Универзитет у Београду
Математички факултет

Магдалена Мићић

Предикција секундарне структуре протеина на основу статистичких података о заступљености аминокиселина у секундарној структури

мастер рад

Београд, 2018.

Ментор: проф. др Саша Малков
Математички факултет, Универзитет у Београду

Чланови комисије: проф. др Миодраг Живковић
Математички факултет, Универзитет у Београду

доц. др Јована Ковачевић
Математички факултет, Универзитет у Београду

Датум одбране: _____

Садржај

1 Увод

Протеини спадају у најзаступљеније органске макромолекуле у природи. Поред тога, они су и најзначајнији органски макромолекули, јер учествују у готово свим битним функцијама у ораганизму. Нпр. неки протеини обезбеђују структуру која даје ћелијама целовитост и облик. Други служе као хормони који преносе сигнал од једне ћелије до друге. Пример је панкреас који лучи хормон инсулин, који сигнализира јетри и мишићним ћелијама да односе шећер глукозу из крви. Такође, протеини могу да везују и носе различите супстанце. Хемоглобин преноси кисеоник из плућа до удаљених делова тела, док миоглобин складишти кисеоник у мишићним ткивима. Протеини, такође, контролишу активности гена. Неки служе као ензими који убрзавају хемијске реакције неопходне за живот. Према томе, различити протеини дају ћелијама различите улоге [18].

Протеини су изграђени од ланаца аминокиселина, повезаних пептидним везама. Број, врста и редослед аминокиселина у овим ланцима представља примарну структуру протеина. Осим примарне, постоје још три нивоа структуре протеина. Секундарна структура се односи на тродимензионални облик локалних сегмената полипептидног ланца, док терцијарна структура описује уређење елемената секундарне структуре у простору. Кватернарна структура је резултат спајања више полипептидних ланаца у један функционалан протеин.

Како су протеини укључени у све најважније процесе у организму, познавање њихове улоге и својстава је од огромног значаја, пре свега у медицини и фармацији, приликом креирања лекова и њихове примене, али и у другим областима, као што је производња хране, козметике, кућне хемије итд.

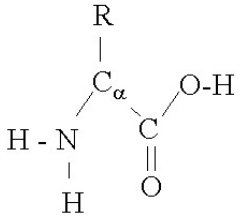
Како би се одредила функција протеина, није довољно само познавање секвенце аминокиселина које га граде. Ове секвенце могу да се, у зависности од разних фактора, увијају у различите облике у простору, што може довести до различитих улога. Због тога је потребно предвидети више облике структуре протеина, јер нам оне дају више информација о његовој функцији.

Циљ овог рада је да се, коришћењем корелација као оцене структуре, предвиди секундарна структура протеина на основу дате примарне структуре. Све предикције се врше узимајући у обзир односе које аминокиселине имају са различитим класама. У раду ће, осим описа алгоритама који су коришћени, бити дат и упоредни приказ резултата које остварују, као и додатне статистике које би могле бити од значаја за даља истраживања.

Сва израчунавања и обраде података, као и приказ резултата, реализовани су коришћењем програмског језика *Python* и помоћних библиотека отвореног кода.

1.1 Аминокиселине

Аминокиселине су једињења која садрже амино групу ($-NH_2$), карбоксилну групу ($-COOH$) и бочни ланац (R-групу, радикал или R-остатак). Док су амино и карбоксилна група заједничке за све аминокиселине, R-група је специфична за сваку. Иако у природи постоји преко 500 аминокиселина, само њих 20 улази у састав генетичког кода и обично се мисли управо на њих када се говори о аминокиселинама. Ове аминокиселине се још називају и алфа-аминокиселине (α-аминокиселине), јер су код њих амино, карбоксилна и R-група везане за α-атом угљеника (C_α). Хемијска структура аминокиселина представљена је на слици 1.1. Осим 20 "стандардних" аминокиселина које учествују у изградњи протеина, постоје и 2 "нестандардне" аминокиселине: селеноцистеин и пиролизин. Списак α-аминокиселина дат је у табели 1.1. Поред сваке аминокиселине, налазе се њена трословна, као и једнословна ознака.



Слика 1.1: Хемијска структура аминокиселина

Постоји 9 аминокиселина које се не могу синтетисати у људском организму, а неопходне су за његово нормално функционисање. То су *есенцијалне аминокиселине* и оне се морају унети путем хране. Ту спадају: фенилаланин, валин, треонин, триптофан, метионин, леуцин, изолеуцин, лизин и хистидин. Осим њих, 6 аминокиселина се сматрају за условно есенцијалне, јер се могу синтетисати у људском организму под одређеним условима, као што су нпр. неки катаболички поремећаји или превремено рођење код деце. У условно есенцијалне спадају: аргинин, цистеин, глицин, глутамин, пролин и тирозин. Преосталих 5 аминокиселина (аланин, аспарагинска киселина, аспарагин, глутаминска киселина и серин), могу се синтетисати у организму у довољним количинама.

1.2 Пептидна веза

Пептидна веза је најважнија хемијска веза која се остварује између две аминокиселине. У њој учествује амино група једне аминокиселине и карбоксилна група друге аминокиселине. У овој реакцији, атом угљеника из карбоксилне групе, везује се за атом азота из амино групе, при чему се ослобађа молекул воде (слика 1.2). Повезивањем аминокиселина на овај начин, добијају се пептидни ланци, који представљају основу за изградњу протеина. Спајањем неколико аминокиселина пептидним везама, добијају се *олигопептиди*, док се спајањем великог броја аминокиселина добијају *полипептиди*. Полипептидни ланац поседује својство поларности - на једном крају је α-амино група, а на другом α-карбоксилна група. По конвенцији, крај на коме се налази амино група, сматра се почетком полипептидног ланца [1].

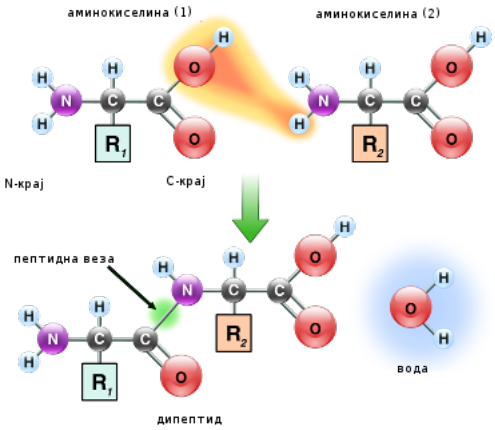
Табела 1.1: Стандардне аминокиселине и њихове ознаке

Аминокиселина	Трословна ознака	Једнословна ознака
аланин	Ala	A
аргинин	Arg	R
аспарагин	Asn	N
аспарагинска киселина	Asp	D
цистеин	Cys	C
фенилаланин	Phe	F
глицин	Gly	G
глутамин	Gln	Q
глутаминска киселина	Glu	E
хистидин	His	H
изолеуцин	Ile	I
леуцин	Leu	L
лизин	Lys	K
метионин	Met	M
пролин	Pro	P
серин	Ser	S
тирозин	Tyr	Y
триптофан	Trp	W
треонин	Thr	T
валин	Val	V

1.3 Протеини

Уобичајени полипептидни ланци садрже између 50 и 2000 аминокиселина и називају се *протеинима* [1]. Протеини су најзначајнији органски макромолекули. Њихов назив потиче од грчке речи *proteios*, што значи први, најважнији, главни. Неке од главних улога које имају у организму су:

- ензимска - ензими су катализатори који убрзавају хемијске реакције. Пример ензима је фосфофруктокиназа.
- регулаторна - регулишу рад других протеина у обављању својих физиолошких функција. Један од најпознатијих регулаторних протеина је инсулин.
- транспортна - преносе одређене супстанце од једног до другог места. Нпр. хемоглобин преноси кисеоник од плућа до ткива.
- складишна - обезбеђују резервоар неопходних нутријената за организам. Казеин је најзаступљенији протеин у млеку и главни извор азота код одојчади.



Слика 1.2: Изградња пептидне везе

- контрактилна - омогућавају покретљивост ћелија, нпр. код ћелијске деобе и контракције мишића. Актин и миозин су примери ове врсте протеина.
- структурна - стварају и одржавају биолошке структуре. Они обезбеђују издржљивост и заштиту ћелијама и ткивима. Нпр. кератин се налази у коси и ноктима, док колаген изграђује кости, зубе, хрскавицу, лигаменте.
- адаптивна - имају улогу у ћелијским одговорима на хормоне и факторе раста. Они имају модуларну организацију, где модули препознају и везују одређене структуралне елементе у друге протеине, кроз интеракцију између два протеина.
- одбрамбена - учествују у заштити ћелија или експлоатацији. Значајну улогу у заштити имају антитела, која препознају и неутралишу "стране" молекуле, који су резултат напада бактерија и вируса на организам. Протеини попут тромбина и фибриногена учествују у процесу згрушавања крви [10].

1.4 Структура протеина

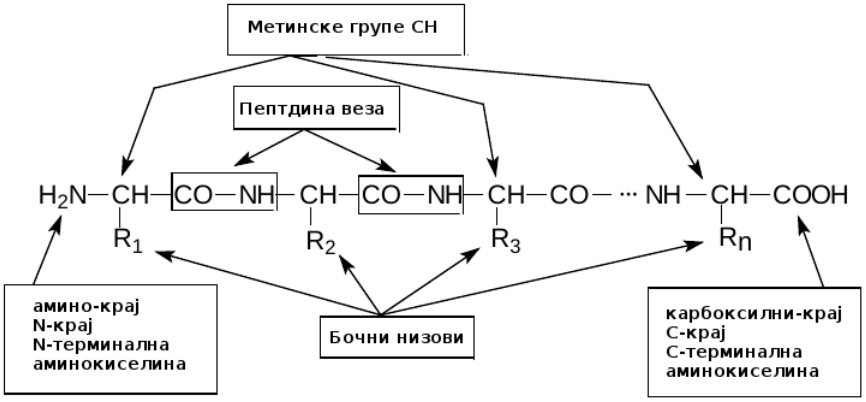
Разликујемо четири нивоа структуре протеина:

- примарна структура
- секундарна структура
- терцијарна структура
- кватернарна структура

Нивои су хијерархијски уређени, тако да сваки представља сложенији облик структуре у односу на претходни.

1.4.1 Примарна структура

Редослед везивања остатака аминокиселина у полипептидном ланцу представља примарну структуру протеина. Основни низ код полипептидног ланца је исти за све молекуле протеина и сачињен је од наизменично поређаних метинских ($-CH-$) и пептидних (амидних) група [19]. Грађа основног низа приказана је на слици 1.3.



Слика 1.3: Грађа основног низа полипептида

Крај основног низа, на коме се налази слободна амино група, обично се назива N-крај, док се супротни крај низа, са слободном карбоксилном групом, назива C-крај. Као што је већ споменуто, по конвенцији, приликом писања примарне структуре, увек се почиње од N-краја.

При грађењу виших нивоа структуре протеина, бочни низови остатака аминокиселина у полипептидном ланцу имају одлучујућу улогу. При подели аминокиселина на основу улоге у стереохемији[1] протеина, узима се у обзир поларност бочних низова, присуство функционалних група у њима, као и способност тих група у погледу електролитичке дисоцијације[2] [19].

1.4.2 Секундарна структура

Секундарна структура протеина представља тродимензионални облик делова протеина, настао увијањем полипептидног ланца. Секундарна структура је у највећој мери одређена водоничним везама између остатака аминокиселина са другим аминокиселинама, које се налазе релативно близу у полипептидном ланцу. Два најчешћа облика секундарне струкутре су α-хеликс (α-завојница) и β-наборана структура (β-плоча).

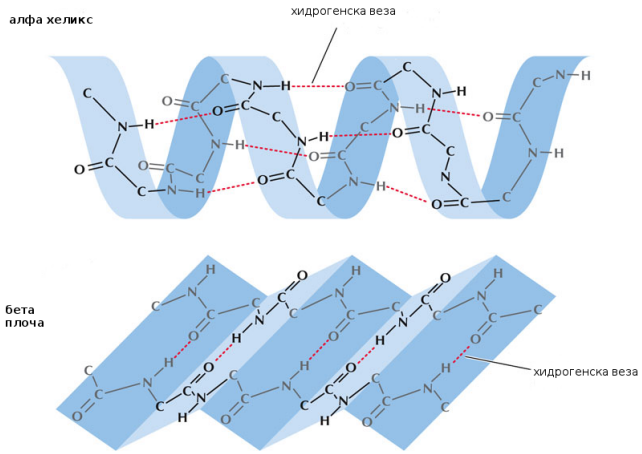
Код α-хеликса, водоничне везе се образују између кисеониковог атома из карбоксилне групе ($C = O$) сваког првог остатка аминокиселине у низу и атома водоника из амино групе ($NH-$) сваког петог остатка. Водоничне везе се протежу паралелно са осом α-хеликса, држећи основни полипептидни ланац у увијеном облику, а бочни ланци су увек усмерени од хеликса ка спољашњости [19, 12]. Ду-

[1] грана хемије која проучава просторну организацију молекула

[2] разлагање електролита на позитивне или негативне јоне, под утицајем молекула растварача

жине α-хеликса варирају од протеина до протеина, али у просеку, приближно имамо 10 остатака по сегменту.

Полипептидни ланац, назван β-плоча, у β-набораним структурама, готово је потпуно издужен, за разлику од α-хеликса, који је уско увијен. Бочни ланци суседних аминокиселина су супротно усмерени. Једна β-набрана структура се формира повезивањем две или више β-плоча водоничним везама. Суседни ланци у овој структури могу да се простиру у истом смеру (паралелне β-структуре) или у супротном (антипаралелне β-структуре). Код антипаралелног уређења, NH група и CO група сваке аминокиселине су, респективно, повезане водоничним веезама са CO групом и NH групом партнерског суседног ланца. У паралелном уређењу, везе су мало компликованије. За сваку аминокиселину, NH група је повезана са CO групом једне аминокиселине на суседној плочи, док је CO група повезана са NH групом аминокиселине која је два остатка даље низ ланац [1]. На слици 1.4 приказан је изглед секундарних структура α-хеликса и β-плоче.



Слика 1.4: Секундарне структуре α-хеликса и β-плоче

Према класификацији *DSSP* (*Define Secondary Structure of Proteins*[3]), постоји 8 типова секундарних структура:

- H - α-хеликс (4-хеликс); минимална дужина је 4 резидуала
- G - 3_{10}-хеликс (3-хеликс); минимална дужина је 3 резидуала
- I - π-хеликс (5-хеликс); минимална дужина је 5 резидуала
- E - продужена трака у паралелним и/или антипаралелним β-плочама; минимална дужина је 2 резидуала
- B - остатак у изолованом β-мосту (формација водоничне везе од једног пара β-структура)
- T - завој, кривина премошћена водоничном везом; 3, 4 или 5 кривина

[3]назив потиче од Pascal програма који су Wolfgang Kabsch и Chris Sander 1983. године имплементирали с циљем стандардизације класификовања секундарних структура

- S - оштрији завој, кривина; једина структура која није заснована на водоничним везама
- C - бланко, празнина, сви остали резидуали[7].

1.4.3 Терцијарна структура

Терцијарна структура описује глобалну конформацију протеина, односно, начин на који су елементи његове секундарне структуре уређени у простору. Терцијарна структура је одређена интеракцијама између аминокиселина, које могу бити веома удаљене у полипептидном низу, али се, захваљујући увијању протеина, нађу близу једна другој [2]. Полипептидни ланци који имају сличне секвенце аминокиселина, углавном дају и сличне тродимензионалне структуре. Испитивање ових структура показало је да су хиљаде познатих протеина изграђене од релативно малог број различитих структуралних мотива [3].

У стабилизацији терцијарне структуре протеина, поред водоничне везе, битну улогу имају и други видови интеракције бочних низова аминокиселина, као што су електростатичке и хидрофобне интеракције, као и успостављање ковалентних веза (нпр. дисулфидне везе) [19].

1.4.4 Кватернарна структура

Кватернарна структура настаје спајањем више полипептидних ланаца, како би се формирао један функционалан протеин [2]. Сваки ланац у оваквом протеину представља једну подјединицу, па је кватернарна структура одређена просторним уређењем ових подјединица и природом њихових интеракција [1]. За разлику од терцијарне структуре, где смо имали интеракције између бочних низова остатака аминокиселина које припадају једном полипептидном ланцу, на овом нивоу структуре, те интеракције се одвијају између бочних ланаца остатака аминокиселина које припадају различитим полипептидним ланцима [19]. На слици 1.5, приказан је однос сва четири нивоа структуре протеина.

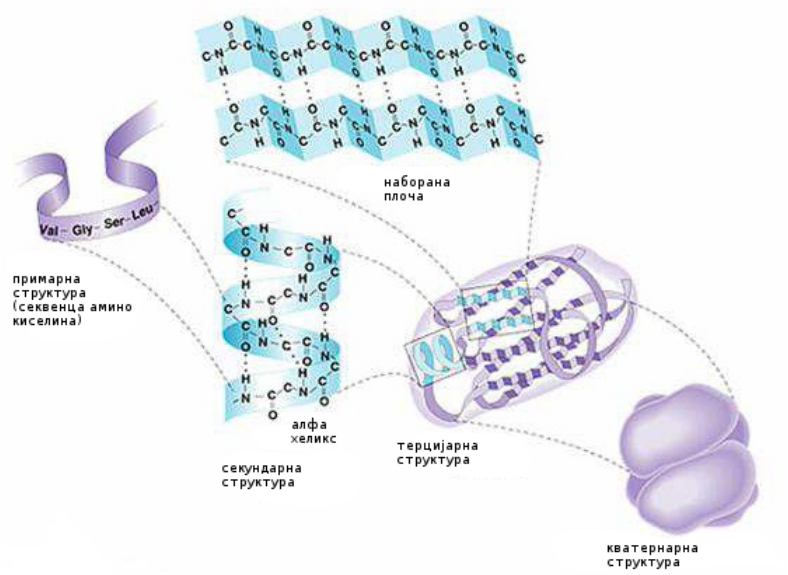
У зависности од броја полипептидних ланаца, протеини могу бити мономери, димери, тримери итд. Уколико се састоје од идентичних полипептидних ланаца, то се означава префиксом хомо- (исти), а уколико се састоје од различитих полипептидних ланаца - префиксом хетеро- (различит). Нпр. хетеродимер је протеин који се састоји од два различита полипептдина ланца [2].

1.5 Класификација аминокиселина

Аминокиселине се разликују према природи R-групе која улази у њихов састав. Ови бочни ланци могу бити хидрофилни или хидрофобни, кисели, базни или неутрални. Хемијски састав јединствене R-групе одговоран је за најважније карактеристике аминокиселина: хемијску реактивност, наелектрисање и релативну хидрофобност [12].

Наелектрисане аминокиселине могу бити киселе или базне. При ниској pH вредности[4], протеини су позитивно наелектрисани захваљујући базним групама,

[4]мера киселости или базности раствора



Слика 1.5: Нивои структуре протеина

као што је случај код лизина и аргинина, док су при високој pH вредности, негативно наелектрисани, услед киселих група, као у случају аспарагинске и глутаминске киселине [12].

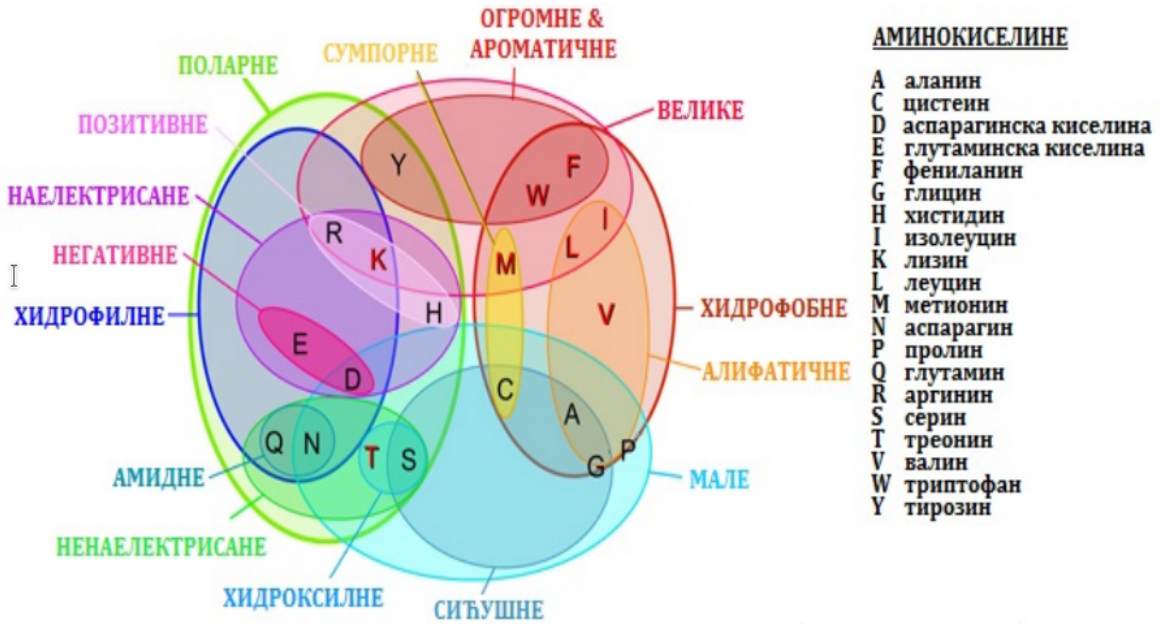
Основа пептидног ланца састављена је од аминокиселина које имају поларне, неполарне, ароматичне и наелектрисане остатке [12].

Осам аминокиселина има неполарне, хидрофобичне R-групе. Њих пет поседује алифатичне хидроугљеничне R-групе (аланин, леуцин, изолеуцин, валин и пролин), две имају ароматичне прстенове (фенилаланин и триптофан), док метионин садржи сумпор. Ове аминокиселине су мање растворљиве у води од поларних [12].

Поларне аминокиселине су растворљивије у води због својих R-група које могу да формирају водоничне везе са водом. Поларност серина, треонина и тирозина је последица њихових хидроксилних ($-OH$) група, док је код аспарагина и глутамина то последица њихових угљено-амидних група, а код цистеина - последица његових сулфхидрилних група ($-SH$). Глицин подразумевано спада у ову групу [12].

Аминокиселине које носе негативно наелектрисање за pH вредности 6.0-7.0 садрже другу карбоксилну групу. То су киселе аминокиселине - аспарагинска и глутаминска киселина. Базне аминокиселине, код којих R-групе имају позитивно наелектрисање за pH вредност 7.0 су лизин, аргинин и хистидин, који је слабо базан [12]. Слика 1.6 илуструје припадност посматраних аминокиселина различитим класама [6].

Осим споменутих, у овом раду коришћене су и додатне класе које представљају склоност појединих аминокиселина ка изградњи одређених секундарних структура. Те класе су: “склоне хеликсу”, “склоне равни”, “склоне заокрету” и “без склоности”. Цела листа коришћених класа налази се у табели 1.2, где је уједно приказан и однос аминокиселина са сваком од класа. Свака аминокиселина може имати један од два односа са класом - припадајући или неприпадајући. Уколико аминокиселина припада посматраној класи, то обележавамо знаком “+”,



Слика 1.6: Венов дијаграм класа аминокиселина

а уколико не припада, знаком “-”.

Табела 1.2: Класификација аминокиселина

Аминокиселине / Класе	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
поларне	-	+	+	+	-	+	+	-	+	-	-	+	-	-	-	+	+	-	+	-
позитивне	-	+	-	-	-	-	-	-	+	-	-	+	-	-	-	-	-	-	-	-
наелектрисане	-	+	-	+	-	-	+	-	+	-	-	+	-	-	-	-	-	-	-	-
негативне	-	-	-	+	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
хидрофилне	-	+	+	+	-	+	+	-	-	-	-	+	-	-	-	-	-	-	-	-
амидне	-	-	+	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ненаелектрисане	-	-	+	-	-	+	-	-	-	-	-	-	-	-	-	+	+	-	-	-
хидроксилне	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-
сићушне	+	-	-	-	+	-	-	+	-	-	-	-	-	-	-	+	-	-	-	-
мале	+	-	+	+	+	-	-	+	-	-	-	-	-	-	+	+	+	-	-	-
алифатичне	+	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	+
хидрофобне	+	-	-	-	+	-	-	-	-	+	+	-	+	+	-	-	-	+	-	+
велике	-	+	-	-	-	-	-	-	-	+	+	+	+	+	-	-	-	+	+	-
огромне и ароматичне	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+	+	-
сумпорне	-	-	-	-	+	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-
склоне хеликсу	+	+	-	-	-	+	+	-	-	-	+	+	+	-	-	-	-	-	-	-
склоне равни	-	-	-	-	-	-	-	-	-	+	-	-	-	+	-	-	+	+	+	+
склоне заокрету	-	-	+	+	-	-	-	+	-	-	-	-	-	-	+	+	-	-	-	-
без склоности	-	-	-	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-

2 Подаци и методе

Предикција секундарне структуре протеина, о којој је реч у овом раду, заснива се на рачунању корелација над скупом података добијених из Протеинске банке података (*Protein Data Bank, PDB*).

2.1 Подаци

Протеинска банка података је јединствени, глобални репозиторијум, експериментално одређених тродимензионалних структура биолошких макромолекула и њихових комплекса. PDB је успостављена 1971. године, чиме је постала први дигитални ресурс са отвореним приступом у оквиру биолошких наука. Архива тренутно броји око 142000 уноса, а за њено одржавање задужена је *Worldwide Protein Data Bank organization* (*wwPDB; wwpdb.org*) [14, 11]. Главна функција ове базе је да организује тродимензионалне структурне податке великих биолошких макромолекула, укључујући протеине и нуклеинске киселине свих организама, бактерија, гљива, биљака, инсеката и других животиња, као и људи. Све ове структуре одређене су експерименталним методама, попут кристалографије X-зрацима, спектроскопије нуклеарном магнетном резонанцом, електронском микроскопијом итд. [13].

Коришћени узорак, на основу кога су рачунате статистике, форматиран је у формату CSV (*comma-separated values*), са симболом "|" као сепаратором (уместо уобичајеног симбола ","). Сваки ред из узорка састоји се из четири блока, раздвојена сепаратором:

- PDBID - четворословни идентификатор протеина у бази PDB
- SEQID - једнословни идентификатор секвенце у оквиру протеина (протеини често имају више од једне секвенце)
- примарна структура - низ једнословних ознака аминокиселина
- секундарна структура - низ једнословних ознака секундарне структуре

Свакој аминокиселини из примарне стурктуре додељена је одговарајућа секундарна структура, на основу алгоритма DSSP. Као последица тога, низови примарне и секундарне структуре у оквиру једног реда, увек су исте дужине. Коришћени узорак има укупно 14483 реда, тј. 14483 секевенце протеина. Пример једне секвенце из узорка изгледа овако:

```
1ACW|A|VSaEDbPEHcSTQKAQAKaDNDKbVcEPI|  SSHHHHHHHTTT EEEEETTEEEEE
```

Дакле, у овом случају:

- `1ACW` је идентификатор протеина у бази PDB
- `A` је идентификатор секвенце у оквиру протеина
- `VSaEDbPEHcSTQKAQAKaDNDKbVcEPI` је примарна структура
- `␣␣SSHHHHHHHTTT␣EEEEETTEEEEE` је секундарна структура

Као што се може видети, ниска секундарне структуре може имати и бланко карактере (празнине), што значи да за ту позицију није позната секундарна структура. Приликом обраде података, свака појава бланко карактера у оквиру секундарне структуре, замењена је карактером C.

Мала слова у оквиру примарне структуре означавају посебне облике аминокиселина и игнорисана су овој анализи. Исти случај је и са појавом карактера Z, X, B, J, који представљају непознате или двосмислене аминокиселине. Селеноцистеин (U) и пиролизин (O), такође су искључене из анализе, па се све статистике рачунају коришћењем 20 стандардних аминокиселина.

Узорак је подељен на тренинг и тест скуп, при чему се 65% случајно изабраних секвенци из узорка користи као тренинг скуп, док је преосталих 35% коришћено као тест скуп. Корелације су рачунате над тренинг скупом, а тест скуп се користи као провера квалитета добијених резултата предикција.

2.2 Корелације

Као оцена структуре при предикцији коришћене су корелације. Корелација представља меру зависности две променљиве. У овом раду, анализира се повезаност пара аминокиселина са изградњом пара секундарних структура, а на основу односа који тај пар аминокиселина гради са сваком од класа.

Разлог због кога се не врши предикција за појединачне аминокиселине је што посматрањем парова суседних аминокиселина (биграма) релативно повећавамо величину узорка који ће бити коришћен за моделирање и нај начин добијамо релевантније резултате. Узимајући две по две аминокиселине из сваке секвенце у узорку, посматран је њихов однос са сваком од класа. Користећи нотацију `"+/-"`, сваки биграм узима један од четири могућа односа са неком класом:

- `"++"` - обе аминокиселине припадају класи
- `"+-"` - прва аминокиселина припада класи, док друга не припада
- `"-+"` - прва аминокиселина не припада класи, док друга припада
- `"--"` - ниједна аминокиселина не припада класи

Како је први корак ка добијању релевантнијих резултата посматрање парова аминокиселина уместо појединачних аминокиселина, други корак је посматрање груписаних биграма секундарне структуре уместо конкретних биграма. Груписани биграм је биграм који на првој или на другој позицији у биграму има посматрану, конкретну, секундарну структуру, док нам преостала позиција није од

значаја, тј. на њој се може наћи било која секундарна структура (може се наћи и иста секундарна структура као и она коју посматрамо). На позицију која нам није од значаја стављамо ознаку "X". Нпр. за секундарну структуру G, два могућа груписана биграма су GX и XG. Списак свих груписаних биграма дат је у табели 2.1.

Табела 2.1: Секундарне структуре и одговарајући груписани биграми

Секундарна структура	Груписани биграми	
H	HX	XH
E	EX	XE
T	TX	XT
G	GX	XG
S	SX	XS
B	BX	XB
I	IX	XI
C	CX	XC

За сваку класу и сваки груписни биграм секундарне структуре, рачунамо четири корелације, по једну за сваки од могућих односа биграма аминокиселина са класом (`"++"`, `"+-"`, `"-+"`, `"--"`). Дакле, рачунање корелације се не врши за конкретне парове аминокиселина, већ се парови замењују релацијом коју остварују са посматраном класом.

Корелација се рачуна помоћу Пирсоновог коефицијента за две промењљиве. Овај коефицијент представља меру линеарне зависности између две промењљиве. Коефицијент може узети вредности из опсега $[-1, 1]$, при чему вредност 1 означава савршену позитивну линеарну зависност између промењљивих, док вредност -1 означава савршену негативну линеарну зависност. Вредност 0 говори да не постоји линеарна зависност између промењљивих [16].

Својство симетричности се односи на то да је корелација између промењљиве X са промењљивом Y, иста као и корелација између промењљиве Y са промењљивом X. Друга важна особина корелације је њена отпорност на линеарне трансформације. То значи да множење једне промењљиве константом и/или сабирање са константом, не утиче на корелацију те промењљиве са другом промењљивом [16].

На примеру релације `"-+"` са класом наелектрисане и груписаног биграма секундарне структуре XH, у табели 2.2, приказане су вредности које је потребно обезбедити за рачунање Пирсоновог коефицијента.

Табела 2.2: Вредности коришћене за рачунање корелација

	Биграм гради секундарну структуру XH	
Биграм има однос `"-+"` са класом наелектрисане	✓	×
✓	A	B
×	C	D

Вредности имају следеће значење:

- A - број биграма аминокиселина из узорка које имају `"-+"` однос са класом наелектрисане, а граде секундарну структуру XH

- B - број биграма аминокиселина из узорка које имају `"-+"` однос са класом наелектрисане, а не граде секундарну структуру ХН
- C - број биграма аминокиселина из узорка које немају `"-+"` однос са класом наелектрисане, а граде секундарну структуру ХН
- D - број биграма аминокиселина из узорка које немају `"-+"` однос са класом наелектрисане и не граде секундарну структуру ХН

Затим би се описани поступак поновио и за преостале односе (`"++"`, `"+-"`, `"--"`) са класом наелектрисане и груписани биграм ХН, а онда и за остале груписане биграме секундарне структуре и остале класе. Према томе, како имамо 8 типова секундарних структура, имамо и 16 груписаних биграма који им одговарају, па за 19 коришћених класа и по 4 односа са сваком од њих, долазимо до $16*19*4 = 1216$ корелација које је потребно израчунати.

Формула која се користи за рачунање коефицијента корелација је следећа:

$$R = \frac{AD - BC}{\sqrt{(A+B)(A+C)(B+D)(C+D)}} \tag{1}$$

[17].

2.3 Python

За рачунање статистика, обраду и приказ резултата, коришћен је програмски језик *Python*. Python је програмски језик високог нивоа, опште намене и отвореног кода. Често се дефинише као објектно орјентисани скрипт језик. Осим објектно орјентисане, Python подржава и друге парадигме, као што су функционална и императивна. Користи се како за скрипт програме, тако и за самосталне програме. Философија језика инсистира на читљивости кода, употреби библиотека, као и на архитектури која оптимизује продуктивност програмера, квалитет софтвера и подршку за различите платформе, међу којима су и оне најчешће коришћене - Linux, Windows, Macintosh, Java, .NET, Android и iOS [8, 9].

У даљем тексту биће описане најважније структуре података језика Python, које су коришћене за потребе овог рада.

2.3.1 Листе

Листа је променљива, уређена секвенца објеката, чијим се члановима приступа помоћу индекса. Елементи листе смештају се унутар угластих заграда (`[]`), а одвајају зарезом [9]. Једна листа може да садржи елементе различитих типова.

Примери креирања листе:

- празна листа: `[]`

- листа стрингова: `['arcade', 'fire']`

- листа са елементима различитих типова:
  `['conversation', 16, ['city', 'middle'], 1.5, {}]`

- листа добијена од елемената стринг објекта:
  `list('high violet')`
  (резултат је: `['h', 'i', 'g', 'h', ' ', 'v', 'i', 'o', 'l', 'e', 't']`)

- листа добијена као резултат "list comperhension" операције:
  `[i.upper() for i in 'Win']`
  (резултат је: `['W', 'I', 'N']`, добијена применом методе `upper()` на сваки од карактера секвенце `'Win'` (метода `upper()` претвара мала слова у велика)

Елементима листе приступа се преко индекса, а бројање почиње од 0. Нпр:

```
a = ['dazed', 'and', 'confused']
    - a[2] → 'confused'
```

Могу се користити и негативни индекси. У том случају, бројање креће са краја листе, а први индекс је -1. Нпр:

```
a = ['dazed', 'and', 'confused']
    - a[-2] → 'and'
```

Променљивост листе односи се на то да се њени елементи могу променити, могу се додати нови елементи или се могу обрисати постојећи.

```
a = ['dazed', 'and', 'confused']
    - a[1] = '&' → ['dazed', '&', 'confused']

a = ['dazed', 'and', 'confused']
    - a.append(1) → ['dazed', 'and', 'confused', 1]

a = ['dazed', 'and', 'confused']
    - a[1:] → ['and', 'confused']
```

2.3.2 Торке

Торка је непроменљива, уређена секвенца објеката, чијим се члановима приступа помоћу индекса. За разлику од листе, чији се елементи записују између угластих заграда, елементи торке смештају се у мале заграде `()`. Као и код листи, елементи се раздвајају зарезом, а елементи једне торке могу бити различитих типова [9].

Примери торки:

- празна торка: `()`

- торка од једног елемента (зарез је обавезан, иначе би се изгубило својство торке): `(6,)`

- торка од три елемента: `(13, 'January', (19, 92))`

- други начин записа исте торке као у претходном примеру: `13, 'beaches', (19, 92)`

Као и код листи, приступ елементима се обавља преко индекса:

```
a = (13, 'January', (19, 92))
    -  a[1] → 'January'
    -  a[2][0] → 19
    -  a[-3] → 13
```

2.3.3 Речници

Речник је неуређена, променљива колекција података, где је сваки елемент облика `кључ: vrednost`. Елементи речника смештени су унутар пара витичастих заграда `{}`, а као и код листи и торки, раздвајају се зарезом. Међутим, за разлику од ових структура, код којих је сваки елемент имао своју позицију, односно, свој индекс у оквиру структуре преко кога му се могло приступити, дохватање елемента из речника обавља се преко кључа. Због тога је неопходно да кључеви у речнику буду јединствени. Такође, кључеви могу бити само непроменљиви објекти (бројеви, стрингови, торке) [9].

Примери речника:

- празан речник: `{}`

- речник од четири елемента:
  `{7: 'days', 5: {'a': 1, 'b': 2}, 'four': 'seasons', 12: 1}`

- речник од 2 елемента: `dict('a': 1, 'b': {'c': 2})`
  (резултат је: `{'a': 1, 'b': {'c': 2}}`)

Примери приступања елементима преко кључа:

```
d = {7: 'days', 5: {'a': 1, 'b': 2}, 'four': 4, (1, 2, 3): 6}}
    - d[7] → 'days'
    - d[5]['b'] → 2
    - d[(1, 2, 3)] → 6
```

Примери мењања речника:

```
d = {'a': 1, 'b': 2, 'c': 3}
    - d['e'] = 15 → {'a': 1, 'b': 2, 'c': 3, 'e': 15}

d = {'a': 1, 'b': 2, 'c': 3}
    - del d['b'] → {'a': 1, 'c': 3}

d = {'a': 1, 'b': 2, 'c': 3}
    - d['a'] = 10 → {'a': 10, 'b': 2, 'c': 3}
```

2.4 Марковљеви модели

Марковљев модел (Марковљев ланац) представља стохастички модел који описује секвенцу могућих догађаја, у коме вероватноћа сваког догађаја зависи само од стања постигнутог у претходном догађају [5]. Ово је Марковљев процес првог реда, где расподела вероватноћа зависи само од претходног стања, а не и од свих осталих. Односно, вероватноћа наредног стања директно зависи само од тренутног стања, а претходна стања постају ирелевантна једном када имамо тренутно стање [4]. Постоје и ланци вишег (n-тог) реда, код којих је наредно стање одређеног на основу претходних n стања.

Скривени Марковљев модел (енг. *Hidden Markov Model, HMM*) је уопштење Марковљевог ланца, у коме ("унутрашње") стање није директно видљиво (дакле, скривено је), али емитује видљиво ("спољашње") стање, које се још назива и

"емисија" или опсервација, на основу датог закона вероватноће. Уколико је број унутрашњих стања N, вероватноће преласка из једног стања у друго, описују се матрицом транзиције, димензија `NxN`. За број емисија M, вероватноће емисија описују се матрицом емисије димензија `NxM` [4].

Нека је:

- T - дужина посматране секвенце опсервација
- N - број стања у моделу
- M - број видљивих симбола (опсервација)
- $Q = \{q_0, q_1, ..., q_{N-1}\}$ - различита стања Марковљевог ланца
- $V = \{0, 1, ..., M - 1\}$ - скуп могућих опсервација
- A - вероватноће преласка између стања
- B - матрица вероватноћа опсервација
- π - почетна расподела вероватноћа
- $O = \{O_0, O_1, ..., O_{T-1}\}$ - секвенца опсервација [15].

Постоје три основна проблема која можемо да решимо користећи HMM.

1. За дати модел $\lambda = (A, B, \pi)$ и секвенцу опсервација O, одредити $P(O|\lambda)$, односно, одредити вероватноћу емитовања секвенце O, ако је дат модел λ. Овај проблем се решава алгоритмом унапред (енг. forward algorithm).

2. За дати модел $\lambda = (A, B, \pi)$ и секвенцу опсервација O, одредити оптималну секвенцу стања која је емитовала дату секвенцу опсервација. За решавање овог проблема, користи се Витерби алгоритам.

3. За дату секвенцу опсервација O и димензије N и M, одредити модел $\lambda = (A, B, \pi)$ који максимизује вероватноћу емитовања секвенце O. Овај проблем се може посматрати као тренирање модела који ће најбоље одговарати емитованој секвенци [15]. За његово решавање, користи се *Baum-Welch* алгоритам (forward-backward algorithm).

3 Имплементација

Како би се истражио утицај припадности аминокиселина одређеним класама на секундарне структуре које те аминокиселине граде, написани су алгоритми који, користећи различите технике комбиновања добијених корелација, дају процену "највероватније" секундарне структуре за дату секвенцу примарне структуре. У наредном делу биће приказани ови алгоритми, као и резултати које дају.

3.1 Предикција у односу на дату класу

Метода `predict_secondary_structure_using_class` за дату секвенцу примарне структуре и дату класу, одређује секундарну структуру која има највећу функцију оцене. У овом случају, функција оцене је збир корелација између односа који биграми аминокиселина остварују са датом класом и груписаних биграма секундарне структуре. Као међурезултат ове предикције, добија се листа груписаних биграма секундарне структуре, да би се, затим, додатном, помоћном методом `concat_summary_bigrams`, претворила у конкретну ниску секундарне структуре.

Аргументи функције су ниска примарне структуре за коју се врши предикција, класа у односу на коју се врши предикција, као и додатни параметар `num_of_max`, којим се задаје број највећих корелација по свакој релацији, који ће учествовати у израчунавању (подразумевана вредност параметра је 1).

Први корак у налажењу најбоље структуре представља налажење свих (валидних) путања, чија дужина одговара датој примарној структури. Путања је уређена листа релација које биграми аминокиселина остварују са посматраном класом. Нпр. ако посматрамо секвенцу GKELRMH у односу на класу хидрофилне, из табеле 1.2 видимо да њене аминокиселине K, E и R припадају класи хидрофилне (имају `"+"` однос са класом), док G, L, M и H не припадају (имају `"-"` однос са класом). Биграми које добијамо од дате секвенце, померајући се увек за једно место удесно су: GK, KE, EL, LR, RM, MH, па је одговарајућа путања (`"-+"`, `"++"`, `"+-"`, `"-+"`, `"+-"`, `"--"`).

Валидна путања подразумева да суседни елементи задовољавају својство да је знак на другој позицији у првом елементу, једнак знаку на првој позицији у другом елементу. Нпр. путања (`"-+"`, `"++"`) је валидна, јер се унутрашњи знакови поклапају (прва релација има знак `"+"` на другој позицији, док друга релација има знак `"+"` на првој позицији). Пример путање која није валидна би био (`"+-"`, `"+-"`), јер се симболи на захтевани позицијама разликују (знак `"-"` код првог елемента и знак `"+"` код другог). Захтев за валидношћу потиче од тога што се све секвенце из скупа података обрађују биграм по биграм, али тако да се увек померамо за једно место удесно. Због тога два суседна биграма увек деле једну аминокиселину, па

самим тим, релација на прелазу између биграма, односно, симбол на прелазу у путањи, мора бити исти код оба елемента.

Путање се добијају позивом методе `generate_paths`, чији је аргумент дужина путање. Дужина путање је увек за један мања од дужине ниске примарне структуре. У претходном примеру, од ниске GKELRMH, дужине 7, добијамо 6 биграма (GK, KE, EL, LR, RM, MH), па бисмо у том случају, тражили све валидне путање дужине 6. Метода `generate_paths` користи рекурзију да креира све валидне путање дате дужине.

Путање се проналазе техником "подели па владај"[5] (енг. *"divide and conquere"*). Задати број, који представља дужину путање, дели се на 2 и рекурзивно се налазе валидне путање за добијене краће дужине, након чега се врши њихово спајање. Излаз из рекурзије је путања дужине 1, што би биле наше базне релације: `"++"`, `"+-"`, `"-+"`, `"--"`. Путање се чувају у Python листама, а за сваку дужину, листе се смештају у једну заједничку листу, која се затим уписује у датотеку облика `paths_of_length_<duzina_putanje>.txt`, у директоријуму `generate_paths_results`.

Садржај датотеке `paths_of_length_1.txt`, која чува све путање дужине 1, изгледа овако:

```
[["++"], ["+-"], ["-+"], ["--"]]
```

док је нпр. за дужину 3, садржај датотеке `paths_of_length_3.txt` следећи:

```
[
  ["++", "++", "++"], ["++", "++", "+-"], ["++", "+-", "-+"],
  ["++", "+-", "--"], ["+-", "-+", "++"], ["+-", "-+", "+-"],
  ["+-", "--", "-+"], ["+-", "--", "--"], ["-+", "++", "++"],
  ["-+", "++", "+-"], ["-+", "+-", "-+"], ["-+", "+-", "--"],
  ["--", "-+", "++"], ["--", "-+", "+-"], ["--", "--", "-+"],
  ["--", "--", "--"]
]
```

При сваком наредном позиву методе, прво се проверава да ли већ постоји датотека за дату дужину путање. Уколико постоји, чита се њен садржај, док се у супротном, путања налази на описани начин. Чувањем резултата у датотекама, знатно се убрзавају даља рачунања, што је посебно битно у случају великог броја рекурзивних позива. Међутим, са дужином, експоненцијално расте број могућих путања за дату дужину, па датотеке у којима се чувају јако брзо постану "превелике". То за последицу има да њихово отварање, читање и затварање постаје споро. Ово се највише манифестује када је потребно програм покренути за велику количину података, односно за велики број секвенци, као што је у случају тестирања квалитета метода, коришћењем издвојеног тест скупа. Због тога су дуге секвенце дељене на краће, што је значајно убрзало проверу.

Након што нађемо путање, прелази се на проналажење највећих корелације за дату класу, по свакој од релација. Корелације се рачунају помоћу методе

[5] парадигма која подразумева да се проблем дели на потпроблеме који се решавају рекурзивно, а њихвоа решења се онда комбинују у решење почетног проблема

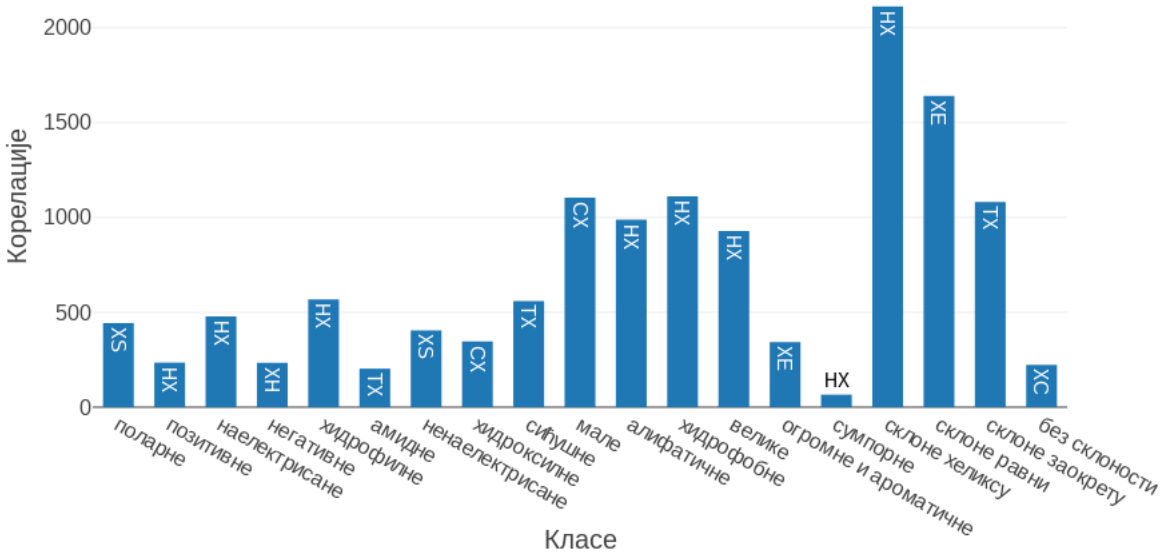
`calculate_summary_correlations`, где се, на већ описани начин, рачунају корелације између груписаних биграма секундарне структуре и релације коју биграми аминокиселина имају са прослеђеном класом, користећи Пирсонов коефицијент корелације.

Резултати се чувају у Python речницима, који се уписују у датотеке, из којих се касније могу читати уместо да се сваки пут изнова рачунају. Једна датотека одговара једној класи. Сваки од речника садржи четири елемента, такође речника, по један за сваку релацију. Нпр. речник за класу амидне изгледа овако:

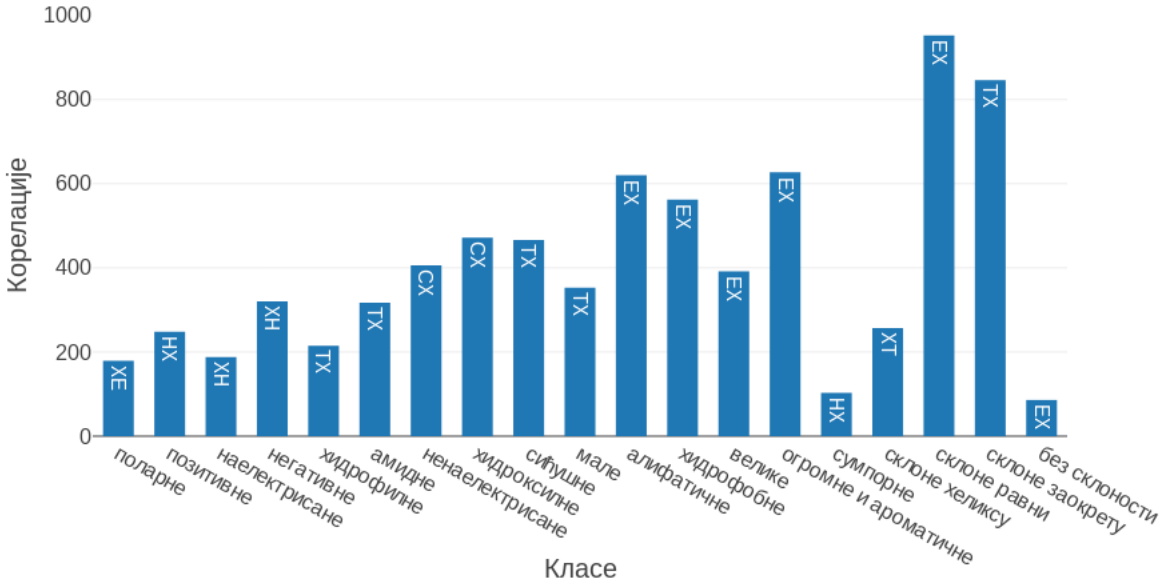
```
{
  "++": {
    "HX": 49, "XH": 12, "GX": 40, "XG": 11, "IX": -6, "XI": -6,
    "EX": -244, "XE": -242, "BX": -20, "XB": -14, "TX": 202,
    "XT": 166, "SX": 82, "XS": 98, "CX": -39, "XC": 31
  },
  "+-": {
    "HX": -23, "XH": 56, "GX": 41, "XG": 17, "IX": -20, "XI": -7,
    "EX": -420, "XE": -321, "BX": -25, "XB": 17, "TX": 316,
    "XT": 146, "SX": 140, "XS": 101, "CX": 97, "XC": 63
  },
  "-+": {
    "HX": 109, "XH": -14, "GX": 96, "XG": 51, "IX": -12, "XI": -20,
    "EX": -308, "XE": -421, "BX": 3, "XB": -27, "TX": 230,
    "XT": 328, "SX": 89, "XS": 134, "CX": -104, "XC": 78
  },
  "--": {
    "HX": -74, "XH": -33, "GX": -109, "XG": -51, "IX": 24, "XI": 21,
    "EX": 584, "XE": 593, "BX": 21, "XB": 10, "TX": -442,
    "XT": -382, "SX": -185, "XS": -193, "CX": 14, "XC": -109
  }
}
```

Као што се може видети, елементи ових "подречника" одговарају паровима `grupisani_bigram: korelacija`. Све корелације у речницима помножене су фактором 10000, а затим заокружене на цео број, како би резултати били једноставнији за тумачење.

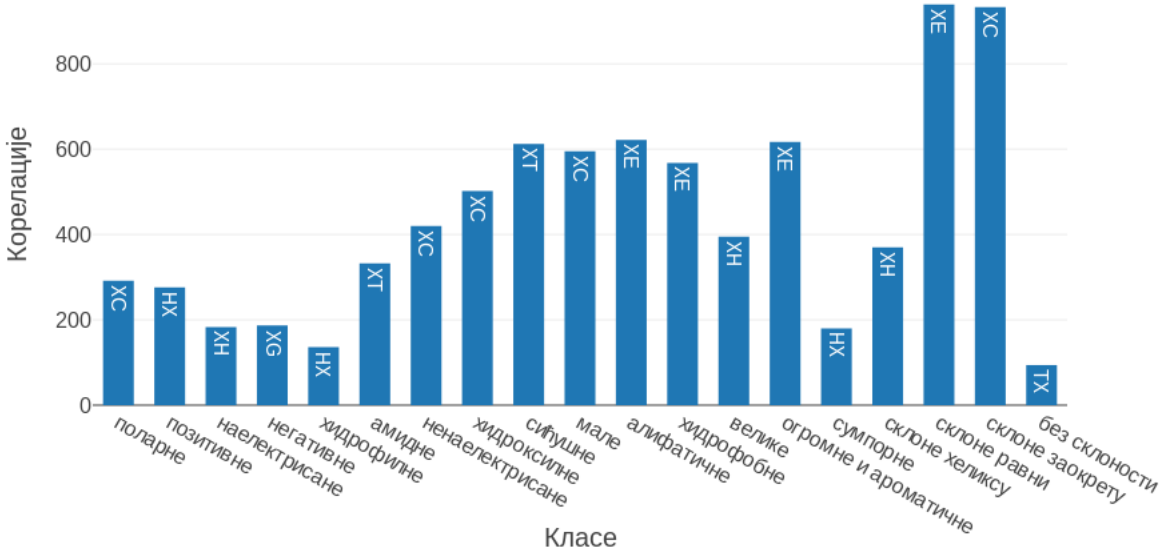
Метода `find_max_correlations` враћа највеће корелације за дату класу. Број највећих корелација из сваке класе који ће бити враћен, одређује се параметром `num_of_max`. Вредност параметра 1, која је и подразумевана вредност, говори да ћемо из сваке класе узимати само по један груписани биграм за сваку релацију и то онај који има највећу корелацију са класом. Због тога, за сваку путању добијамо тачно један могући резултат. На графицима 3.1, 3.2, 3.3 и 3.4 могу се видети груписани биграми који имају највећу корелацију са сваком класом, а за тражену релацију.



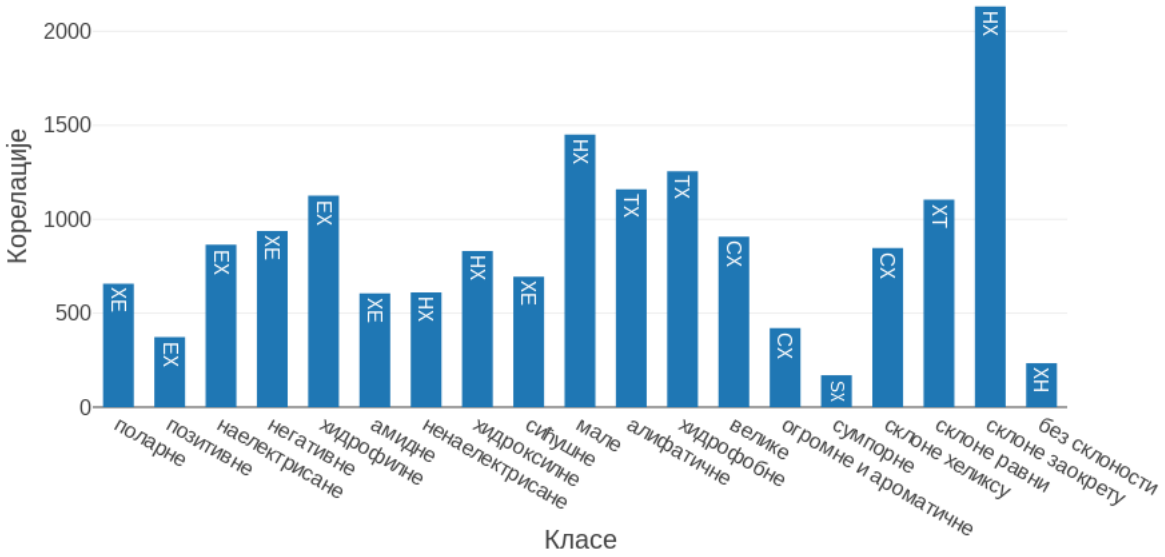
Слика 3.1: Груписани биграми са највећим корелацијама за релацију "++"



Слика 3.2: Груписани биграми са највећим корелацијама за релацију "+-"



Слика 3.3: Груписани биграми са највећим корелацијама за релацију `"-+"`



Слика 3.4: Груписани биграми са највећим корелацијама за релацију `"- -"`

Како узимање искључиво највеће корелације за једну релацију и једну класу може да не даје праву слику, идеја је да се параметризује број највећих корелација

по свакој релацији који ће бити узет у обзир. На тај начин, постављањем параметра `num_of_max` на вредност већу од 1, даје се прилика и осталим груписаним биграмима секундарне структуре да учествују у суми и у комбинацији са другим биграмима, евентуално буду изабрани као највероватнији. Да бисмо добили све могуће комбинације нађених максималних груписаних биграма, које одговарају овим путањама, налазимо њихов Декартов производ. Следећи пример илуструје ове кораке за класу хидрофобне и дужину секвенце 4 (тј. дужину путање 3) и вредност параметра `num_of_max` 2.

Све могуће путање дужине 3 су:

```
("++", "++", "++"), ("++", "++", "+-"), ("++", "+-", "-+"),
("++", "+-", "--"), ("+-", "-+", "++"), ("+-", "-+", "+-"),
("+-", "--", "-+"), ("+-", "--", "--"), ("-+", "++", "++"),
("-+", "++", "+-"), ("-+", "+-", "-+"), ("-+", "+-", "--"),
("--", "-+", "++"), ("--", "-+", "+-"), ("--", "--", "-+"),
("--", "--", "--").
```

Дакле, узимамо у обзир само валидне путање, тј. оне које немају недозвољене прелазе. Помоћу функције `find_max_correlations` за нашу класу хидрофобне и `num_of_max` 2, налазимо по 2 највеће корелације за сваки од односа са класом. Резултат је речник:

```
{
  "++": [{"HX": 1111}, {"EX": 1061}],
  "+-": [{"EX": 562}, {"HX": 224}],
  "-+": [{"XE": 568}, {"XH": 355}],
  "--": [{"TX": 1256}, {"XT": 1224}]
}
```

Након тога, обрађује се једна по једна путања, налазећи за сваку Декартов производ свих могућих груписаних биграма из добијеног речника који могу да је формирају. Нпр. за путању `("++", "+-", "--")`, на првој позицији, за однос `"++"` са класом хидрофобне, могу се наћи груписани биграми HX и EX (са корелацијама 1111 и 1061, респективно), на другој позицији, за однос `"+-"`, могу се наћи груписани биграми EX и HX (са корелацијама 562 и 224, респективно), док се на последњој, трећој позицији, за однос `"--"`, могу наћи TX и XT (са корелацијама 1256 и 1224, респективно). Декартов производ за ову путању дужине 3 и за по 2 могућа груписана биграма на свакој од позиција, даје $2^3 = 8$ начина за њихово комбиновање:

- HX, EX, TX (сума корелација је $1111 + 562 + 1256 = 2929$)
- HX, EX, XT (сума корелација је $1111 + 562 + 1224 = 2897$)
- HX, HX, TX (сума корелација је $1111 + 224 + 1256 = 2591$)
- HX, HX, XT (сума корелација је $1111 + 224 + 1224 = 2559$)
- EX, EX, TX (сума корелација је $1061 + 562 + 1256 = 2879$)

- EX, EX, XT (сума корелација је $1061 + 562 + 1224 = 2847$)
- EX, HX, TX (сума корелација је $1061 + 224 + 1256 = 2541$)
- EX, HX, XT (сума корелација је $1061 + 224 + 1224 = 2509$)

Ови резултати се чувају у речнику, а његов део, који се односи на путање представљене изнад, изгледа овако:

```
[
  ...,
  {"sum":2929,"path":["++","+-","--"],"bigrams":["HX","EX","TX"]},
  {"sum":2897,"path":["++","+-","--"],"bigrams":["HX","EX","XT"]},
  {"sum":2591,"path":["++","+-","--"],"bigrams":["HX","HX","TX"]},
  {"sum":2559,"path":["++","+-","--"],"bigrams":["HX","HX","XT"]},
  {"sum":2879,"path":["++","+-","--"],"bigrams":["EX","EX","TX"]},
  {"sum":2847,"path":["++","+-","--"],"bigrams":["EX","EX","XT"]},
  {"sum":2541,"path":["++","+-","--"],"bigrams":["EX","HX","TX"]},
  {"sum":2509,"path":["++","+-","--"],"bigrams":["EX","HX","XT"]},
  ...
]
```

За сваку класу, дужину путање и број највећих корелација који се узима за сваку од њих, добијени речник уписује се у датотеку, чије је име облика `seq_est_cls_<ime_klase>_length_<duzina_putanje>_num_of_max_<broj_maksimalnih>.txt`, а која се чува у директоријуму `predict_sec_str_using_class_results`.

Након што се из речника узме листа груписаних биграма са највећим збиром корелација, потребно је извршити њихово спајање у конкретну ниску секундарне структуре. За то је задужена метода `concat_summary_bigrams`, која уз помоћ полазне ниске примарне структуре формира коначну секундарну структуру.

У зависности од тога да ли при спајању два груписана биграма имамо конкретну секундарну структуру или ознаку "X" на преклапајућој позицији, разликујемо 4 случаја:

(1) $XS_1, S_1X \rightarrow \_S_1\_$

(2) (a) $XS_1, XS_2 \rightarrow \_S_1S_2$
(b) $S_1X, S_2X \rightarrow S_1S_2\_$

(3) $XS_1, S_2X \rightarrow \_S_1\_$ или $\_S_2\_$

(4) $S_1X, XS_2 \rightarrow S_1S_3S_2$

У првом случају, суседни груписани биграми дају исту секундарну структуру на унутрашњим позицијама, па ће она и бити изабрана за резултујућу ниску. Код другог случаја, један од груписаних биграма има конкретну секундарну структуру на унутрашњој позицији, док други има "X", па ће конкретна секундарна структура бити изабрана.

Трећи и четврти случај су ситуације у којима се морају применити додатне технике како би се изабрала секундарна структура. У трећем случају имамо колизију

на унутрашњим позицијама, јер су суседни груписани биграми дали различите секундарне структуре за исту примарну структуру. Овај проблем се решава тако што се изабере онај груписани биграм који има већу суму корелација по свим класама за ту позицију. Дакле, за оба конфликтна биграма, ради се следеће: узме се биграм аминокиселина који се налази на тој позицији и за сваку класу, нађе се његов однос са том класом, а онда и корелација коју конфликтни груписани биграм има са нађеном релацијом и текућом класом. Корелације се сабирају и добијени збир се пореди са збиром који је исти поступак дао за други конфликтни биграм. У случају да су збирови једнаки, можемо изабрати произвољни груписани биграм као резултујући, па тако овде узимамо секундарну структуру из првог груписаног биграма.

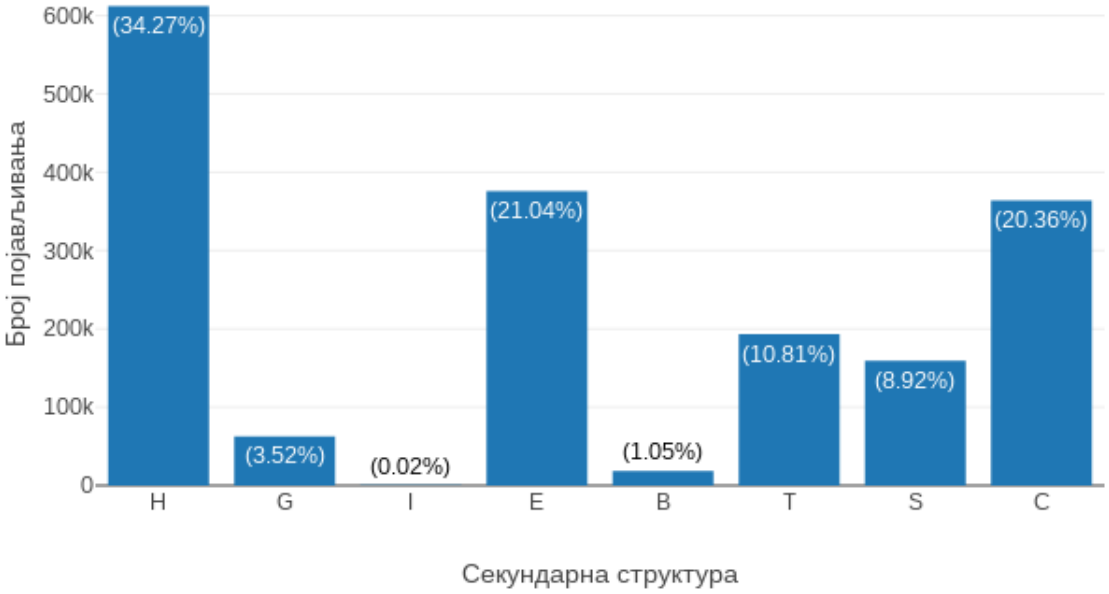
Четврти случај даје најмање информација о томе која би секундарна структура требало да се налази на траженој позицији, јер су унутрашњи чланови груписаних биграма "X". За решавање ове ситуације, поново се користи примарна структура од које смо кренули. Нека су нпр. наши конфликтни груписани биграми потекли од примарне структуре $P_1P_2P_3$. Налазимо коју секундарну структуру та ниска примарне структуре најчешће гради у нашем скупу података. Нека је то ниска $S_4S_3S_5$. Дакле, на позицији која је у питању, налази се S_3, па ће она бити изабрана, независно од тога што се спољне секундарне структуре не поклапају са оним што су дали претходно одређени груписани биграми.

Преостали необрађени случајеви су када први груписани биграм почиње симболом "X" и када се последњи груписани биграм завршава симболом "X". Оба случаја решавају се на сличан начин као и случај (4). Нађу се секундарне структуре које најчешће граде биграми примарних структура на тим позицијама, а које се поклапају са оним што су груписани биграми дали на другој позицији у првом случају, односно, на првој позицији у другом случају. Ако нпр. добијемо да је први груписани биграм XG, а да је биграм примарне структуре на тој позицији FL, тражићемо биграм секундарне структуре који FL најчешће гради, а који на другој позицији у биграму има G. Налазимо да је то биграм GG, па ће G бити резултат. Уколико не нађемо ниједан биграм секундарне структуре који има тражени облик (у овом случају, биграм који има G на другој позицији), за резултујућу секундарну структуру стављамо H, јер се најчешће јавља у скупу података (што се може видети на графику 3.5). Сличан поступак примењује се и за случај када је последња секундарна структура у резултату недефинисана.

Један пример би био спајање листе груписаних биграма `["XC", "CX", "SX", "XT", "GX", "XH"]`, која даје ниску секундарних структура CCSHGHH.

Након добијања конкретне ниске секундарне структуре, ради се "поправљање" резултата. Као што се може видети на графику 3.5, H и E су најзаступљеније секундарне структуре у скупу података, а додатно, најчешће се јављају у дугим, непрекинутим секвенцама. Секундарна структура H се ретко јавља у секвенцама краћим од 3-4, док је обично тај број између 6-10. Ове секвенце могу бити и доста дуже, али обично не прелазећи дужину од 45. Слична ситуација је и са секундарном структуром E - најчешће се јавља у секвенцама дужине 4-10 [1]. Ова својства искоришћена су за потенцијално побољшавање резултата.

Наиме, уколико се у добијеном резултату, пронађе секвенца H, прекинута једним појављивањем секундарне структуре T или 1-2 појављивања секундарне



Слика 3.5: Учесталост појављивања секундарних структура у скупу података

структуре C, ти случајеви се сматрају мало вероватним, па је претпоставка да је и на тим позицијама, такође, највероватније H. Ови прекиди, односно секундарне структуре које су се нашле између H, биће "претворене" у H. Исти поступак примењује се и код секвенци E. Нпр. уколико након спајања добијемо секвенцу THHHHCCHHSS, она ће овим поступком бити претворена у THHHHHHHHSS, чиме је добијен вероватнији резултат.

3.2 Предикција у односу на све класе

За разлику од методе `predict_secondary_structure_using_class`, метода `predict_secondary_structure_using_all_classes` не прихвата класу за коју ће се вршити предикција, већ само ниску примарне структуре протеина, док однос њених аминокиселина са сваком од класа, учествује у предикцији. Осим ниске примарне структуре, додатни аргумент методе је и параметар `num_of_max`, са истим значењем као и до сада - да одреди број максималних груписаних биграма (број груписаних биграма са максималним корелацијама) из сваке класе, који ће учествовати у предикцији.

Нека је $P_1P_2P_3$ дата ниска примарне структуре и нека је `num_of_max = 2` и нека постоје само две класе - K_1 и K_2. За сваку позицију у резултујућој секундарној структуру (број позиција одговара дужини полазне ниске примарне структуре), пратићемо онолико сума корелација, колико се различитих секундарних структура јави на тим позицијама кроз груписане биграме.

Прво обрађујемо биграм P_1P_2. Нека је његов однос са класом K_1 једнак rel_1. Помоћу методе `find_max_correlations` за класу K_1 и `num_of_max = 2` добијамо по 2 максимална груписана биграма за сваку релацију и од њих узимамо само оне који имају исту релацију као P_1P_2 са K_1, тј. узимамо само листу груписаних биграма за релацију rel_1. Нека је та листа `[{"SX": 520}, {"XT": 311}]`. У овом случају, за прву позицију у коначној секундарној структуру, имамо само једног кандидата, тј. само једну конкретну секундарну структуру - S, па ћемо збир за прву позицију и за S ($S_{1,S}$), иницијализовати са 520. За другу позицију, имамо конкретну секундарну структуру T, па иницијализујемо збир за другу позицију и секундарну структуру T ($S_{2,T}$) на 311. У овом тренутку, имамо следеће суме:

- за прву позицију: $S_{1,S} = 520$
- за другу позицију: $S_{2,T} = 311$

Затим се исти поступак понавља и за другу класу. Нека је однос P_1P_2 са K_2 једнак rel_2 и нека је листа груписаних биграма са највећим корелацијама за K_2 и rel_2 једнака `[{"TX": 110}, {"XU": 95}]`. Сада ћемо за прву позицију пратити и збир корелација за T ($S_{1,T}$) и иницијализовати га на 110. Биграм XU нам даје $S_{2,U}$ за другу позицију, па нове суме изгледају овако:

- за прву позицију: $S_{1,S} = 520$, $S_{1,T} = 110$
- за другу позицију: $S_{2,T} = 311$, $S_{2,U} = 95$

Нека за наредни биграм примарне структуре P_2P_3, добијемо листу `[{"TX": 342}, {"VX": 118}]` са класом K_1. За другу позицију у коначној секундарној структури, посматрајући последњу листу, добијамо конкретну секундарну структуру T, па ћемо вредност 342 додати на суму $S_{2,T}$ и добити $311 + 342 = 653$. Други елемент листе каже да се на другој позицији у кончаној секундарној структури може наћи и V, па иницијализујемо нову суму $S_{2,V}$ са 118. Тренутне суме изгледају овако:

- за прву позицију: $S_{1,S} = 520$, $S_{1,T} = 110$
- за другу позицију: $S_{2,T} = 653$, $S_{2,U} = 95$, $S_{2,V} = 118$

За класу K_2 и биграм P_2P_3 добијемо листу `[{"XT": 368}, {"VX": 344}]`, па након ажурирања постојећих сума и иницијализовања нових, имамо:

- за прву позицију: $S_{1,S} = 520$, $S_{1,T} = 110$
- за другу позицију: $S_{2,T} = 653$, $S_{2,U} = 95$, $S_{2,V} = 462$
- за трећу позицију: $S_{3,T} = 368$

За сваку позицију у резултату узимамо ону секундарну структуру за коју смо добили највећи збир - за прву позицију то ће бити S, за другу T и за трећу, такође T. Према томе, коначна секундарна структура је STT.

Слично као и код `predict_secondary_structure_using_class` и овде се може десити да за неку позицију немамо ниједну суму, односно ниједну секундарну структуру као кандидата за ту позицију. До тог случаја долази када добијемо само груписане биграме секундарне структуре који имају "X" на позицији која је у питању. У овим ситуацијама, када није очигледно који конкретни биграми треба да се нађу на датим местима, морамо да употребимо додатне методе избора.

Могуће ситуације са недефинисаним секундарним структурама су:

(1) $\_S$

(2) $S\_$

(3) $\_\_$

Први случај се јавља када је прва позиција у резултату недефинисана, док за другу позицију имамо конкретну секундарну структуру. У тој ситуацији, проналазимо најчешћи биграм секундарне структуре, који биграм примарне структуре са те позиције гради, а да на другој позицији има S и узимамо оно што се налази на његовој првој позицији. Уколико не нађемо биграм који на другој позицији има S, изабраћемо H као резултујућу секундарну структуру за ту позицију, јер се, као што смо видели на графику 3.5, H најчешће јавља у скупу података, па има смисла претпоставити да се и овде налази.

Други случај настаје када након конкретне секундарне структуре, имамо недефинисану. Слично као и код првог случаја, налазимо најчешћи биграм секундарне структуре који биграм примарне структуре на тој позицији гради, а који на првој позицији има S и узимамо оно што се налази на његовој другој позицији. Уколико не нађемо одговарајући биграм, поново бирамо H као резултујућу секундарну структуру.

До трећег случаја долази када су прве две позиције у резултату недефинисане. Примењујемо сличан поступак као у претходним ситуацијама - налазимо најчешћи биграм секундарне структуре који биграм примарне структуре за прве две позиције гради и узимамо секундарну структуру са његове прве позиције. Друга позиција остаје недефинисана, али је проблем сада сведен на други случај, па примењујемо одговарајући поступак.

Након добијања конкретне ниске секундарне структуре, као и код методе `predict_secondary_structure_using_class`, ради се "поправљање" резултата. Секвенце секундарних структура H и E, које су прекинуте једним појављивањем секундарне структуре T или 1-2 појављивања секундарне структуре C, биће кориговане замењивањем ових "прекида", структурама H и E, респективно.

3.3 Најбоља класа за сваки груписани биграм

Метода `best_class` не врши предикцију секундарне структуре, већ пружа увид у успешност предвиђања на основу израчунатих корелација и параметризације методе `find_max_correlations`. Она за дати груписани биграм секундарне структуре и параметар `num_of_max`, враћа два Python речника из којих се може прочитати колико пута је свака класа тачно, односно нетачно, предвидела дати биграм у скупу података. `num_of_max`, као и у претходно описаним метода, има

улогу у контролисању броја максималних биграма који ће бити коришћени за сваку релацију. Резултујући речници, осим кључева који представљају класе, имају и додатни кључ `count`, који означава колико је укупно пута дати биграм предвиђен тачно, односно нетачно, независно од класе. Део речника који приказује тачна предвиђања за `num_of_max = 1` изгледа овако:

```
{
  "HX": {
    "count": 1350991
    "Hydroxylic": 266535, "Uncharged": 219844, "Turn-pref": 215298,
    "Small": 137655, "Hydrophillic": 114412, "Helix-pref": 94057,
    "Positive(Basic)": 90164, "Hydrophobic": 64238, "Large": 58623,
    "Aliphatic": 42281, "Charged": 28522, "Sulphur Containing": 19362,
    "Polar": 0, "Negative(Acidic)": 0, "Amide": 0, "Tiny": 0,
    "Very Large & Aromatic": 0, "Strand-pref": 0, "Noss-pref": 0
   },
   "XH": {
     "count": 646767,
     "Noss-pref": 303132, "Charged": 130782, "Large": 83468,
     "Helix-pref": 82111, "Negative(Acidic)": 47274, "Polar": 0,
     "Positive(Basic)": 0, "Hydrophillic": 0, "Amide": 0, "Uncharged": 0,
     "Hydroxylic": 0, "Tiny": 0, "Small": 0, "Aliphatic": 0,
     "Hydrophobic": 0, "Very Large & Aromatic": 0, "Sulphur Containing": 0,
     "Strand-pref": 0, "Turn-pref": 0
   },
  ...
}
```

Речници су сортирани опадајуће, по броју тачних, односно, нетачних предвиђања, а чувају се у датотекама облика
`best_class_true_counter_num_of_max_<num_of_max>.txt` и
`best_class_false_counter_num_of_max_<num_of_max>.txt`.

Само пребројавање обавља функција `find_best_classes`, са параметром `num_of_max`, која пролази кроз сваку секвенцу у скупу података, биграм по биграм, узимајући у обзир само валидне биграме примарне структуре и њима одговарајуће биграме секундарне структуре. Од сваког биграма секундарне структуре формирају се два груписана биграма, тако да је код једног задржавамо прву секундарну структуру, а на другу позицију стављамо "X", док код другог на прву позицију стављамо "X", а другу секундарну структуру задржавамо. Нпр. уколико у секвенци из скупа података наиђемо на биграм TE, од њега формирамо 2 груписана биграма - TX и XE.

Затим се користи метода `find_max_correlations` да се за сваку класу пронађу предикције - груписани биграми који дају највеће корелације са том класом и то, онолико највећих за сваку релацију, колико је одређено параметром `num_of_max`. Нпр. за класу алифатичне, `num_of_max = 2` и релацију `"--"`, метода враћа `[{"TX": 1160}, {"XT": 1076}]`. Како је предикција, између осталог, дала и биграм TX, увећавамо бројач за тренутну класу и TX, као и бројач који броји укупан број успешних предвиђања. У случају биграма XT, који се не поклапа ни са једним од очекиваних биграма (TX и XE), увећавамо бројач неуспешних пред-

вићања за тренутну класу и биграм ХТ, као и бројач укупног броја неуспешних предвићања.

4 Резултати и дискусија

У наредно делу, дата је анализа понашања имплементираних алгоритама за различите вредности коришћених параметара.

4.1 Алгоритам унапред

Као што је споменуто у 2.4, алгоритам унапред се користи за одређивање вероватноће емитовања дате секвенце, уколико имамо модел. У тренинг скупу су пребројани преласци између стања у оквиру сваке класе, где су стања заправо релације (`"++"`, `"+-"`, `"-+"`, `"--"`), а затим је извршена нормализација. На тај начин, формирана је матрица транзиције. За добијање матрице емисије, уместо вероватноћа, коришћене су израчунате корелације. Међутим, због релативно малог тренинг скупа, имплементација алгоритма на овај начин није дала значајне резултате, па је приступ напуштен. Уместо њега, акценат је стављен на имплементацију метода `predict_secondary_structure_using_class` и `predict_secondary_structure_using_all_classes`, где је покушано другачијим приступом остварити релевантније резултате за дати скуп података.

4.2 Предикција у односу на дату класу

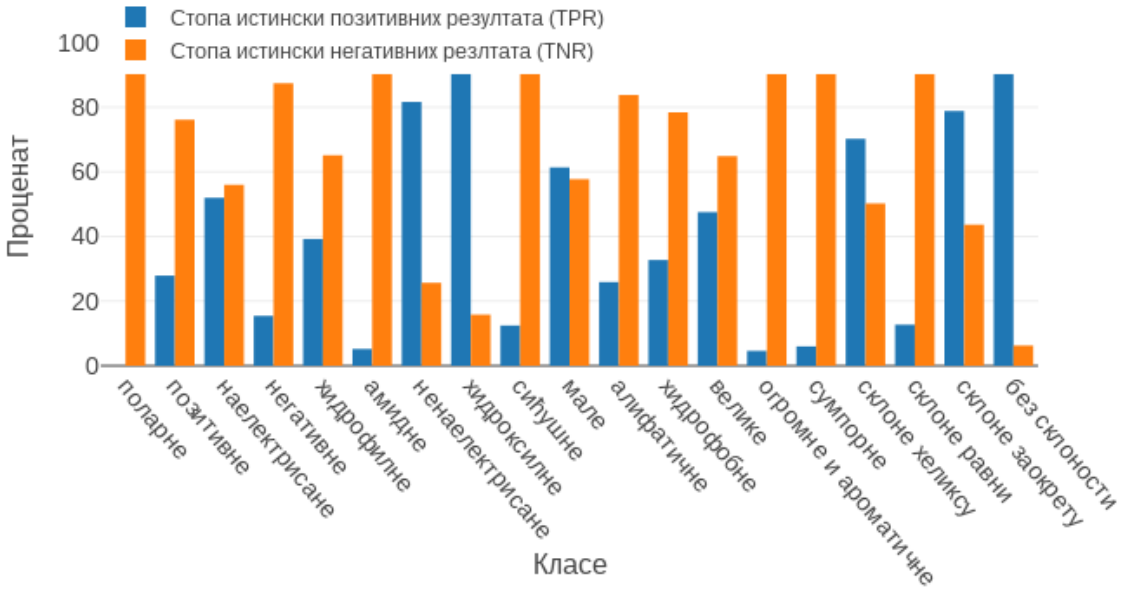
Анализа понашања методе `predict_secondary_structure_using_class`, обављена је над тест скупом, раније одвојеним за проверу квалитета резултата. Извршене су две оцене - једна која у фокус ставља секундарне структуре и друга, која се бави секвенцама.

За сваку секундарну структуру и сваку класу, израчунате су две статистичке оцене - стопа истински позитивних резултата и стопа истински негативних резултата. Стопа истински позитивних резултата (одзив, енг. *true positive rate, sensitivity, recall*) је удео стварно позитивних резултата који су препознати као позитивни. Стопа истински негативних резултата (енг. *true negative rate, specificity, selectivity*) је удео стварно негативних резултата који су препознати као негативни.

У контексту који се овде користи, стопа истински позитивних резултата (TPR) и стопа истински негативних резултата (TNR) израчунате су за сваку секнударну структуру и сваку класу. TPR је удео посматране секундарне структуре исправно препознате користећи дату класу, у односу на укупан број њених јављања у тест скупу. Нпр. секундарна структура H је у односу на класу сићушне (користећи корелације израчунате за H и класу сићушне) тачно препозната 40356 пута и још 284110 пута нетачно препозната као нека од преосталих секундарних структура, па је TPR у овом случају $40356/(40356 + 284110) = 0,1244$, односно 12,44%, што

се може видети на графику 4.1. TNR је удео секундарних структура различитих од посматране, препознатих као заиста различите од посматране користећи дату класу, а у односу на укупуан број јављања преосталих секундарних структура у скупу. У случају секундарне структуре H и и класе сиђушне, тај проценат износи 90,89%.

Различите вредности параметра `num_of_max` нису дале битније разлике у резултатима, па су све наредне статистике дате за `num_of_max = 1`.



Слика 4.1: Однос вредности TPR и TNR за секундарну структуру H и различите класе

Са графика 4.1 видимо још да класе ненаелектрисане, хидроксилне и класа аминокиселина без склоности, остварују високе резултате за тачну предикцију секундарне структуре H, али са друге стране, доста лоше откривају случајеве када није у питању H. Супротно од њих, класе поларне, негативне, амидне, сиђушне, алифатичне, огромне и ароматичне, сумпорне, као и класа аминокиселина склоних равни, са високим процентом откривају секундарне структуре које нису H, али са доста малим процентом предвиђају H. Класе које имају релативно уједначене резултате за оцене TPR и TNR (између 40% и 70%) су наелектрисане, мале, велике и класа аминокиселина склоних хеликсу.

На графицима 4.2, 4.3 и 4.4 може се видети да је за секундарне структуре G, I и B, респективно, стопа истински негативних резултата за све класе између 99% и 100%, а стопа истински позитивних резултата тек 0-1%. Из овога се може закључити да је метода `predict_secondary_structure_using_class` немоћна у њиховом детектовању.

Класе поларне, наелектрисане, хидрофилне и класа аминокиселина склоних равни дају приближно уједначене резултате за оцене TPR и TNR (између 40% и

Слика 4.2: Однос вредности TPR и TNR за секундарну структуру G и различите класе

70%) и секундарну структуру E. Негативне, амидне и сићушне добро предвиђају E (70-90%), али са доста ниским процентима откривају остале секундарне структуре (15-40%). Упадљиво високе проценте за TNR, а ниске за TPR дају остале, што се може видети на 4.5.

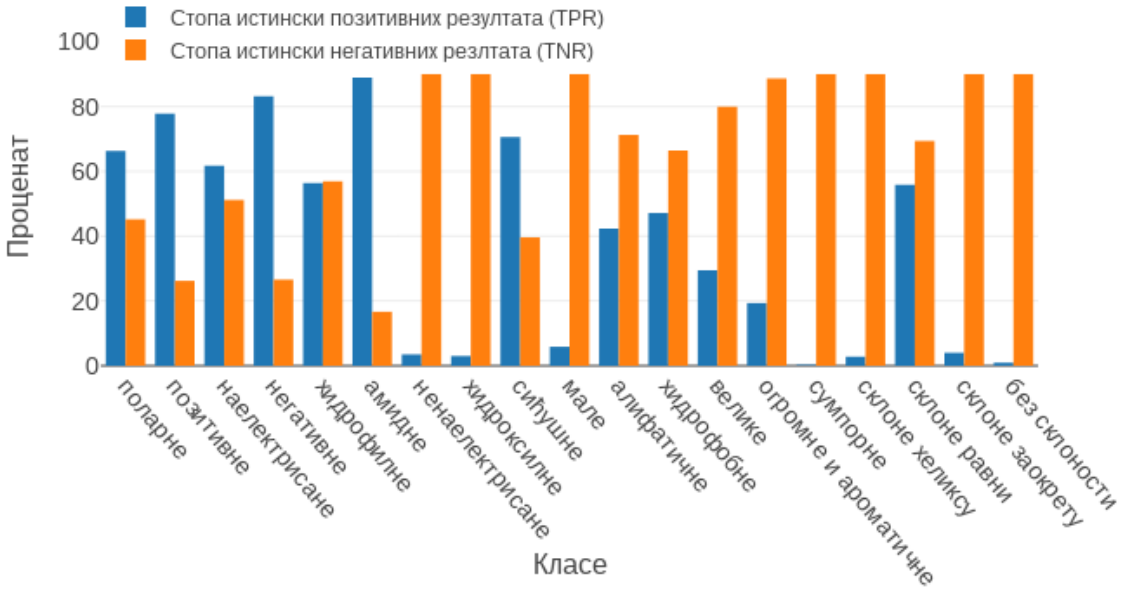
За секундарну структуру T, најбоље резултате дају класе алифатичне, хидрофобне и класа аминокиселина склоних равни - њихови проценти за обе мере су између 50% и 70%. Класе аминокиселина склоних заокрету и сићушних аминокиселина са око 35% успевају да тачно предвиде T, а са 75-80% откривају да није у питању T. Остале класе готово да никада не детектују исправно секундарну структуру T, већ само успешно откривају случајеве када су у питању остале секундарне структуре, што је приказано на 4.6.

Код секундарне структуре S, класа сумпорних аминокиселина је успешно предвиђа са процентом од 90,37, али доста слабо исправно препознаје остале секундарне структуре (11,89%). Осим ње, једино још класа поларне успева да у мало значајнијој мери препозна S - 29,45%. Са графика 4.7 видимо још да све остале класе готово никада исправно не детектују S, већ само случајеве када није у питању S.

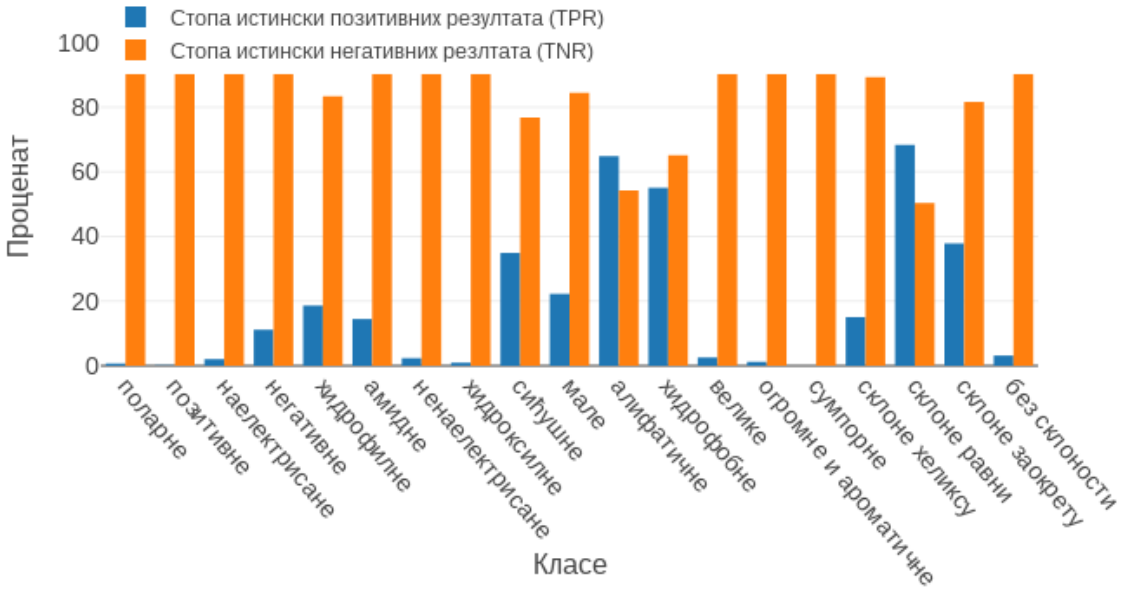
На 4.8 уочавамо да класе мале, велике и класа аминокиселина склоних хеликсу дају најбоље резултата када је у питању предикција секндарне структуре C. Класа малих даје 43,95% за TPR и 71,58% за TNR, велике дају 46,76% за TPR и 64,70% за TNR, а класа аминокиселина склоних хеликсу 39,98% за TPR и 71,86% за TNR. Остале класе само поуздано откривају негативне инстанце (секундарне структуре различите од C), а у релативно занемарљивом броју случајева откривају C.

Слика 4.3: Однос вредности TPR и TNR за секундарну структуру I и различите класе

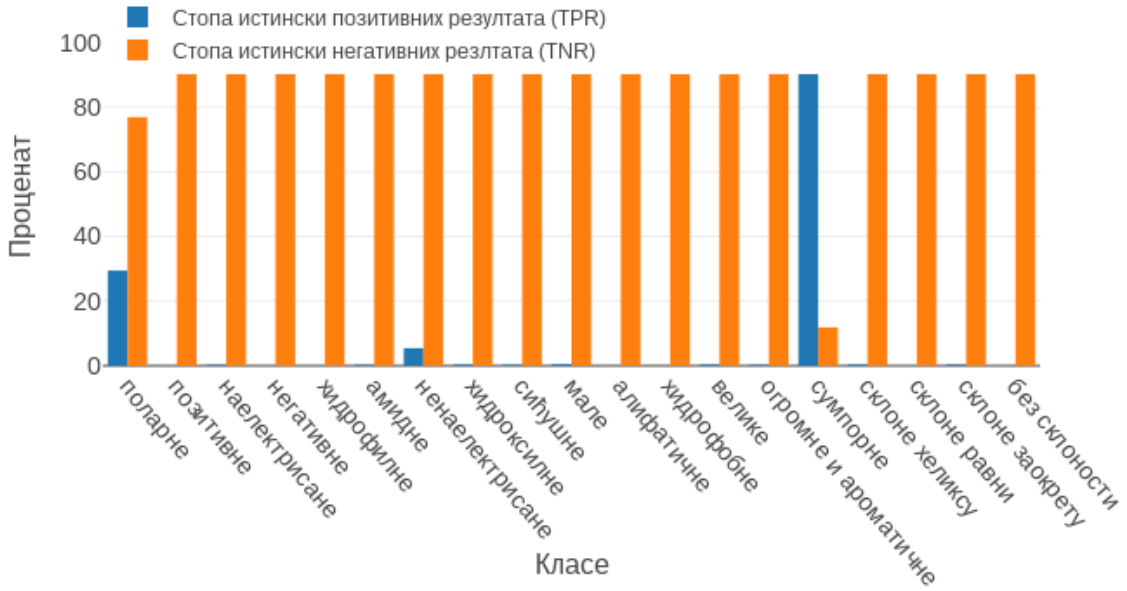
Слика 4.4: Однос вредности TPR и TNR за секундарну структуру B и различите класе



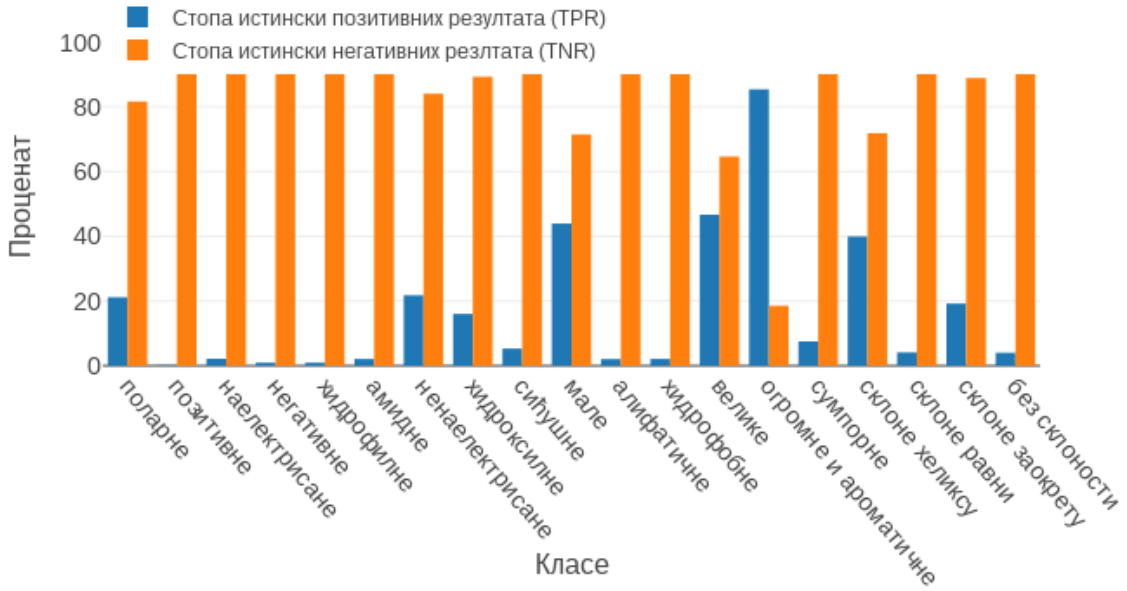
Слика 4.5: Однос вредности TPR и TNR за секундарну структуру E и различите класе



Слика 4.6: Однос вредности TPR и TNR за секундарну структуру T и различите класе

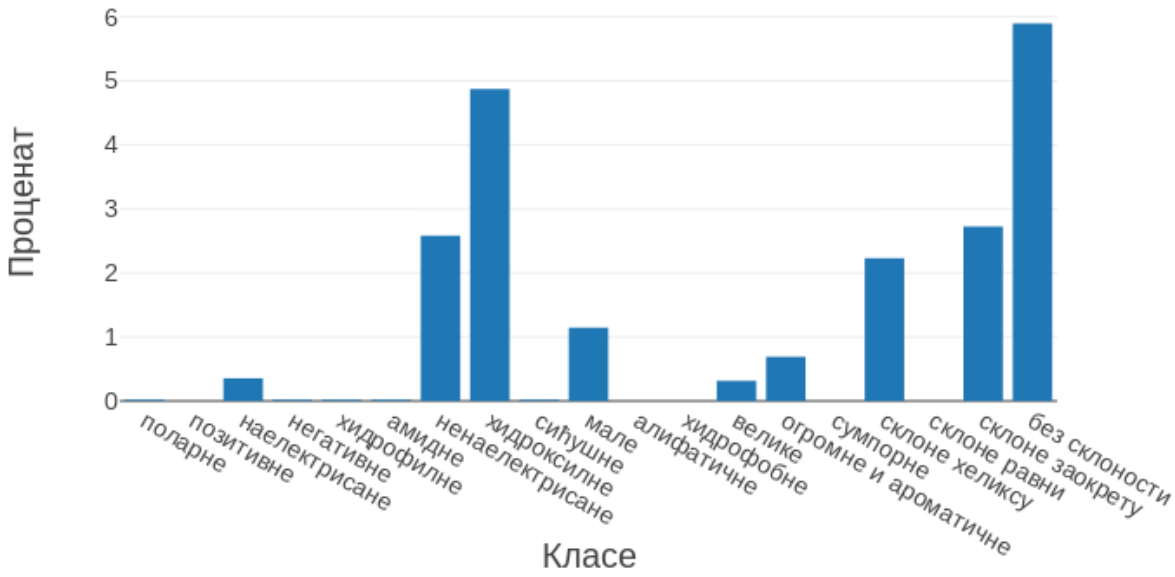


Слика 4.7: Однос вредности TPR и TNR за секундарну структуру S и различите класе



Слика 4.8: Однос вредности TPR и TNR за секундарну структуру C и различите класе

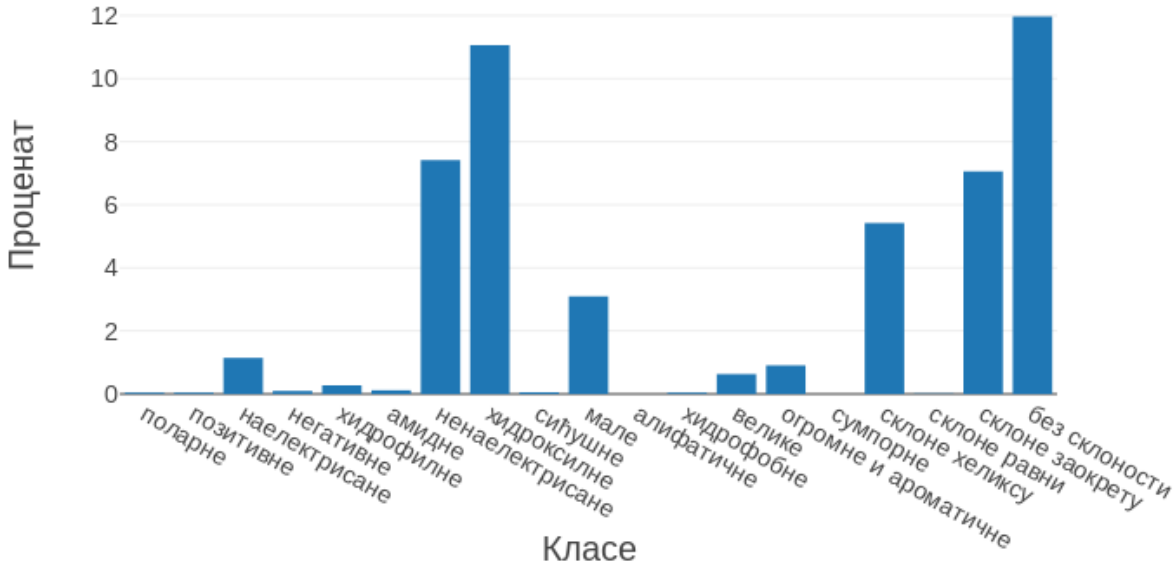
Анализа понашања методе `predict_secondary_structure_using_class` извршена је и у односу на секвенце из тест скупа. За сваку секвенцу за коју је извршена предикција, пребројане су исправно погођене секундарне структуре у оквиру ње. Уколико је број погодака већи од неког задатог прага, секвенца се сматра успшено предвиђеном. На графицима 4.9, 4.10 и 4.11 може се видети проценат успешно предвиђених секвенци, за различите вредности задатог прага.



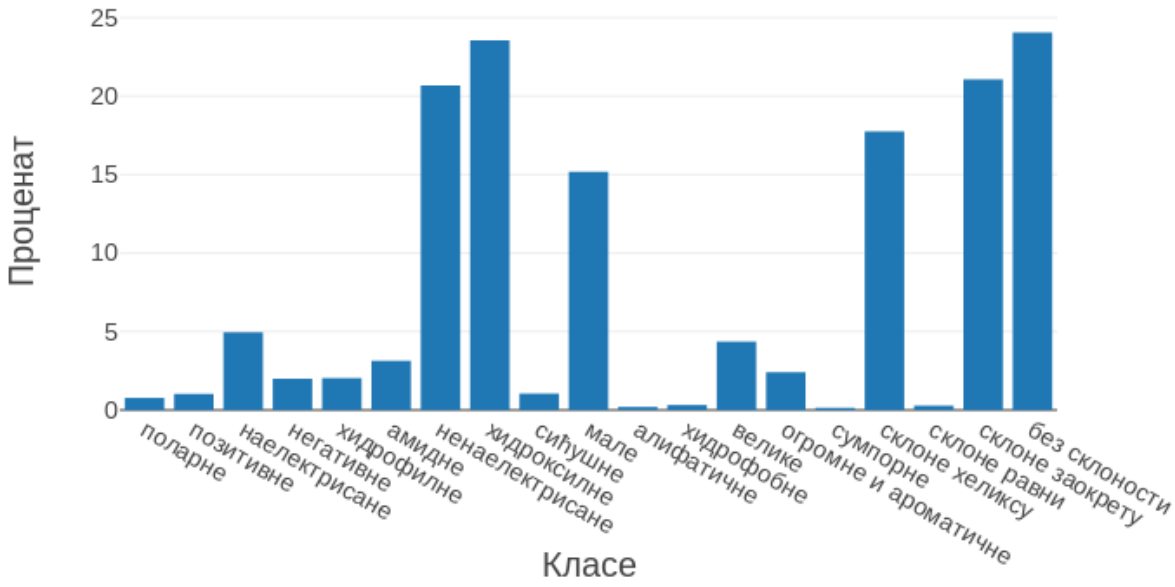
Слика 4.9: Проценат тачно предвиђених секвенци из тест скупа у односу на различите класе (праг је 75%)

Са 4.9 видимо да, без обзира у односу на коју класу радимо предикцију, проценат секвенци из тест скупа, у којима је број исправно погођених секундарних структура изнад 75%, не прелази 6%.

Ако спустимо праг на 65%, проценат исправно предвиђених секвенци се готово дуплира за сваку класу (график 4.10). Последња вредност прага за коју је извршена провера је 50%. Очекивано, резултати се опет побољшавају - сада проценат тачно предвиђених секвенци долази до готово 25%. Предикција у односу на класу аминокиселина без склоности у 24,06% случајева успешно предвиђа секвенце из тест скупа за праг од 50%, а одмах иза ње по успешности су класа хидроксилних (23,55%), затим класа аминокиселина склоних заокрету (21,08%), па класа ненаелектрисаних (20,69%). Класа аминокиселина склоних хеликсу има 17,75% успешности, а класа малих 15,19%. Предикција у односу на остале класе не даје никада више од 5% успешности, што се може видети на 4.11.



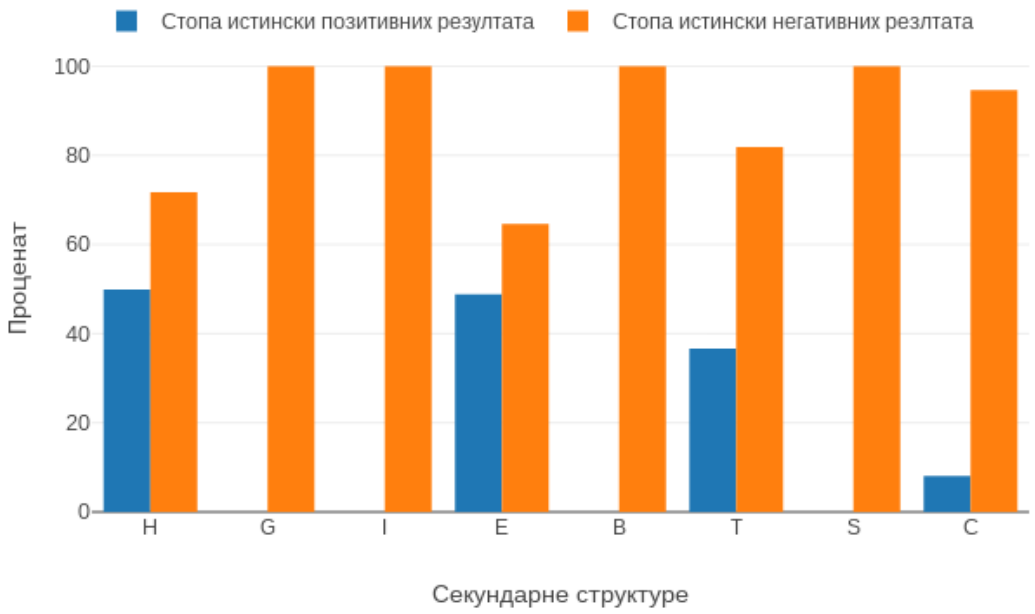
Слика 4.10: Проценат тачно предвиђених секвенци из тест скупа у односу на различите класе (праг је 65%)



Слика 4.11: Проценат тачно предвиђених секвенци из тест скупа у односу на различите класе (праг је 50%)

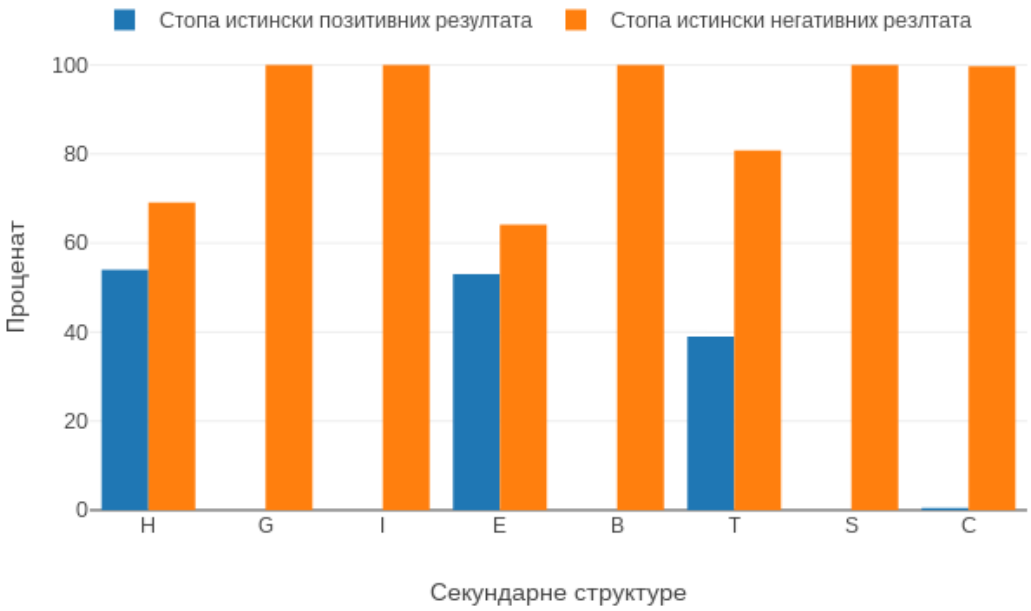
4.3 Предикција у односу на све класе

Резултати које метода `predict_secondary_structure_using_all_classes` постиже, испитани су над истим скупом као за `predict_secondary_structure_using_class`. Као и код `predict_secondary_structure_using_class` и овде су извршене две анализе - једна за појединачне секундарне структуре и друга за секвенце. Оно што је другачије је то што су све анализе извршене збирно, користећи корелације са свим класама.

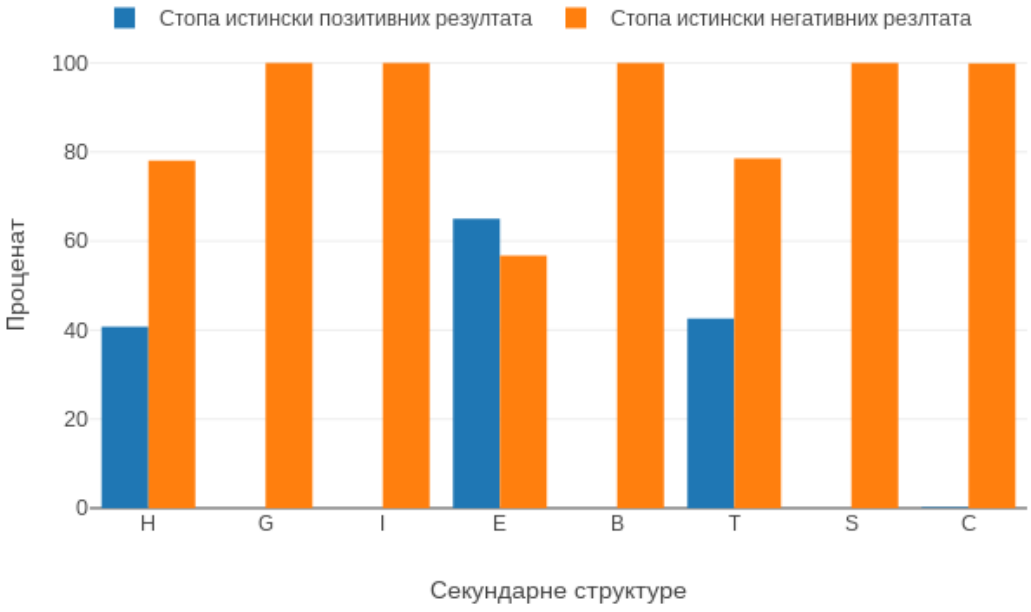


Слика 4.12: Однос вредности TPR и TNR за различите секундарне структуре и `num_of_max = 1`

Са графика 4.12 видимо да `predict_secondary_structure_using_all_classes` најбоље детектује секундарне структуре H и E. Стопа истински позитивних резултата за H изоси 49,93%, а стопа истински негативних даје 71,69%, док су код E ти проценти 48,84 и 64,59, респективно. Нешто слабији резултати предвиђања добијају се за секундарну структуру T - 36,61% за TPR и 81,84% за TNR. Једина преостала секундарна структура која се у неким случајевима успешно предвиђа је C, али тек са стопом од 8,07% (стопа откривања секундраних структура различитих од C је 94,68%). Све остале секундарне структуре имају 100% за TPR, али 0% за TNR, користећи `num_of_max = 1`.



Слика 4.13: Однос вредности TPR и TNR за различите секундарне структуре и `num_of_max = 2`

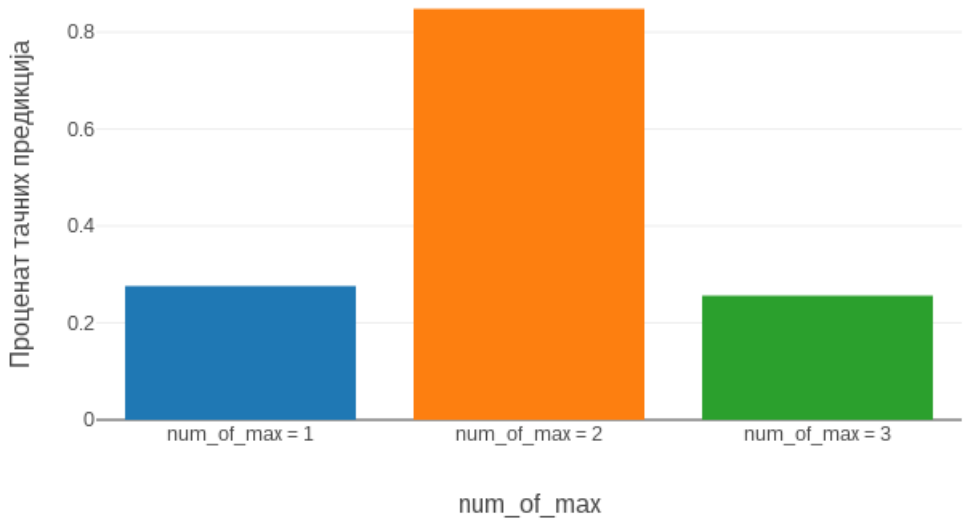


Слика 4.14: Однос вредности TPR и TNR за различите секундарне структуре и `num_of_max = 3`

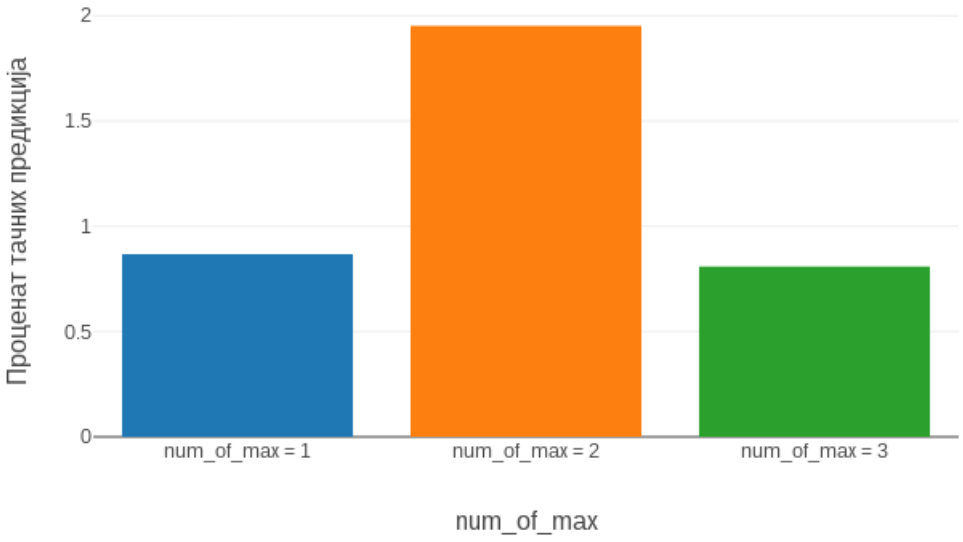
Метода `predict_secondary_structure_using_all_classes` ни за `num_of_max = 2` не успева да тачно предвиди секундарне структуре G, I, B и S. Код све четири, TPR је 0%, а TNR 100%, што је приказано на 4.13. За вредност параметра 2, погоршао се проценат успшености детектовања секундарне структуре C - сада износи само 0,51%, док је TNR 99,72%. За структуре H, E и T, оцена TPR је благо повећана и то на 53,98%, 53,02% и 39,03%, респективно. Одговарајуће вредности за TNR сада износе 69,09%, 64,11% и 80,80%.

Сличну ситуацију имамо и за `num_of_max = 3`. Оцена TPR код H је мања него за `num_of_max = 2` и износи 40,69%, док је TNR повећана на 78,10%. Код секундарне структуре E, TPR је такође повећана и сада износи 65,01%, а TNR смањена на 56,75%. И за T, стопа позитивних резултата је у порасту (42,60%), а стопа негативних у паду (78,59%), у односу на случај када је параметар `num_of_max` имао вредност 2. Као и за претходни случај, секундарне структуре G, I, B и S имају 100% за TNR и 0% за TPR. Лоше резултате имамо и за секундарну структуру C - 0,18% за TPR и 99,93% за TNR. Сви резултати приказани су на графику 4.14.

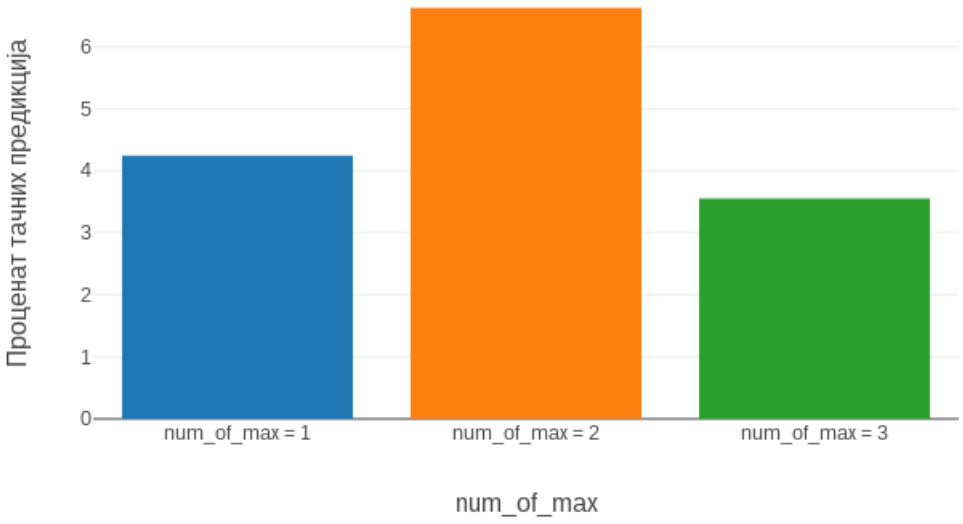
Процена квалитета методе `predict_secondary_structure_using_all_classes` за целе секвенце, приказана је на графицима 4.15, 4.16 и 4.17. На сваком од графика дат је проценат успешно предвиђених секвенци за различите вредности параметра `num_of_max` и различите прагове. Као и за `predict_secondary_structure_using_class`, секвенца се сматра тачно предвиђеном уколико је број погођених секундарних структура у оквиру ње већи од датог прага.



Слика 4.15: Проценат тачно предвиђених секвенци из тест скупа за различите вредности `num_of_max` (праг је 75%)



Слика 4.16: Проценат тачно предвиђених секвенци из тест скупа за различите вредности `num_of_max` (праг је 65%)

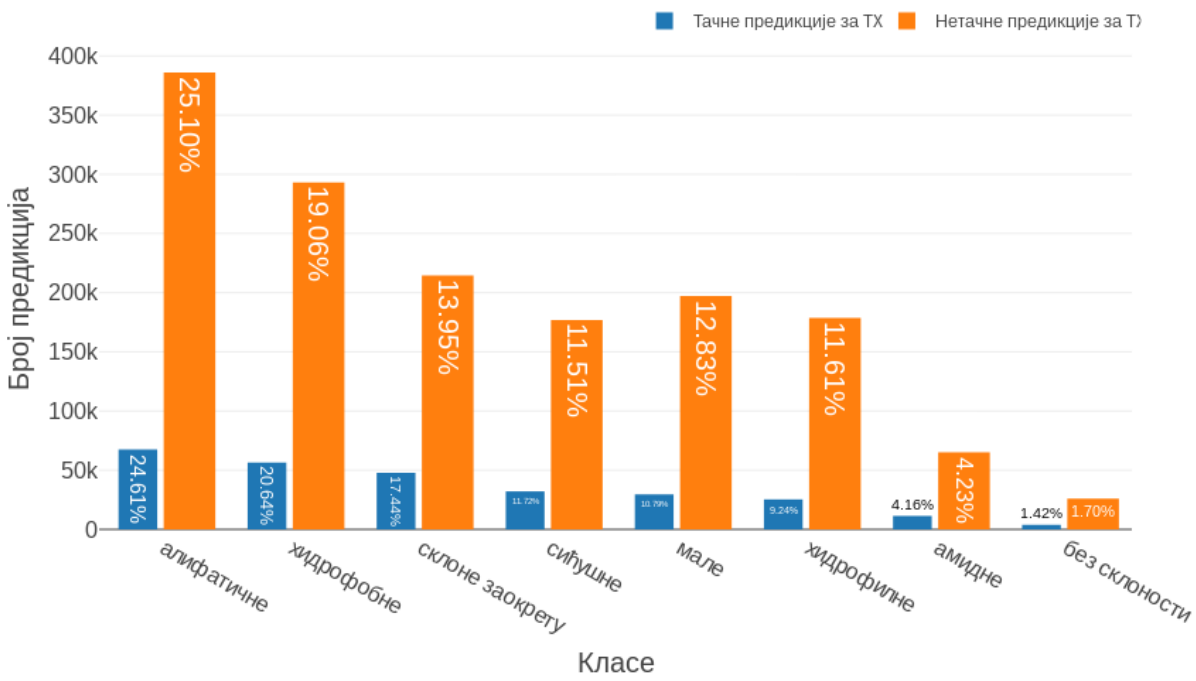


Слика 4.17: Проценат тачно предвиђених секвенци из тест скупа за различите вредности `num_of_max` (праг је 50%)

На 4.15 видимо да ни за једну вредност `num_of_max` и праг од 75%, метода није погодила више од 1% секвенци из тест скупа. Када се праг спусти на 65%, број тачно предвиђених секвенци иде до 2% (график 4.16). Уколико секвенцу сматрамо тачно погођеном када је број погодака у оквиру ње бар 50%, онда метода има успешност од 6,63% тачно погођених секвенци у скупу података и то за `num_of_max = 2` (график 4.17). За све три вредности параметра за које је метода тестирана, резултати су увек најбољи за `num_of_max = 2`, док се за `num_of_max = 1` и `num_of_max = 3` добијају углавном слични резултати и то приближно дупло лошији од оних за `num_of_max = 2`.

4.4 Најбоља класа за сваки груписани биграм

Како је рачунање корелација, а затим и њихово даље коришћење у предикцијама, засновано на груписаним биграмима, познавање резултата које различите класе дају за њих може пружити значај увид за њихово ефективније коришћење.

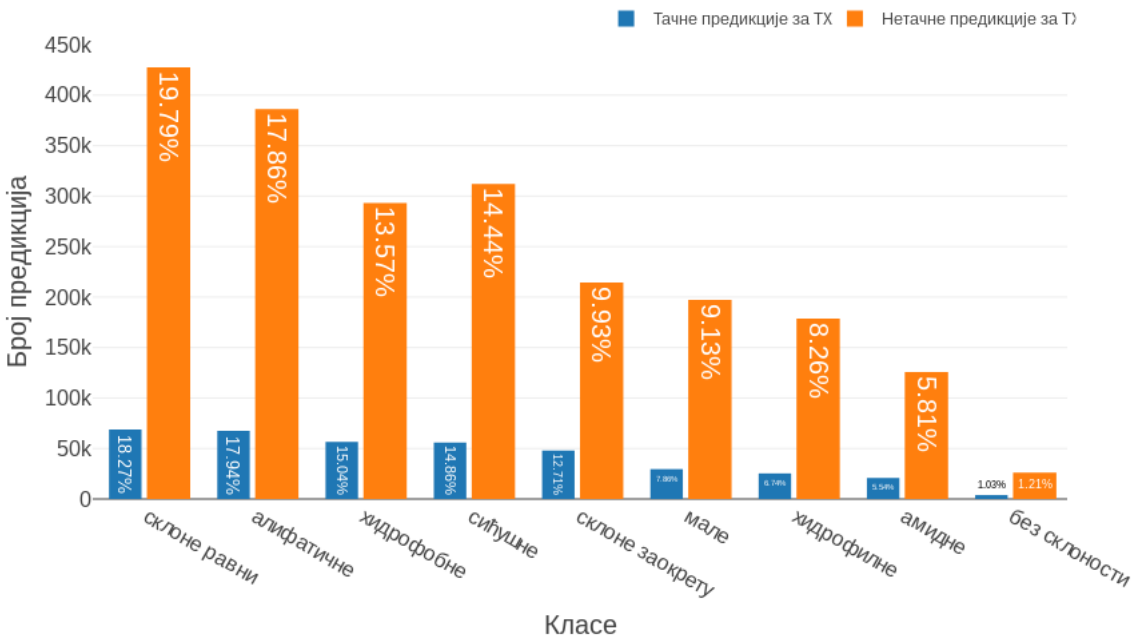


Слика 4.18: Однос тачних и нетачних предвиђања по класама за груписани биграм ТХ и `num_of_max = 1`

На графицима 4.18 и 4.19, може се видети однос тачних и нетачних предвиђања по класама за груписани биграм ТХ, за вредности параметра `num_of_max = 1` и `num_of_max = 2`, користећи податке из тест скупа. Са 4.18 видимо да груписани биграм ТХ најчешће тачно предвиђа класа алифатичне (24,61% од свих успешних предвиђања), док су одмах иза ње по успешности класе хидрофобне (20,64%) и склоне заокрету (17,44%). Нешто мање проценте постижу класе сићушне (11,72%), мале (10,79%), хидрофилне (9,24%), амидне (4,16%) и на крају, класа аминокиселина без склоности (1,42%). Преостале класе ниједном нису ни

коректно, ни некоректно предвиделе ТХ, узимајући увек само највећу корелацију између релације и класе.

Сличан редослед процената остварен је и за погрешна предвиђања - класа алифатичне је највише пута погрешно предвидела ТХ (25,10% од свих неуспешних предвиђања), затим следе хидрофобне (19,06%), склоне заокрету (13,95%), мале (12,83%), хидрофилне (11,61%), сиђушне (11,51%), амидне (4,23%) и класа аминокиселина без склоности (1,70%).

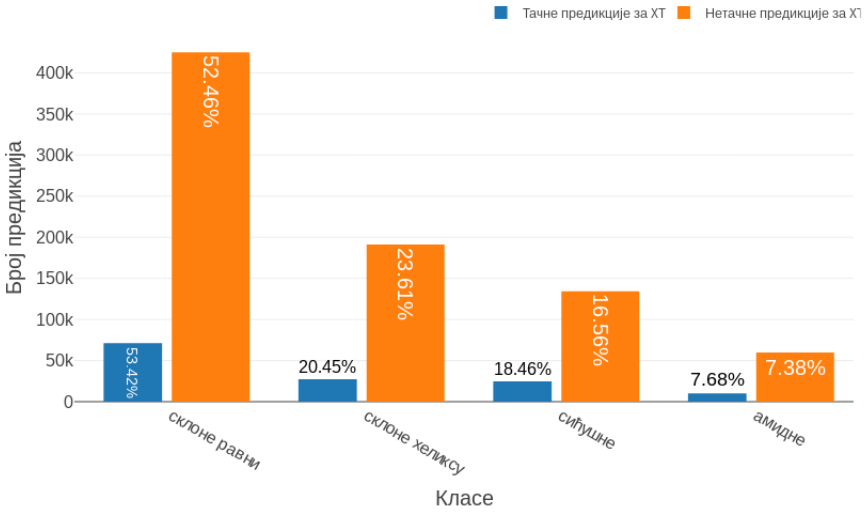


Слика 4.19: Однос тачних и нетачних предвиђања по класама за груписани биграм ТХ и `num_of_max = 2`

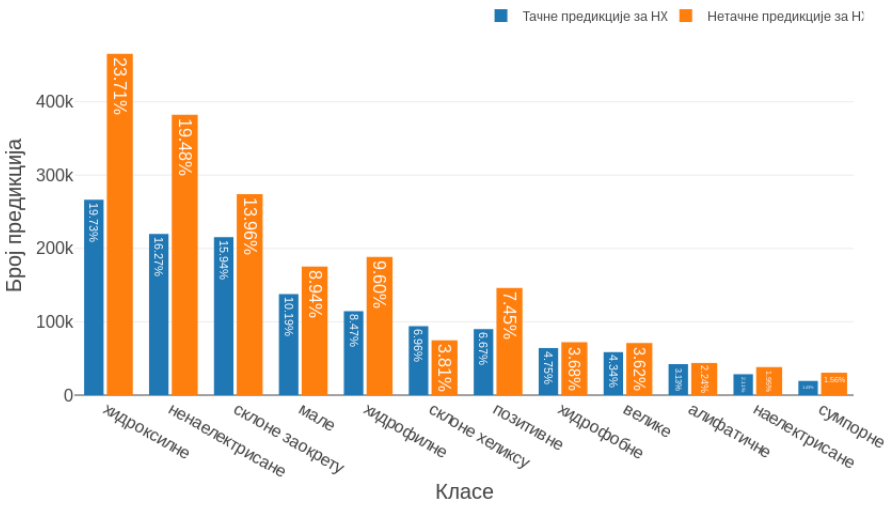
Уколико се у обзир узимају по 2 највеће корелације између релација и класа, резултати су другачији, што се може видети са графика 4.19. Сада највише тачних предвиђања за биграм ТХ остварује класа аминокиселина склоних равни (18,27% од свих успешних предвиђања), док за њом следе алифатичне (17,94%), хидрофобне (15,04%), сиђушне (14,86%), склоне заокрету (12,71%), мале (7,86%), хидрофилне (6,74%), амидне (5,54%) и на крају класа аминокиселина без склоности (1,03%).

Класа аминокиселина склоних равни и алифатичне дају највише нетачних предвиђања (19,79% и 17,86%, респективно), док за њима следе сиђушне (14,44%) и хидрофобне (13,57%). Нешто мање проценте даје класа аминокиселина склоних заокрету (9,93%), затим класа малих (9,13%), хидрофилних (8,26%), амидних (5,81%) и класа аминокиселина без склоности (1,21%).

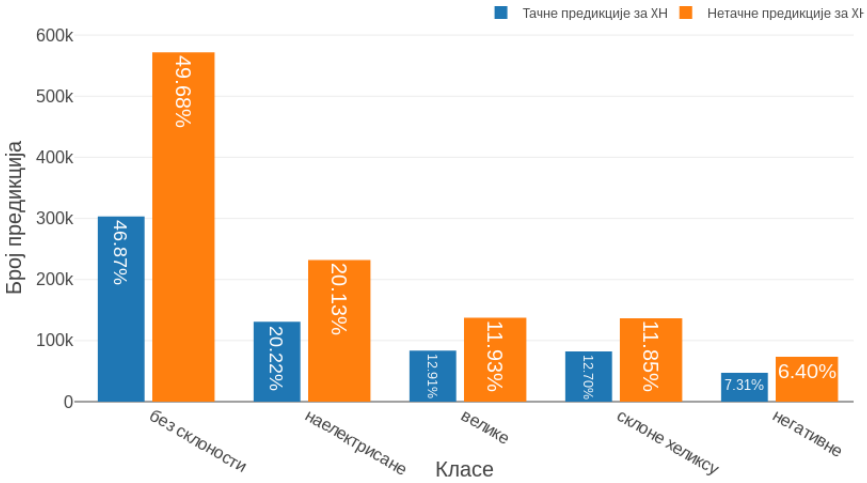
На наредним графицима, приказани су резултати предвиђања осталих груписаних биграма секундарне структуре, када је `num_of_max = 1`. Груписани биграми GX, IX, XI, BX, XB, никада нису предвиђени ни тачно, ни нетачно, користећи вредност параметра 1, па су ти графици изостављени.



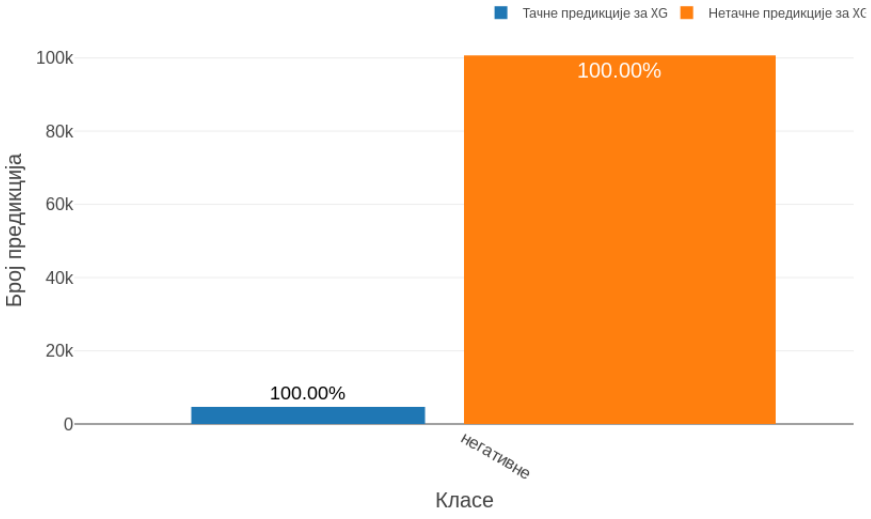
Слика 4.20: Однос тачних и нетачних предвиђања по класама за груписани биграм XT и `num_of_max = 1`



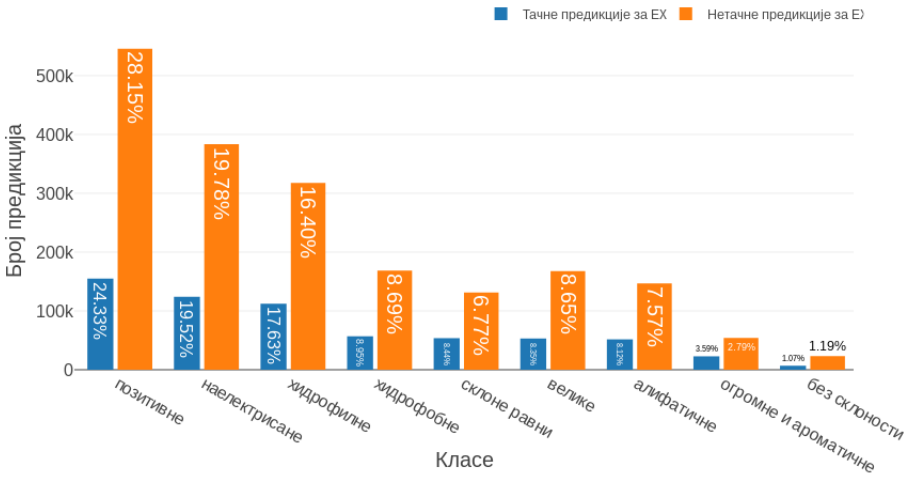
Слика 4.21: Однос тачних и нетачних предвиђања по класама за груписани биграм HX и `num_of_max = 1`



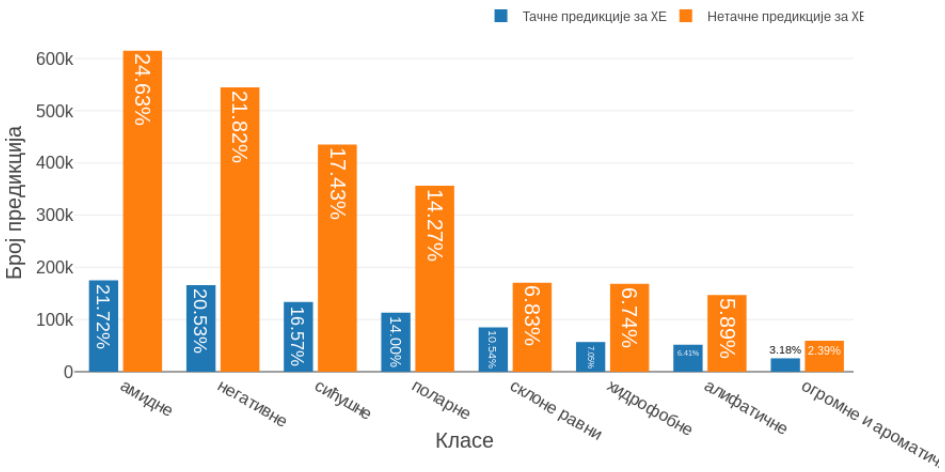
Слика 4.22: Однос тачних и нетачних предвиђања по класама за груписани биграм XH и `num_of_max = 1`



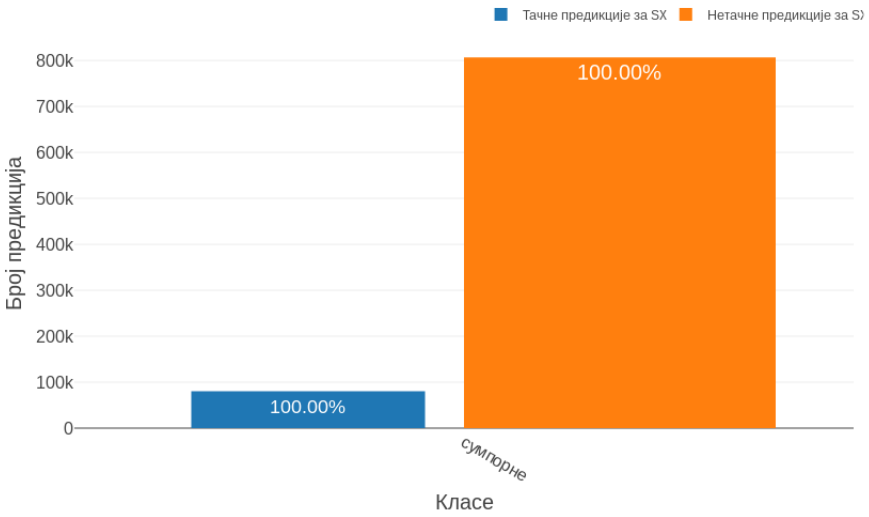
Слика 4.23: Однос тачних и нетачних предвиђања по класама за груписани биграм XG и `num_of_max = 1`



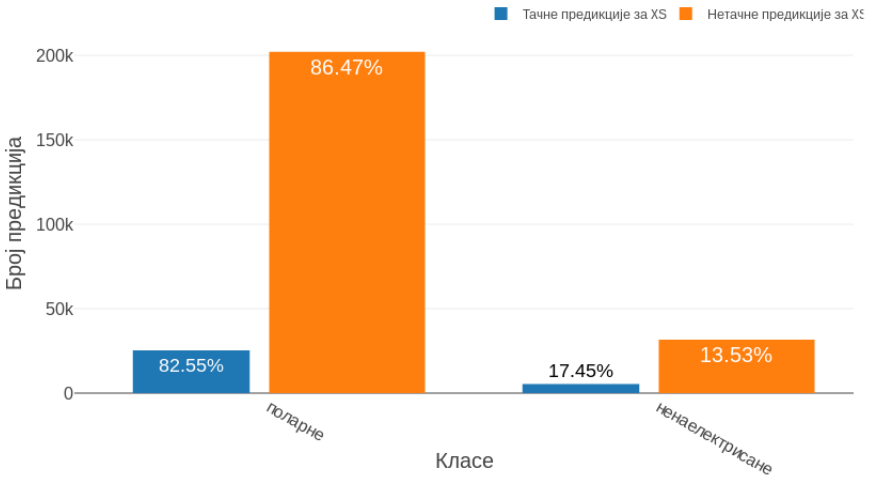
Слика 4.24: Однос тачних и нетачних предвиђања по класама за груписани биграм EX и `num_of_max = 1`



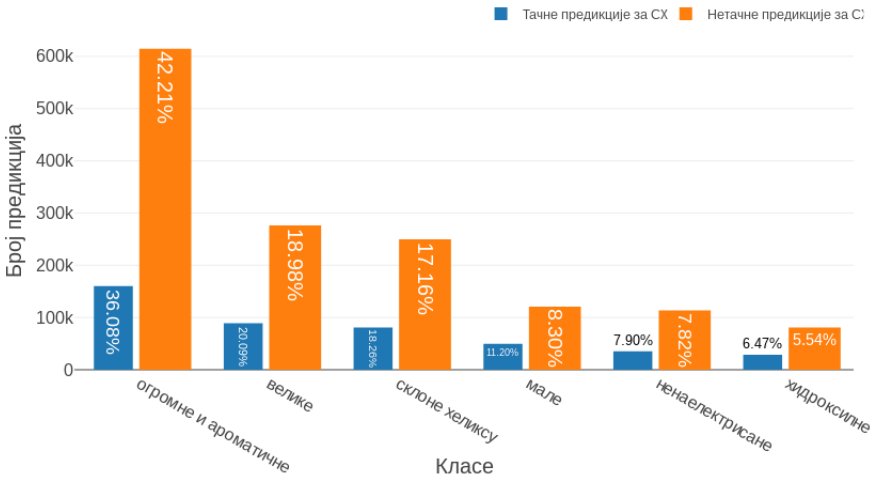
Слика 4.25: Однос тачних и нетачних предвиђања по класама за груписани биграм XE и `num_of_max = 1`



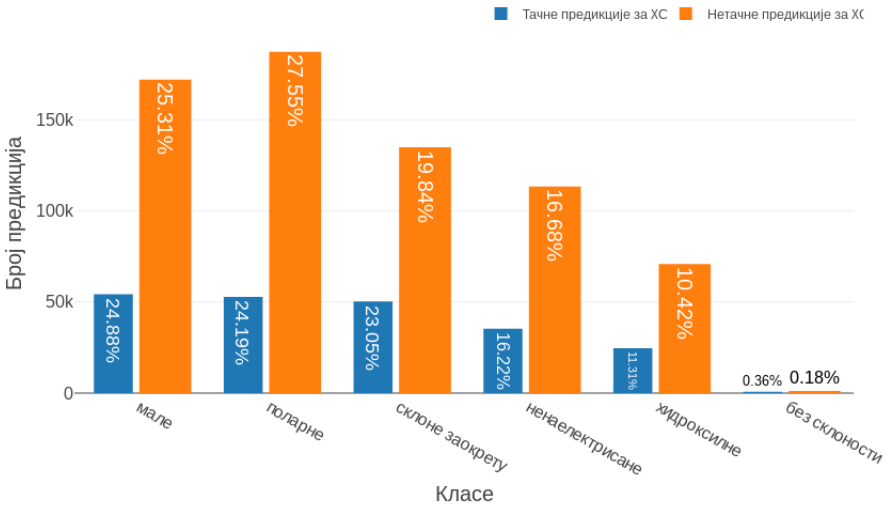
Слика 4.26: Однос тачних и нетачних предвиђања по класама за груписани биграм SX и `num_of_max = 1`



Слика 4.27: Однос тачних и нетачних предвиђања по класама за груписани биграм XS и `num_of_max = 1`



Слика 4.28: Однос тачних и нетачних предвиђања по класама за груписани биграм CX и `num_of_max = 1`



Слика 4.29: Однос тачних и нетачних предвиђања по класама за груписани биграм XC и `num_of_max = 1`

5 Закључак

Разумевање улоге протеина у свакодневном животу, значајно је из више аспеката, међу којима су најбитније примене у медицини и фармацији, али и у областима као што су прехрамбена индустрија, кућна хемија, козметика и сл. Како је за успешно разумевање и коришћење различитих својстава протеина неопходно познавање њихове улоге, намеће се потреба за анализом структуре протеина, која би нам дала информације о тој улози. Виши нивои структуре носе више информација о функцији протеина, па је предвиђање ових, виших нивоа, од кључног значаја, а први корак у томе представља предвиђање секундарне структуре на основу примарне.

У овом раду, представљене су различите методе, које имају за циљ да предвиде секундарну структуру протеина, посматрајући односе које дата примарна структура има са различитим класама. Уместо рачунања вероватноћа, као функција оцене коришћене су корелација и то управо корелације између секундарне структуре и односа који примарна структура има са сваком од класа. Различите методе, на различите начине користе ове корелације, па дају и различите резултате. У поглављу 4 дат је приказ добијених резултата и њихово поређење.

На основу добијених статистика, закључује се да израчунате корелације могу бити значајан параметар у процесу предвиђања секундарне структуре, али да, коришћене на овај начин, саме нису довољне за потпуно поуздане резултате. Наиме, алгоритми у својим предвиђањима увек форсирају секундарне структуре које су постигле већу корелацију, односно имају већу функцију оцене, па се на тај начин потискују друге структуре, које су најчешће, мање заступљене у коришћеном узорку. Овај проблем је, у извесној мери, умањен параметризацијом коришћених метода, с циљем да се ослаби значај секундарних структура са већим оценама и тако да прилика осталима. Међутим, ни то није у потпуности отклонило проблем. Управо због тога, постоји још доста простора за унапређење ових алгоритама, највише у смеру који би подразумевао другачији приступ мање заступљеним секундарним структурама. Уколико би се побољшале методе које на другачији начин процењују изгледе за њихово појављивање, потенцијално на основу неких потпуно других параметара, то би у значајној мери могло да побољша општи квалитет и поузданост алгоритама.

Литература

[1] Jeremy M. Berg, John L. Tymoczko, and Lubert Stryer. *Biochemistry.* W.H.Freeman & Co Ltd, 5 edition, 2002.

[2] Engelbert Buxbaum. *Fundamentals of Protein Structure and Function.* Springer, 2 edition, 2015.

[3] David Clark. *Molecular Biology.* Elsevier Academic Press, 2005.

[4] Valeria De Fonzo, Filippo Aluffi-Pentini, and Valerio Parisi. Hidden Markov Models in Bioinformatics. 200, 2007.

[5] Oxford Dictionaries. English Oxford Dictionaries. on-line at: `https://en.oxforddictionaries.com/definition/us/markov_chain`.

[6] Rodolfo Esquivel, Moyocoyani Molina-Espiritu, Frank Salas, Catalina Soriano-Correa, Carolina Barrientos, Jesus Dehesa, and Jose Dobado. *Decoding the Building Blocks of Life from the Perspective of Quantum Information.* 2013.

[7] Wolfgang Kabsch and Christian Sander. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. 22, 1983.

[8] Mark Lutz. *Learning Python.* O'Reilly Media, Inc., 5 edition, 2013.

[9] Mark Lutz. *Python Pocket Reference.* O'Reilly Media, Inc., 5 edition, 2014.

[10] The National Institute of Open Schooling. Proteins. on-line at: `http://www.nios.ac.in/media/documents/dmlt/Biochemistry/Lesson-04.pdf`.

[11] The RCSB PDB. The Protein Data Bank. on-line at: `https://www.rcsb.org/`.

[12] Ian M. Rosenberg. *Protein Analysis and Purification.* Birkhäuser Basel, 2 edition, 2005.

[13] Parasuraman S. Protein data bank. *J Pharmacol Pharmacother*, 3(4), 2012.

[14] Burley S.K., Berman H.M., Kleywegt G.J., Markley J.L., Nakamura H., and Velankar S. *Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive.* Humana Press, 2017.

[15] Mark Stamp. A Revealing Introduction to Hidden Markov Models. 2015.

[16] David M. Lane Rice University. Online Statistics Education: A Multimedia Course of Study. on-line at: `http://onlinestatbook.com/`.

[17] Matthijs J. Warrens. *Similarity Coefficients for Binary Data: properties of coefficients, coefficient matrices, multi-way metrics and multivariate coefficients*. PhD thesis, 2008.

[18] Robert F. Weaver. *Molecular Biology*. McGraw-Hill, 5 edition, 2012.

[19] Dr Dubravka Štajner and Dr Slavko Kevrešan. *Hemija*. Poljoprivredni fakultet Novi Sad, 2014.