реч је о запажањима над четири детаљно анализирана примера, а не о статистички провереним тврдњама о целој бази података. Њихова провера на свим структурама наведена је међу правцима даљег рада.

Преношењем података о растојањима у графовску базу података омогућено је њихово непосредно анализирање упитима. Тип и гранична вредност растојања задају се у самом упиту, без потребе за прављењем нове матрице за сваку анализу. На тај начин иста база служи за различите анализе просторне организације протеина.

# Литература

- [1] Bruce Alberts et al. “The Shape and Structure of Proteins.” In: *Molecular Biology of the Cell*. 4th ed. Garland Science, 2002. Chap. 3. URL: <https://www.ncbi.nlm.nih.gov/books/NBK26830/>.
- [2] Maciej Besta et al. “Demystifying Graph Databases: Analysis and Taxonomy of Data Organization, System Designs, and Graph Queries.” In: *ACM Computing Surveys* 56.2 (2023), pp. 1–40. DOI: 10.1145/3604932.
- [3] N. V. Bhagavan and Chung-Eun Ha. *Essentials of Medical Biochemistry: With Clinical Cases*. 2nd ed. Academic Press, 2015. ISBN: 978-0-12-416687-5.
- [4] Carl Boettiger. “An introduction to Docker for reproducible research.” In: *ACM SIGOPS Operating Systems Review* 49.1 (2015), pp. 71–79. DOI: 10.1145/2723872.2723882.
- [5] Alessio Del Conte et al. “RING 4.0: faster residue interaction networks with novel interaction types across over 35,000 different chemical structures.” In: *Nucleic Acids Research* 52.W1 (2024), W306–W312. DOI: 10.1093/nar/gkae337.
- [6] Django Software Foundation. *Django Documentation*. Version 5.2. 2026. URL: <https://docs.djangoproject.com/en/5.2/> (visited on March 2026).
- [7] Docker, Inc. *Docker Compose Documentation*. 2026. URL: <https://docs.docker.com/compose/> (visited on March 2026).
- [8] José Guia, Valéria Gonçalves Soares, and Jorge Bernardino. “Graph Databases: Neo4j Analysis.” In: *Proceedings of the 19th International Conference on Enterprise Information Systems (ICEIS 2017)*. Vol. 1. SCITEPRESS, 2017, pp. 351–356. ISBN: 978-989-758-247-9. DOI: 10.5220/0006356003510356.

- [9] Wolfgang Kabsch and Christian Sander. “Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features.” In: *Biopolymers* 22.12 (1983), pp. 2577–2637. DOI: 10.1002/bip.360221211.
- [10] Koes Lab. *3Dmol.js Documentation*. 2026. URL: <https://3dmol.org/doc/index.html> (visited on March 2026).
- [11] Mirjana M. Maljković. “DistAA: Database of amino acid distances in proteins and web application for statistical review of distances.” In: *Computational Biology and Chemistry* 83 (2019), p. 107130. DOI: 10.1016/j.compbiolchem.2019.107130.
- [12] Antonio Melé. *Django 4 By Example: Build powerful and reliable Python web applications from scratch*. 4th ed. Packt Publishing, 2022. ISBN: 978-1-80181-305-1.
- [13] Alexey G. Murzin et al. “SCOP: a structural classification of proteins database for the investigation of sequences and structures.” In: *Journal of Molecular Biology* 247.4 (1995), pp. 536–540. DOI: 10.1006/jmbi.1995.0159.
- [14] Anupam Nath Jha, Saraswathi Vishveshwara, and Jayanth R. Banavar. “Amino acid interaction preferences in proteins.” In: *Protein Science* 19.3 (2010), pp. 603–616. DOI: 10.1002/pro.339.
- [15] Neo4j, Inc. *Neo4j Documentation*. 2026. URL: <https://neo4j.com/docs/> (visited on August 2026).
- [16] Ashwin Ragupathi, Alexander J. Mastrogiannis, and Nader Rahimi. “Biochemistry, Tertiary Protein Structure.” In: *StatPearls*. updated 1 December 2025. StatPearls Publishing, 2025. URL: <https://www.ncbi.nlm.nih.gov/books/NBK470269/>.
- [17] Nicholas Rego and David Koes. “3Dmol.js: molecular visualization with WebGL.” In: *Bioinformatics* 31.8 (2015), pp. 1322–1324. DOI: 10.1093/bioinformatics/btu829.
- [18] Ibraheem Rehman, Mustafa Farooq, and Salome Botelho. “Biochemistry, Secondary Protein Structure.” In: *StatPearls*. updated 11 December 2022. StatPearls Publishing, 2022. URL: <https://www.ncbi.nlm.nih.gov/books/NBK470235/>.

- [19] Ian Robinson, Jim Webber, and Emil Eifrem. *Graph Databases*. 1st ed. O'Reilly Media, 2013. ISBN: 978-1-449-35626-2.
- [20] Chad Vicknair et al. "A comparison of a graph database and a relational database: a data provenance perspective." In: *Proceedings of the 48th Annual ACM Southeast Regional Conference (ACM SE '10)*. ACM, 2010, pp. 1–6. DOI: 10.1145/1900008.1900067.
- [21] Саша Малков. *Пројектовање база података*. Материјали са курса, Универзитет у Београду, Математички факултет. 2023–2024.

# Употреба алата заснованих на вештачкој интелигенцији

У изради рада коришћена су два алата заснована на великим језичким моделима: *ChatGPT* и *Claude*. У делу који се односи на програмску имплементацију оба су служила за разматрање могућих имплементационих решења, писање и измену појединих делова кода и за откривање узрока грешака. *Claude* је коришћен и при обликовању текста, за преформулисање и уједначавање стила појединих делова.

Ниједан предлог добијен на овај начин није преузет без провере. Исправност кода потврђивана је покретањем и поређењем добијених резултата са очекиваним, а бројчане вредности наведене у раду проверене су упитима над самом базом података. Тврдње изнете у теоријском делу преузете су из извора наведених у литератури.

# Биографија аутора

Емилија Стевановић рођена је 28. октобра 2000. године у Врању. Завршила је природно-математички смер гимназије у Врању као носилац Вукове дипломе. Основне студије завршила је 2024. године, након чега је уписала мастер студије на Математичком факултету Универзитета у Београду, смер Информатика.

Интересовања ауторке усмерена су ка раду са подацима. Радно искуство у индустрији стекла је на пословима из области науке о подацима и као инжењер машинског учења.