Универзитет “Свети Кирил и Методиј” во Скопје

Факултет за информатички науки и компјутерско инженерство, Скопје, Р. Македонија

МЕТОДИ ЗА ДЕТЕКЦИЈА И АНАЛИЗА НА СВРЗНИТЕ ДЕЛОВИ ОД ПРОТЕИНСКИТЕ СТРУКТУРИ И НИВНА ПРИМЕНА ЗА ОДРЕДУВАЊЕ НА ФУНКЦИИ НА ПРОТЕИНИ

-докторска дисертација-

Ментор:

вон. проф. д-р Андреа Кулаков

Кандидат:

м-р Георгина Мирчева

Скопје, 2014

Апстракт

Знаењето за функциите на протеинските молекули е многу важно за разбирање и регулирање на процесите во живите организми, па затоа целта на оваа докторска дисертација е да се развијат методи за одредување на функциите на протеинските структури. Во оваа дисертација, презентирани се четири методи за одредување на структурната сличност помеѓу протеинските структури. Направена е детална анализа на перформансите на предложените методи за пребарување на протеински структури, а дополнително методите се споредени со неколку постоечки методи за порамнување на протеински структури. Со овие методи се одредува структурната сличност помеѓу протеинските структури, што може да се примени при класификација на протеини, како и при функционално анотирање на протеински структури. Во оваа докторска дисертација е предложен нов метод за предвидување на функциите на протеините кој своите одлуки ги базира на анотациите на најблиските соседи на испитуваниот примерок. Анотирањето на протеинските структури, покрај преку структурно порамнување, може да се направи и преку анализа на карактеристиките на сврзните региони каде што испитуваната протеинска структура стапува во интеракција со друга структура. Во оваа дисертација се предложени неколку методи за детекција на сврзните делови од протеинската структури. При тоа предвид се земени методи кои се базираат на класичната теорија на множества, како и методи кои се базираат на непрецизираната логика. Со цел да се подобри предиктивната моќ на моделите, изградени се и ансамбли, а исто така направени се и анализи во кои со користење на техники за избор и трансформација на карактеристики е направена селекција на најрелевантните карактеристики на амино-киселинските остатоци. Предложените методи за одредување на сврзните региони се споредени со неколку постоечки методи кои се користат за оваа намена. Потоа, предложен е нов метод за одредување на протеинските функции кој се базира на локалните карактеристики на сврзниот регион, како и на глобалните карактеристики на структурата. При градењето на моделите предвид се земаат неколку методи за повеќезначна класификација. На крај, направена е споредба на двата предложени методи за анотирање на протеински структури.

Abstract

The knowledge about the functions of the protein molecules is very important in order to understand and regulate the processes in living organisms. Therefore, the aim of this PhD thesis is to develop new methods for determining the functions of the protein structures. In this thesis, four methods for determining the structural similarity between protein molecules are presented. Also, the performances of these methods are analyzed in details, and additionally these methods are compared with several existing methods for aligning protein structures. By using these methods, the structural similarity between the comparing protein structures could be determined, and they could be used for classifying and annotating protein structures. In this theses, a novel method for protein function prediction is proposed, where the decisions about the annotations of the inspected structure are based on the annotations of its nearest neighbors. Besides annotation based on structural alignment, the protein function prediction could be made by analysis of the characteristics of the binding sites, which are the regions where the inspected structure interacts with another structure. In this thesis several methods for protein binding sites detection are proposed. Besides the methods based on the classical theory of sets, also the methods based on the fuzzy set theory are taken into consideration. In order to improve the prediction power of the models, ensembles are induced, and also by using feature selection and transformation techniques the most relevant characteristics of the amino acids residues are determined in order to find out which features should be considered in the model induction. The proposed methods for protein binding sites prediction are compared with several existing methods used for this purpose. In this thesis, a novel method for protein function prediction is proposed that is based on the local characteristics of the binding sites, as well as the global characteristics of the protein structure. The model induction is made by using several methods for multi-label classification. Finally, a detailed comparison of the two proposed methods for determining protein functions is performed.

СОДРЖИНА

1

ВОВЕД

Протеините се макромолекули кои ги синтетизираат живите клетки, и истите се изградени од мономерни единици наречени амино-киселини. Во градбата на протеините учествуваат 20 амино-киселини кои се разликуваат според градбата на остатокот (R-групата). Алфа карбоксилната група од една амино-киселина преку ковалентна пептидна врска се поврзува со алфа карбоксилната група од друга амино-киселина, и на тој начин амино-киселините се поврзани и градат протеинска нишка (ланец).

Во молекулите на протеините има четири нивоа на организација: примарна, секундарна, терциерна и кватернерна структура. *Примарната структура* на протеините, која уште се нарекува и секвенца, се дефинира преку редоследот на амино-киселините. *Секундарната структура* се однесува на локалните извиткувања на протеинскиот ланец во тродимензионалниот простор, кои настануваат како резултат на водородните врски помеѓу амино-киселинските остатоци кои се блиску во просторот. Најстабилни облици на секундарната структура се α-хеликсите и β-рамнините. *Терциерната структура* се однесува на тродимензионалната поставеност на протеинскиот ланец во тродимензионалниот простор. *Кватернерната структура* ја опишува просторната поставеност на протеинските молекули кои имаат повеќе полипептидни ланци. Кватернерната структура го опишува начинот на кој што овие протеинските ланци се поврзуваат во протеинската молекула.

Протеините се едни од основните градбени компоненти на клетките во живите организми при што вршат различни витални функции, како на пример транспорт (пр. хемоглобинот пренесува кислород и јаглерод диоксид преку крвта), катализа (протеини кои се ензими и учествуваат во биокатализирачки процеси), заштита (пр. антителата се имуноглобулини), регулација и контрола во организмот (голем број од хормоните имаат протеинска природа), мускулна контракција (пр. актин, миозин) и многу други. Познавањето на функциите на протеинските молекули е од особено значење бидејќи ова знаење може да се примени при дизајн на лекарства со цел да се овозможат или оневозможат различни процеси во живите организми.

При еволуција настануваат одредени промени кај протеинските структури, но сепак може да се забележи значителна сличност помеѓу протеинските структури кои имаат ист предок. Протеините кои имаат заеднички предок делат исти функции, па врз основа на тоа може да се утврди дека делот од протеинската структура кој останува непроменет е делот кој ја одредува функцијата на протеинската структура. Со цел да се добие што е можно повеќе знаење за улогата на протеините во процесите во живите организми, човекот развил повеќе техники со кои се откриваат нови протеински структури. Овие структури по соодветна верификација се складираат во најголемата база на протеински структури Protein Data Bank (PDB) (1), (2). Оваа база е креирана во 1971 година во Brookhaven National Laboratories во САД, и тогаш содржела седум структури, а до 4 февруари 2014 година во неа се сместени 97591 структура. PDB базата содржи податоци за примарната, секундарната и терциерната структура на протеините. Врз основа на овие податоци, биолозите по експериментален пат ги одредуваат функциите на протеинските молекули. Сепак, мануелното одредување на функцијата на протеинските структури е скапа и временски обемна процедура. Со развојот на техниките за одредување на протеинските структури, сè повеќе расте бројот на новооткриени протеински структури чии функции сè уште не се познати. Па затоа сè поатрактивни и понеопходни се компјутерските методи за одредување на протеинските функции.

Голем број на истражувачки групи работат во областа на одредување на протеинските функции, па затоа неопходно било да се дефинира стандард за унифицирана репрезентација на откриеното знаење за протеинските функционални анотации. За таа цел воведена е онтологијата Gene Ontology (GO) (3), и таа претставува структуриран и контролиран речник на термини за протеинските функции. Овие термини се поделени во три групи: молекуларни функции, биолошки процеси и клеточна локација. Во GO за секоја одредена анотација се чува код за евиденција (evidence code) кој покажува на кој начин е одредена дадената анотација. Во (4) направена е анализа на анотациите во GO откриени до април 2010 година. На сличен начин, во овој докторски труд направена е анализа на анотациите во GO достапни на 12 јули 2013 година. Во Табела 1.1 прикажан е бројот на анотации во секоја од категориите на кодови за евиденција.

Може да се забележи дека само 3.53% од анотациите се одредени по експериментален пат, а дури 95.15% се одредени со користење на компјутерски методи, а сè уште не се потврдени од човек. Од тука очигледна е потребата за развој на прецизни и брзи компјутерски методи за одредување на протеинските функции.

Категорија	Код за евиденција	Број на анотации	Процент (%)
Curator Statement Evidence Codes	(IC, ND)	3256	0.27
Author Statement Evidence Codes	(TAS, NAS)	3748	0.31
Computational Analysis Evidence Codes	(ISS, ISO, ISA, ISM, IGC, IBA, IBD, IKR, IRD, RCA)	8927	0.74
Experimental Evidence Codes	(EXP, IDA, IPI, IMP, IGI, IEP)	42866	3.53
Automatically-assigned Evidence Codes	(IEA)	1154206	95.15

Табела 1.1 Број на протеински анотации за различни категории на кодови за евиденција.

Во литературата постои широк спектар на компјутерски методи за анотирање на протеинските структури. Според типот на информација што ја земаат предвид, методите генерално може да се поделат во четири главни групи. Првата група на методи се фокусира на одредување на сличноста на протеинските секвенци, и истите се базираат на претпоставката дека протеинските молекули кои имаат слична секвенца е поверојатно да имаат и слични функции. За оваа намена може да се користат различни методи за порамнување на секвенци (5), (6), (7), (8), како на пример Basic Local Alignment Search Tool (BLAST) методот (9). Со користење на оваа група методи генерирани се податоци за повеќе бази на податоци, како PROSITE (10), (11), Pfam (12), PRINTS (13), SMART (14), CDD (15), PRODOM (16) и други.

Методите во втората група ја испитуваат структурната сличност помеѓу протеините, бидејќи структурата содржи значително повеќе информации за протеинската молекула. Уште повеќе, постојат протеински структури кои имаат значително различна секвенца, а имаат слична структура (17). Постојат бројни методи за порамнување на протеински структури, и тие се базираат на претпоставката дека деловите од протеинската структура кои ги одредуваат функциите на протеинот се стабилни и не се менуваат во текот на еволуцијата. Во оваа група на методи спаѓаат методите: SSAP (Secondary Structure Alignment Program) (18), DALI (Distance Alignment) (19), CE (Combinatorial Extension) (20), MAMMOTH (21), MAMMOTH-mult (22), MUSTA (Multiple Structure Alignment Algorithm) (23) и други. Постојат и методи кои прават комбинација од порамнување на протеинска секвенца и структура (24) со цел да овозможат

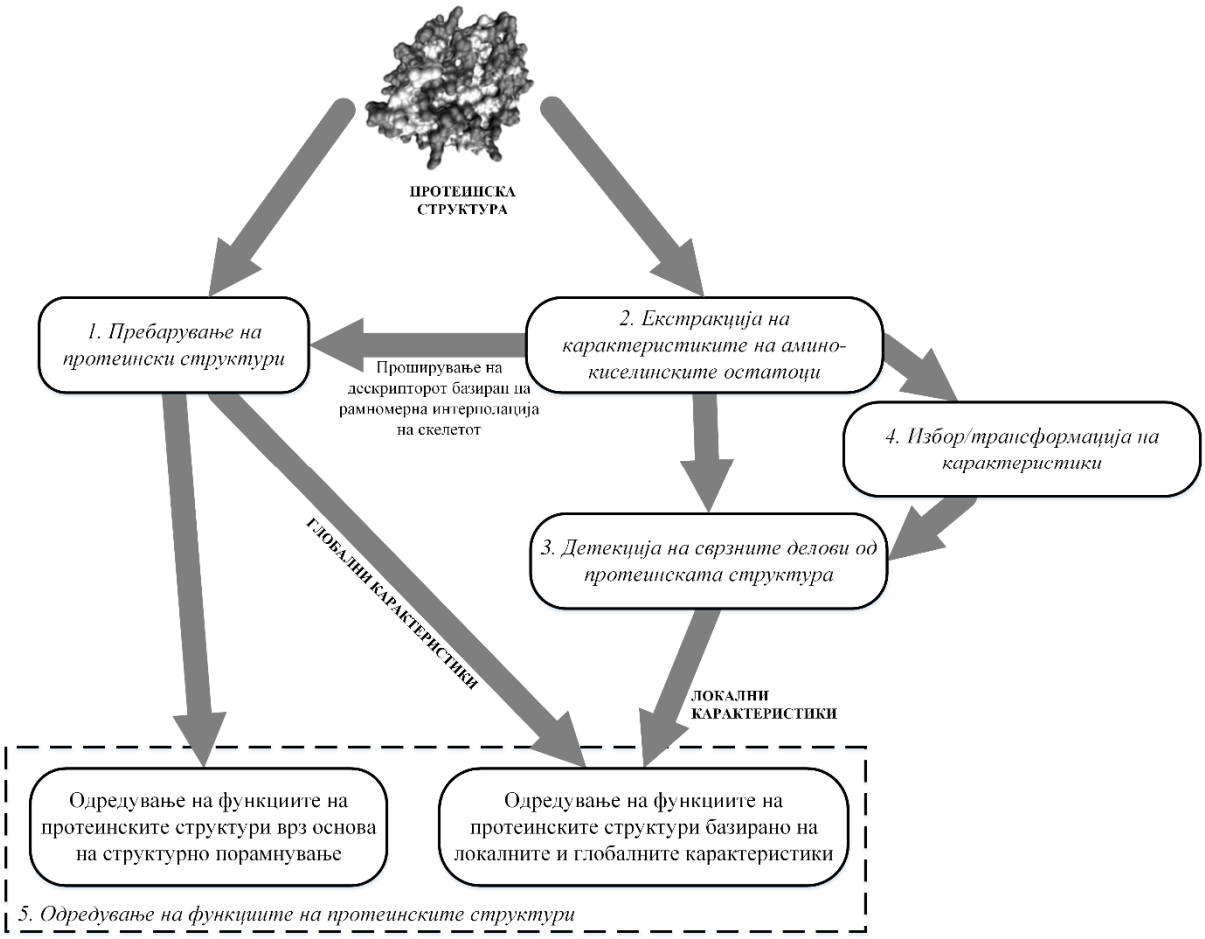
попрецизно одредување на хомологни протеини. Сепак, некои новооткриени протеини може да немаат доволно висока сличност со протеините чии функции се веќе познати, па затоа често има потреба од друг пристап кој не се базира на испитување на хомологноста на протеините.

Третата група методи врши анотација на протеинските структури преку детекција и анализа на сврзните делови од протеинската структура. При детекција на сврзните делови каде што настанува интеракција помеѓу протеините, предвид може да се земат различни својства на протеинската структура. Амино-киселинските остатоци кои се дел од сврзен регион имаат висока конзервација. Оваа стабилност се должи на хидрогенските врски, Ван дер Валсовите сили, електростатските интеракции и хифрофобичните интеракции (25). Методите за одредување на сврзните делови од протеинската структура предвид земаат различни карактеристики како: големината на сврзните региони (26), (27), (28), растојанието помеѓу сврзните региони каде настанува интеракција помеѓу две протеински структури (29), (30), (31), комплементарноста на структурите (32), електростатските и хемиските карактеристики на протеините (33), (34), (35). Во (36) и (37) е даден преглед на постојните алатки и веб сервери за предвидување на сврзните делови од протеинската структура, како и на постоечките бази со протеински интерфејси.

Во четвртата група спаѓаат методите кои вршат анотација на протеинските структури преку анализа на протеинските интеракциски мрежи (38), (39), (40), (41), (42), (43), (44), (45), (46), (47). Овие методи предвид земаат кои протеински структури стапуваат во интеракција, без притоа да ја анализираат примарната, секундарната и терциерната структура на протеинските структури. Постојат повеќе бази кои содржат податоци за протеинските интеракции, како што се BIND (48), DIP (49), MIPS (50), MINT (51), BioGRID (52) и други. Сепак, откривањето на интеракциите помеѓу протеинските структури по експериментален пат е финансиски скапа процедура, па затоа ваквите протеински интеракциски мрежи не содржат информација за сите протеини помеѓу кои може да настане интеракција.

Во [A2] беше направена анализа на тродимензионалната структура на протеините со цел да се овозможи пребарување и класификација на протеини. Оваа докторска дисертација се фокусира на функционално анотирање на протеинските структури базирајќи се на податоците за тродимензионалната протеинска структура. Продолжувајќи во таа насока, во оваа докторска дисертација фокусот е ставен на методите за анотирање на протеинските структури кои припаѓаат во втората и третата група, односно тоа се методите кои се базирани на испитување на протеинската структура. Во [A2] беа презентирани повеќе алгоритами за екстракција на протеински дескриптори кои носат информација за распореденоста на амино-киселинските остатоци во тродимензионалниот простор, и истите беа применети за пребарување и класификација на

протеините. Во оваа докторска дисертација ќе бидат презентирани резултатите од целокупната анализа на овие методи за пребарување на протеинските терциерни структури и истите ќе бидат споредени со два познати методи, DALI (19) и CE (20). Потоа, ќе биде предложен нов метод за повеќезначна класификација за анотирање на протеинските структури врз основа на анотациите на најблиските соседи на испитуваната протеинска структура. При тоа ќе се користат претходно предложените методи за одредување на сличноста помеѓу протеинските структури. Во однос на методите од третата група, прво ќе бидат предложени повеќе методи за детекција на сврзните делови од протеинската структура врз основа на својствата на амино-киселинските остатоци. По идентификацијата на сврзните региони, потоа користејќи постоечки методи за повеќезначна класификација ќе се врши анотирање на протеинските структури. На Слика 1.1 е даден шематски приказ кој го опишува текот на процесите кои се прават при одредување на протеинските функции.



Слика 1.1 Шематски приказ на процесите кои се прават при одредување на функциите на протеинските структури.

Во поглавје 2 на оваа дисертација ќе биде даден преглед на постоечките методи за пребарување на протеински структури, и детекција на сврзните делови каде што настануваат интеракции помеѓу протеините. Во поглавје 3 ќе биде даден краток опис на методите за пребарување на протеински структури кои се предложени за одредување на структурната сличност помеѓу протеините, и истите ќе бидат споредени со неколку постоечки методи. Дополнително, дескрипторот базиран на рамномерна интерполација на протеинскиот скелет ќе биде проширен со неколку својства на амино-киселинските остатоци со цел да се обезбеди попрецизен дескриптор. Во поглавје 4 ќе бидат воведени неколку методи за детекција на сврзните делови од протеинската структура, и ќе биде направена нивна споредба со неколку познати методи наменети за идентификација на сврзните региони. Дополнително, со користење на техники за избор и трансформација на карактеристики ќе биде направена селекција на најрелевантните карактеристики, со цел да се подобри предиктивната моќ на моделите. Во поглавје 5 ќе биде воведен и евалуиран нов метод за анотирање на протеински структури врз основа на структурно порамнување. При тоа структурната сличност помеѓу протеинските структури се одредува со користење на методите за пребарување на протеински структури презентирани во поглавје 2. Во поглавје 5, исто така ќе биде воведен нов метод за анотирање на протеински структури врз основа на локалните карактеристики на амино-киселинските остатоци кои го формираат сврзниот регион, како и глобалните карактеристики на целата протеинска структура. При тоа предиктивните модели ќе се градат со користење на методи за повеќезначна класификација. Во рамки на истото поглавје ќе биде направена споредба на двата предложени методи за одредување на функциите на протеинските структури. На крај, во поглавје 6 ќе бидат сумирани заклучоците од направените анализи за перформансите на предложените методи, и ќе бидат идентификувани насоки за нивно понатамошно подобрување. На Слика Д1 во Додатокот е даден детален преглед на методите кои се предложени и кои се користат во оваа докторска дисертација.

2

АНАЛИЗА НА ПРОБЛЕМИТЕ ЗА ОБРАБОТКА И ПРЕГЛЕД НА ЛИТЕРАТУРА

Со развивањето на повеќе техники за откривање на структурата на протеините, овозможен е брз пораст на бројот на новооткриени протеини. Проектите поврзани со геномите продуцираат податоци кои потоа се складираат во биолошките бази на податоци. Различни бази на податоци содржат различни типови на информации. Така на пример GenBank (53), EMBL (54) и DDBJ (55) содржат податоци за нуклеотидните секвенци, SWISS-PROT (56), PIR (57) и ENZYME (58) се користат за складирање на протеински секвенци, во PDB (1), (2), и MMDB (59) се чуваат податоци за тродимензионалната структура, а PROSITE (11), PRINTS (60) и BLOCKS (61) содржат податоци за протеинските фамилии. Врз основа на податоците кои се добиени по експериментален пат, потоа биолозите треба да ја одредат функцијата на протеините со што би се дошло до знаење како да се предизвика или спречи даден процес во живите организми. Мануелното анотирање на протеинските структури е скапа и бавна процедура, па како резултат на тоа само 3.53% од анотациите во GO се добиени по експериментален пат. Од тука, очигледна е потребата за развој на брзи и прецизни компјутерски методи за анотирање на протеински структури.

Во литературата се среќаваат различни методи за анотирање на протеински структури кои предвид земаат различни својства на протеинската молекула. Како што беше кажано во

воведното поглавје, методите за функционално анотирање на протеини генерално може да се поделат на:

1) методи кои се базираат на порамнување на протеински секвенци,
2) методи кои се базираат на споредба на протеински структури,
3) методи базирани на детекција и анализа на сврзните делови од протеинската структура и
4) методи кои вршат анализа на протеинските интеракциски мрежи.

Првата и втората група на методи вршат анотирање на протеинските структури врз основа на анотациите на протеините чии функции се познати, при тоа земајќи ги предвид протеините кои имаат висока сличност со испитуваниот протеин. Ова се обезбедува со методи за пребарување на хомологните протеини. За одредување на хомологни протеини може да се користат методите кои користат техники од динамичко програмирање (19), (20), (62). Овие методи се временски захтевни, па затоа кај FASTA (63) и BLAST (9) методите се прави апроксимација на Smith-Waterman алгоритамот. BLAST методот врши порамнување на протеински секвенци, но не ја достигнува предиктивната моќ на динамичкото програмирање, а од друга страна е побрз. Досега предложени се повеќе методи кои претставуваат дополнување на BLAST методот со цел да се зголеми брзината и точноста на методот (64), (65), (66). Сепак, постојат протеински молекули кои имаат значително различни секвенци кои во тродимензионалниот простор се извиткуваат така што формираат слични структури (17). Затоа поголем дел од истражувањата се фокусираат на анализа на информациите содржани во протеинската структура.

2.1. Преглед на методите за пребарување на протеински структури

За да се одреди сличноста помеѓу две протеински структури, се врши екстракција на карактеристиките на споредуваните структури. Подоцна, во фазата на пребарување, споредбата се врши врз основа на претходно извлечените карактеристики на протеински структури. Методите за пребарување на протеински структури може да се поделат на методи кои се базираат на секундарната структура и методи кои се базираат на терциерна структура. Во продолжение ќе биде даден краток преглед на најпознатите методи за пребарување на протеински структури.

SSAP (Secondary Structure Alignment Program) методот (18) користи двонивовско динамичко програмирање (67) за порамнување на две протеински структури. Подоцна, во (68) е предложена модификација на SSAP методот за истовремено порамнување на повеќе протеински структури. Овој модифициран алгоритам (68) се користи за класификација на ниво на фолд во CATH (Class, Architecture, Topology, Homology) методот (69).

DALI (Distance Alignment) (19) методот предвид ја зема матрицата на растојанија која содржи информација за Евклидовото растојание помеѓу секој пар од Cα атоми. Со цел да се поедностави порамнувањето, матрицата на растојанија се декомпонира на контактни делови помеѓу два хексапептидни фрагменти. Потоа се врши порамнување на матриците на растојанија на двата испитувани протеина преку пронаоѓање на слични контактни делови, и на крај со Монте Карло симулација се здружуваат паровите од контактните делови.

Слично како и DALI методот, така и MatAlign методот (70) прави порамнување на матриците на растојанија. MatAlign користи brute force пристап со кој се обидува секоја редица од матрицата на растојанија на едната протеинска структура да ја порамни со секоја редица од матрицата на растојанија на втората протеинска структура.

Постојат методи кои порамнувањето на протеинските структури го прават преку порамнување на нивните контактни мапи. Контактна мапа е граф чии јазли соодветствуваат на Cα атомите од протеинската структура, додека врските во графот покажуваат за кои парови од Cα атоми важи дека се "во контакт", т.е. се на растојание помало од некој предефиниран праг. Сличноста на две протеински структури се одредува како степен на поклопување на контактните мапи на двете протеински структури кои се споредуваат. Постојат повеќе методи кои се обидуваат да го решат овој проблем, од кои повеќето се егзактни и временски скапи. MultiStart VNS (MSVNS) методот (71) е евристички метод за порамнување на две контактни мапи, и истиот се базира на Variable Neighborhood Search (VNS) методот, при што дополнително предвид се земаат динамичките промени во соседството.

CE (Combinatorial Extension) методот (20) врши комбинаторното проширување на порамнувачката патека која е составена од порамнети фрагментни парови. Се врши суперпозиција на обликот на протеините, а потоа средната квадратна девијација се одредува според растојанијата на внатрешните амино-киселински остатоци.

MAMMOTH методот (21) го пресметува URMS растојанието помеѓу сите парови од хепта-пептидите на протеините кои се споредуваат, а потоа ги одредува ротациските матрици за секој пар. Преку URMS растојанието се одредува локалното порамнување за кое има максимална локална сличност во двете протеински структури, а потоа се формира најголемото подмножество од слични локални структури. MAMMOTH методот може да се користи за споредување на две протеински структури. Во (22) е предложен MAMMOTH-mult методот кој е модификација на MAMMOTH методот и овозможува истовремено порамнување на повеќе протеински структури.

MUSTA (Multiple Structure Alignment Algorithm) методот (23) го генерира векторот на трансформации преку порамнување на k амино-киселински остатоци. Потоа се групираат

сличните вектори на трансформации и се врши спојување на *k*-те остатоци во заедничко под-структурно јадро.

Постојат методи кои прво извлекуваат вектор кој ги содржи најрелевантните карактеристики на протеинската структура, таканаречен карактеристичен вектор (дескриптор), а потоа пребарувањето на слични протеински структури го прават преку споредба на овие дескриптори во векторскиот простор користејќи одредена мерка за растојание.

Протеинскиот фрактален дескриптор (72) предвид ја зема волуменската фрактална димензија (3Д самосличноста) на протеинската структура. Сепак фракталната димензија не е доволно дискриминаторна карактеристика за да обезбеди робусна репрезентација на геометриските карактеристики на протеинската структура. Па затоа во овој дескриптор дополнително предвид се зема и радиусот на протеинот кој е еднаков на радиусот на најмалата сфера во која може да се смести испитуваната протеинска структура.

Кај протеинскиот Haar дескриптор предложен во (73), прво матрицата на растојанија се скалира до димензии 128x128, а потоа се применува декомпозиција на бранчиња до четвртото ниво користејќи го Haar бранчето. На крај дескрипторот се формира од 36 апроксимативни коефициенти.

Покрај овие методи постојат уште голем број други методи за одредување на сличноста помеѓу протеинските структури, како FATCAT (74), VAST (75), FAST (76), Matras (77), DaliLite (78), GRATH (79) и други.

Со двата начини на порамнување целта е да се одредат деловите од протеинската секвенца или структура кои имаат висока стабилност (конзервација). Во литературата можат да се најдат и методи кои предвид ја земаат конзервацијата и на протеинската секвенца и на протеинската структура (24) со цел да обезбедат подобра предикција на деловите кои имаат голема стабилност и кои ги одредуваат функциите кои ги врши протеинската молекула во процесите во живите организми.

2.2. Преглед на методите за детекција на сврзните делови од протеинската структура

Методите за анотација на протеински структури кои се базираат на одредување на хомологни протеини може да се применат доколку испитуваниот протеин има висока сличност со протеините чии функции се веќе познати. Но не за сите протеини постојат нивни хомологни протеини кои се веќе функционално анотирани, па затоа често има потреба од друг пристап кој нема да се базира на испитување на секвентната или структурната сличност. Токму затоа методите кои спаѓаат во третата група и кои вршат анотирање на протеинските структури преку

детекција и анализа на сврзните делови од протеинската структура се сè позастапени во литературата.

Сврзните делови од протеинската структура, кои уште се нарекуваат и протеински интерфејси, се региони од протеинскиот ланец каде настанува интеракција помеѓу испитуваниот протеински ланец со некој друг протеински ланец. Протеинските интерфејси имаат голема стабилност (конзервација) која произлегува од хидрогенските врски, електростатските интеракции, Ван Дер Валсовите сили и хидрофобичните интеракции (25). Во литературата постојат различни пристапи за одредување на протеински интерфејси, кои предвид ги земаат следниве карактеристики: површината на регионот каде настанува интеракцијата (26), (27), (28), растојанието помеѓу амино-киселинските остатоци кај кои настанува интеракцијата (29), (30), (31), комплементарност на површините (32) и хемиските карактеристики (33), (34), (35). Во продолжение ќе биде даден преглед на поважните методи за одредување на сврзните делови од протеинската структура.

Jones и Thornton (28) разгледувале 59 интеракции и истите ги групирале во четири категории: хомодимери, ензимско-инхибиторни комплекси, антитела и хетеро-комплекси. При тоа за секој комплекс одредени се следните својства: можност за дофатливост, склоност на амино-киселините да учествуваат во интеракции, хидрофобичност, планарност, густина на пополнетост на околниот волумен и дофатлива површина. Во (28) утврдено е дека хомодимерите имаат голема хидрофобичност и голема дофатлива површина, а хетеродимерите имаат значително помала хидрофобичност. Антителата имаат значително помала планарност за разлика од останатите типови на интеракции. Сепак, според овие својства неможе да се пронајде прецизен шаблон кој идеално ќе ги разграничува амино-киселинските остатоци кои се дел од сврзен регион во однос на преостанатите амино-киселински остатоци кој лежат на површината од протеинската молекула (28).

Во (80) анализирани се 136 хомодимери, и утврдено е дека околу една третина од протеинските интерфејси имаат поголема хидрофобичност, поларни контакти и интеракции со посредство на вода.

Една од клучните карактеристики за одредување на типот на интеракцијата е склоноста на амино-киселините да стапуваат во интеракции. Во (81) анализирани се шест типови на протеински интеракции: хомодимери наспрема хетеродимери, трајни наспрема привремени димери, и интеракции во кои учествуваат протеини од ист домен наспрема интеракции помеѓу протеини кој припаѓаат во различни домени. Врз основа на композицијата на амино-киселинските остатоци и склоноста на амино-киселините да стапуваат во интеракција постигнато е 63% прецизност во одредувањето на типот на интеракцијата.

Комплементарноста е многу важно својство на силните интеракции, при тоа може да се земе предвид комплементарноста на формата (82), (83), (84), (85), (86), (87), (88), комплементарноста на хемиските карактеристики (89), (90) или комбинација од нив (91), (92), (93). Кај силните интеракции забележана е голема комплементарност на формата и висока хидрофобичност, а кај краткотрајните интеракции има помала површина на сврзниот регион и помала комплементарност на формата (81), (94), (95).

Во (96) разгледувани се 70 протеински комплекси така што малите сврзни региони се разгледувани како еден интерфејс, а поголемите сврзни региони се разгледувани како спој од повеќе протеински интерфејси. Користејќи ја претпоставката дека хемиските карактеристики на интерфејсите се слични со хемиските карактеристики на амино-киселинските остатоци во останатите дел од површината, а густината на интерфејсите е слична со густината во центарот на протеинската структура, во (97) направено е разграничување на централниот и површинскиот дел на протеинските интерфејси.

Друга важна карактеристика која се користи во одредување на сврзните региони е конзервацијата на амино-киселинските остатоци. Користејќи ја претпоставката дека амино-киселинските остатоци кои се дел од сврзен регион имаат повисока конзервација од останатите остатоци, во (98) Valdar и Thornton направиле анализа кај шест хомодимери. Подоцна, на сличен начин во (99) направена е анализа врз поголемо множество. Во (100) со користење на Баесов пристап утврдено е дека централните амино-киселински остатоци и амино-киселинските остатоци кои го формираат интерфејсот имаат значителни разлики во конзервацијата. Постојат повеќе веб сервери кои овозможуваат одредување на конзервацијата на амино-киселинските остатоци врз основа на протеинската секвенца (101) или структура (102), (103). Сепак, конзервацијата не е доволно дискриминаторна карактеристика со која може да се разграничат амино-киселинските остатоци кои се дел од сврзен регион, но сепак оваа карактеристика може да се користи во комбинација со останатите карактеристики на амино-киселинските остатоци за идентификација на сврзните региони.

Забележано е дека амино-киселинските остатоци кои формираат сврзен регион немаат униформна енергетска распределба. Имено, дел од остатоците значително придонесуваат во енергетската размена која настанува при стапување во интеракција, и таквите амино-киселински остатоци се нарекуваат жешки точки (англ. hot spots). Овие амино-киселински остатоци најчесто се одредуваат со следење на тенденцијата на поврзување при мутација во аланин. На овој принцип формирана е ASEdb базата (104). Постојат и други бази со жешки точки кои се откриени по експериментален пат, како на пример BID базата (105), но овие бази содржат податоци за мал дел од интеракциите. Во (106) даден е преглед на предизвиците и

достигнувањата во однос на идентификацијата на жешки точки. Robetta (107) и FoldX (108) се веб сервери со кои се врши одредување на жешките точки врз основа на енергијата. Постојат повеќе методи кои предвидувањето на жешките точки го базираат на молекуларната динамика (109), (110), (111), но овие методи се временски захтевни. Забележана е голема корелација помеѓу жешките точки и остатоците со висока конзервација, па затоа конзервацијата може да се користи како клучна карактеристика при идентификација на жешки точки (112), (113), (114). Така на пример HotSprint методот (115) предвидувањата ги базира на конзервацијата на амино-киселинските остатоци, при што предвид се зема склоноста на амино-киселините да бидат дел од жешка точка, како и површината на жешките точки. KFC веб серверот (116) врши предвидување на жешките точки користејќи метод кој предвид ги зема контактите меѓу атомите и хидрогенските врски.

Не сите протеински интерфејси имаат биолошко значење. Имено, оние интерфејси кои немаат биолошко значење не ја дефинираат функцијата на протеините, па истите може да внесат шум во предвидувањето. Затоа постојат истражувања кои се однесуваат на одредување дали даден протеински интерфејс има биолошко значење или не. Ова претставува многу сложен процес, особено доколку олигомерната состојба на протеинот не е позната (98). Некои истражувања за одредување на биолошките интерфејси се базираат на големината на површината, па така на пример Henric и Thornton имаат дефинирано минимален праг за големината на површината на интерфејсот за тој да има биолошко значење. На овој начин добиена е PQS базата (117) со биолошки интерфејси. Сепак големината на површината не е доволна карактеристика за одредување дали даден интерфејс има биолошко значење или не, па така во (118), покрај големината на површината, предвид се зема и конзервацијата на амино-киселинските остатоци. Од истражувањата во (119), (120) може да се донесе заклучок дека генерално протеинските интерфејси кои имаат биолошко значење имаат голема површина и висока конзервација. Постојат и други методи за идентификација на интерфејсите кои имаат биолошко значење кои дополнително предвид земаат и други својства на амино-киселинските остатоци, како што се на пример NOXclass (121) и DiMoVo (122) методите.

Во (36) и (37) е даден широк преглед на постојните методи и веб сервери за предвидување на сврзните делови од протеинската структура, како и на постоечките бази со протеински интерфејси. Авторите на трудот (123) ги презентираат најчесто користените карактеристики за предвидување на сврзните деловите од протеинската структура. Покрај претходно споменатите методи, во литература може да се најдат уште голем број други методи за одредување на сврзните делови од протеинската молекула, како на пример ProMate (124), PPI-Pred (125), SHARP2 (126), SPPIDER (127), cons-PPISP (128), PRISM (129), (130), (131) и други. Истражувањата за предвидување на протеинските интерфејси произвеле знаење кое потоа е

сместено во повеќе бази со протеински интерфејси, меѓу кои спаѓаат: PRINT (132), PiBASE (133), InterPare (134), SCOWLP (135), SCOPPI (136), 3DID (137) и други. Сепак во овие бази со протеински интерфејси нема податоци за сите протеински молекули чија структура е веќе позната, а дел од нив содржат податоци само за сврзните места каде што настанува интеракција, но нема придружено соодветни функционални анотации за тие интеракции. Од тука очигледна е потребата за развој на брзи и прецизни компјутерски методи за идентификација на сврзните делови од протеинските структури, како и развој на методи за анотација на протеини врз база на карактеристиките на сврзните делови од протеинската структура.

3

ПРЕБАРУВАЊЕ И КЛАСИФИКАЦИЈА НА ПРОТЕИНСКИ СТРУКТУРИ

Структурата на откриените протеински структури се чува во посебен формат во Protein Data Bank (PDB) (1), (2) базата на податоци. Во оваа база покрај податоци за структурата на протеините се чуваат податоци и за откриените нуклеински киселини и макромолекуларни комплекси. Податоците за протеинските структури најчесто се чуваат во pdb формат, односно во полуструктурирани текстуални датотеки во кои се чуваат податоци за примарната, секундарната и терциерната структура на протеините. До 4 февруари 2014 година во PDB базата на податоци се сместени 97591 структура.

Методите кои ќе бидат презентирани и предложени во овој докторски труд се базираат на тродимензионалната протеинска структура. Секој протеин може да се разгледува како тродимензионален објект, па потоа користејќи различни техники за споредба на 3Д објекти може да се прави пребарување на истите. За таа цел, ќе се применат неколку техники за екстракција на карактеристиките на протеинската структура со што ќе се формира карактеристичен вектор - таканаречен дескриптор. Потоа, во процесот на пребарување споредбата помеѓу две протеински структури ќе се направи користејќи различни метрики за сличност со цел да се споредат двата карактеристични вектори во векторскиот простор. При тоа во процесот на екстракција на карактеристичниот вектор на дадена протеинска структура треба да се обезбеди инваријантност на транслација, скалирање и ротација.

Во [А2] беа презентирани неколку методи за екстракција на протеински дескриптори. Во продолжение ќе биде даден кус опис на овие методи. Исто така ќе биде презентиран и Matrix Alignment by Sequence Alignment within Sliding Window (MASASW) методот [А31] кој може да се користи за споредба на две протеински структури.

3.1. Протеински воксел-базиран дескриптор

Бидејќи протеинските структури во оваа дисертација се разгледуваат како тродимензионални објекти, може да се применат различни техники за екстракција на нивните геометриски својства. Во (138) се презентирани различни техники за екстракција на геометриските карактеристики на тродимензионални објекти, со што се формира еднозимензионален карактеристичен вектор. Во (138) овие техники се применети за пребарување на разни типови на објекти (пример: животни, авиони итн.), но не и за пребарување на протеински структури, чие споредување е релативно посложен процес бидејќи во овој случај постои поголема геометриска сличност помеѓу тродимензионалните протеински објекти кои се споредуваат.

Користејќи го воксел дескрипторот презентиран во (138), во [А4] беше предложен протеинскиот воксел-базиран дескриптор. Со цел да се овозможи повисока прецизност, покрај геометриските карактеристики кои се извлечени од протеинската терциерна структура, дополнително предвид беа земени и неколку карактеристики на примарната и секундарната структура, како што ќе биде опишано подоцна.

Екстракција на карактеристиките на терциерната структура

Екстракцијата на геометрискиот дел од дескрипторот се одвива во четири фази: триангулација, нормализација, вокселизација и фаза на примена на Дискретна Фуриева трансформација. Амино-киселинските остатоци се состојат од одреден број на атоми, и за секој атом се знае неговиот радиус и точната позиција во просторот. Од тука, секој атом може да се претстави како сфера. Во фазата на **триангулација** сферата на секој атом се претставува преку триаголна решетка, односно секоја сфера се апроксимира со унија од триаголни полигони чии јазли се компланарни. Од тука, протеинската структура се претставува со следниот модел

$$I = \bigcup_{i=1}^{m} T_i = \bigcup_{i=1}^{m} \Delta p_{Ai} p_{Bi} p_{Ci}$$

$$T = \{T_1, ..T_m\}, T_i \subset R^3 \tag{3.1}$$

$$P = \{p_i \mid p_i = (x_i, y_i, z_i) \in R^3,\ 1 \le i \le n\},$$

каде што триаголникот T_i се дефинира преку јазлите p_{Ai}, p_{Bi} и p_{Ci}, а T и P се множество од триаголници и множество од јазли во тродимензионалниот простор. Секој јазел p_i е дефиниран преку неговите координати (x_i, y_i, z_i). Потоа со равенството (3.2) се пресметува површината S_i на даден триаголник T_i, и се пресметува вкупната површина S на целата решетка.

$$S_i = \frac{1}{2}\left|(p_{Ci} - p_{Ai}) \times (p_{Bi} - p_{Ai})\right| \qquad S = \sum_{i=1}^{m} S_i \tag{3.2}$$

После триангулацијата следи **нормализација** при што се врши транслација и скалирање со цел да се обезбеди инваријантност на транслација и скалирање. Имено, со нормализацијата се обезбедува секоја протеинска структура да се претстави во унифициран референтен координатен систем. За таа цел прво се одредува центарот на маса на протеинската структура, а потоа сите јазли се транслираат така што центарот на маса да се наоѓа во координатниот почеток. Потоа се наоѓа Евклидовото растојание помеѓу центарот на маса и најоддалечениот јазел, и моделот се скалира за ова растојание.

По нормализација, следи фазата на **вокселизација** со што континуалниот 3Д простор се дели во елементарни ќелии наречени воксели. Прво, се врши дискретизација, односно континуалниот 3Д простор се дели во N^3 воксели. Во овој докторски труд N се поставува на 8 согласно препораките во (138), па со тоа се добиваат 512 воксели. Вокселот μ_{abc} е кубоиден регион со димензии $d_x \times d_y \times d_z$ и во себе ја содржи точката $(a,b,c) \in Z^3$ при што

$$\mu_{abc} = \{(x,y,z) \mid (x,y,z) \in R^3, \quad a \le \frac{x - x_{\min}}{d_x} \le a+1,$$

$$b \le \frac{y - y_{\min}}{d_y} \le b+1, \quad c \le \frac{z - z_{\min}}{d_z} \le c+1\}, \tag{3.3}$$

$$d_x = \frac{x_{\max} - x_{\min}}{N}, \quad d_y = \frac{y_{\max} - y_{\min}}{N}, \quad d_z = \frac{z_{\max} - z_{\min}}{N},$$

$$a = 0, \ldots, N-1, \; b = 0, \ldots, N-1, \; c = 0, \ldots, N-1$$

каде x_{min} , y_{min}, z_{min}, x_{max} , y_{max} и z_{max} се минималните и максималните вредности на x, y и z координатите. По дискретизацијата, се врши семплирање со што на секој воксел μ_{abc} му се доделува вредност v_{abc} која означува колкав дел од вкупната површина S на решетката се наоѓа во вокселот μ_{abc}

$$\upsilon_{abc} = \frac{area\{\mu_{abc} \cap I\}}{S}, \qquad 0 \le a,b,c \le N-1. \tag{3.4}$$

За пресметување на вредноста на даден воксел се користи следниов алгоритам. Секој триаголник T_j се дели на p_j^2 триаголни партиции со површина $\delta = S_j / (p_j^2 S)$, каде S_j е површината на триаголникот T_j, а p_j е параметар кој го дефинира бројот на партиции на триаголникот T_j. Доколку сите јазли на триаголникот T_j лежат во ист воксел, тогаш $p_j = 1$. Во спротивно, бројот на партиции на триаголникот се пресметува како

$$p_j = \left\lceil \sqrt{p_{\min} \frac{S_j}{S}} \right\rceil, \tag{3.5}$$

каде со параметарот p_{min} се дефинира финоста на апроксимација. Во оваа докторска дисертација параметарот p_{min} е поставен на 32000, како во (138). По поделбата на триаголникот на p_j^2 триаголници, за секој од новодобиените триаголници се одредува во кој воксел припаѓа центарот на дадениот триаголник, па соодветно се инкрементира вредноста на соодветниот воксел за δ. Алгоритамот за апроксимација на вредноста на даден воксел е опишан со Алгоритамот Д1 даден во Додатокот. Вредностите на вокселите се запишуваат во тродимензионална матрица која може да се користи како карактеристичен вектор во пребарувањето.

Сепак, за две идентични протеински структури кои се заротирани во тродимензионалниот простор ќе се добијат различни матрици. Затоа врз оваа матрица се применува **3Д Дискретна Фуриева трансформација** со што се обезбедува инваријантност на ротација и покомпактна репрезентација на протеинската структура. Прво, се врши поместување на индексите со што (a,b,c) се транслира во $(a - N/2,\ b - N/2,\ c - N/2)$, а потоа се применува 3Д дискретна фуриева трансформација.

$$\upsilon'_{a-N/2,\, b-N/2,\, c-N/2} = \upsilon_{abc}$$

$$f'_{pqs} = \frac{1}{\sqrt{N^3}} \sum_{a=-N/2}^{N/2-1} \sum_{b=-N/2}^{N/2-1} \sum_{c=-N/2}^{N/2-1} \upsilon'_{abc} e^{-2\pi j(ap+bq+cs)/N} \tag{3.6}$$

$$p = 0, \ldots, N-1,\ q = 0, \ldots, N-1,\ s = 0, \ldots, N-1,$$

каде j е имагинарна единица.

Потоа, коефициентите се нормализираат со делење со коефициентот $|f'_{000}|$ и протеинскиот воксел-базиран дескриптор се формира од коефициентите со ниска фреквенција. Бидејќи постои

симетрија помеѓу добиените коефициенти, во дескрипторот предвид се земаат само несиметричните коефициенти за кои е исполнет условот $1 \leq |p|+|q|+|s| \leq k \leq N/2$. Постапката на формирање на протеинскиот воксел-базиран дескриптор е опишана со Алгоритамот Д2 во Додатокот. Димензијата на карактеристичниот вектор зависи од параметарот k и се пресметува како $k(2k^2+3k+4)/3$. Во овој докторски труд $k = 8$, со што се добива карактеристичен вектор со 416 елементи. Овие 416 елементи ги опишуваат геометриските карактеристики на протеинската терциерна структура.

Екстракција на карактеристиките на примарната и секундарната структура

На овие карактеристики дополнително се додаваат уште 34 карактеристики на примарната и секундарната структура на протеинот, како што е направено во (139), па димензијата на новодобиениот вектор е $dim = 450$. Од примарната структура предвид се зема фреквенцијата на појавување на дваесетте различни амино-киселини, како и процентуалната застапеност на хидрофобичните амино-киселини. Последната карактеристика предвид го зема хидрофобичниот ефект кој укажува дека хидрофиличните амино-киселини почесто се лоцирани во близината на површината на протеинот, додека хидрофобичните амино-киселини почесто се наоѓаат во централниот дел од протеинската структура. Од секундарната структура предвид се земаат: бројот на α-хеликси, β-рамнини и превиткувања, како и бројот на појавувања на секој од десетте различни типови на α-хеликси. Дополнително, на овие карактеристики им се доделуваат различни тежини во фазата на споредување (Табела 3.1), како што е направено во (139).

Карактеристики на примарната структура	**Тежина**
Однос на хидрофобични амино-киселини	6 %
Однос на амино-киселини	90 %
Карактеристики на секундарната структура	**Тежина**
Број на α-хеликси	1 %
Број на β-рамнини	1 %
Број на превиткувања	1 %
Тип на α-хеликси	1 %

Табела 3.1 Тежински фактори кои се доделуваат на карактеристиките на примарната и секундарната структура на протеинот.

Метрика за растојание

Во процесот на споредба се пресметува растојанието D помеѓу дескрипторот на протеинот за тестирање f' и дескрипторот на протеинот за обука со кој се споредува f''. Геометриските делови од дескрипторите на протеинот за тестирање f_T' и протеинот за обука f_T'' се споредуваат со

равенството (3.7) како што е опишано во (138), со што се одредува растојанието на карактеристиките од терциерната структура D_T.

$$D_T = \sqrt{\sum_{i=1}^{416}[f_T(i)' - \alpha f_T(i)'']^2}\ , \qquad \alpha = \sum_{i=1}^{416} f_T'(i) * f_T''(i) \Big/ \sum_{i=1}^{416} f_T''(i)^2 \tag{3.7}$$

Растојанието во примарната и секундарната структура D_{PS} се одредува како

$$D_{PS} = \sqrt{\sum_{i=1}^{34} W_i * [f_{PS}'(i) - f_{PS}''(i)]^2}\ , \tag{3.8}$$

каде f_{PS}' и f_{PS}'' се дескрипторите со карактеристиките на примарната и секундарната структура на протеинот за тестирање и протеинот за обука со кој се прави споредба. Со W_i се означени тежините кои се доделуваат на различните карактеристики од примарната и секундарната структура, и истите се дадени во Табела 3.1. Вкупната сличност D се одредува како во (139)

$$D = k_T D_T + k_{PS} D_{PS}, \qquad k_T = 90\%,\ k_{PS} = 10\%, \tag{3.9}$$

со што споредбата се базира на карактеристиките од терциерната структура (геометриските карактеристики учествуваат со 90% тежина), наспроти карактеристиките на примарната и секундарната структура. Споредбата на два дескриптори има линеарна комплексност $O(dim)$.

3.2. Дескриптор базиран на интерполација на скелетот на протеинот

Во претходните истражувања беше покажано дека попрецизно пребарување се обезбедува доколку во предвид се земат само Cα атомите, кои го формираат скелетот на протеинот, наместо да се земат предвид сите атоми кои се дел од дадената протеинска структура [A6], [A31]. На овој начин не само што се обезбедува попрецизна репрезентација на протеинот, туку дополнително се добива и поедноставена претстава на истиот. За таа цел, понатаму се фокусираме на методи за пребарување на протеински структури кои предвид ги земаат само Cα атомите.

Екстракција на карактеристиките

Со цел да се обезбеди инваријантност на транслација и скалирање, прво се врши транслација на протеинската структура така што центарот на маса се транслира во координатниот почеток. Потоа се одредува Евклидовото растојание d_{max} од центарот на маса до најоддалеченият Cα атом, и моделот се скалира за d_{max}, со што протеинот се сместува во сфера со единечен радиус. Потоа карактеристиките на скелетот на протеинот може да се добијат како Евклидови растојание помеѓу секој Cα атом и центарот на маса. На овој начин се обезбедува инваријантност на ротација.

Бидејќи различен протеин има различен број на Cα атоми, на овој начин ќе се добијат дескриптори со различна големина, па затоа треба да се обезбеди унифициран начин за репрезентирање на сите протеински структури преку дескриптори кои имаат еднаква должина. За таа цел се врши интерполација на протеинскиот скелет, со што истиот се претставува преку $N = 2^k$ интерполациски точки. Во [A6] беше предложен протеинскиот дескриптор базиран на рамномерна интерполација на скелетот на протеинот, а во [A31] дополнително беше воведен и метод кој врши нерамномерна интерполација.

Со ***рамномерната интерполација***, секој дел од скелетот на протеинот се анализира со иста резолуција, во зависност од бројот на интерполациски точки N. Интерполацијата се врши во два чекори:

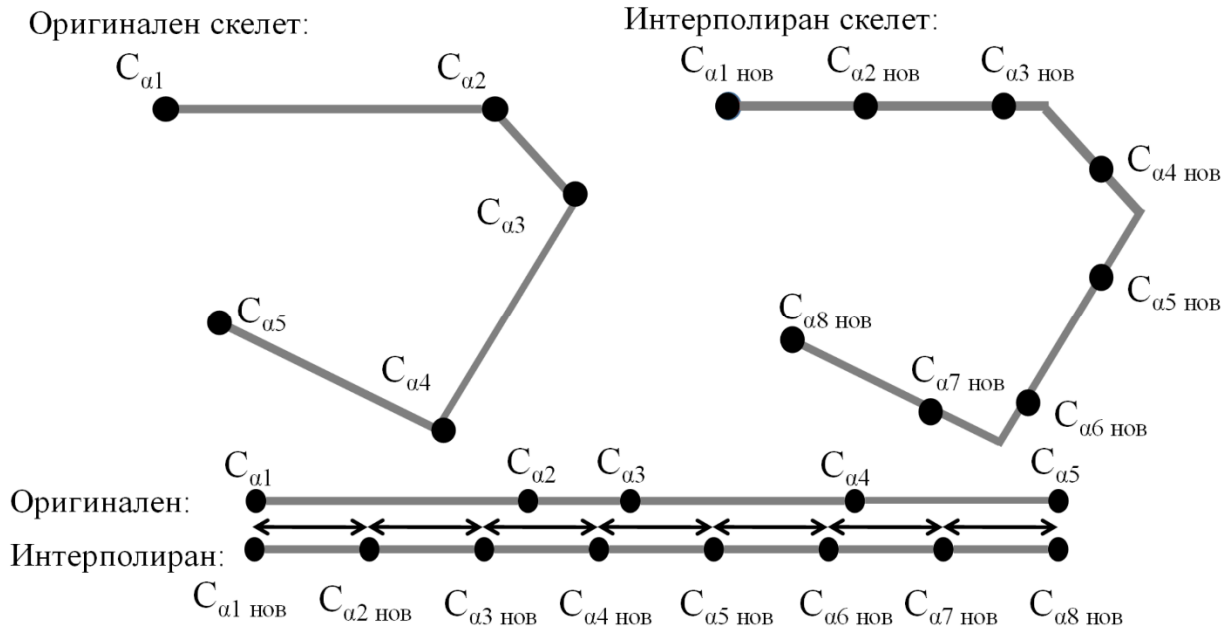
1. Се пресметува вкупната должина на протеинскиот скелет

$$L = \sum_{i=1}^{n_{C_\alpha}-1} d_{i,i+1}\,, \tag{3.10}$$

каде што $n_{C\alpha}$ е бројот на Cα атоми, а со $d_{i,i+1}$ е означено Евклидовото растојание помеѓу i-тиот и (i+1)-виот Cα атом.

2. Потоа протеинскиот скелет се дели на 2^k сегменти со должина $l=L/N$, каде што $N=2^k$.

На Слика 3.1 е илустриран процесот на рамномерна интерполација на протеинскиот скелет.

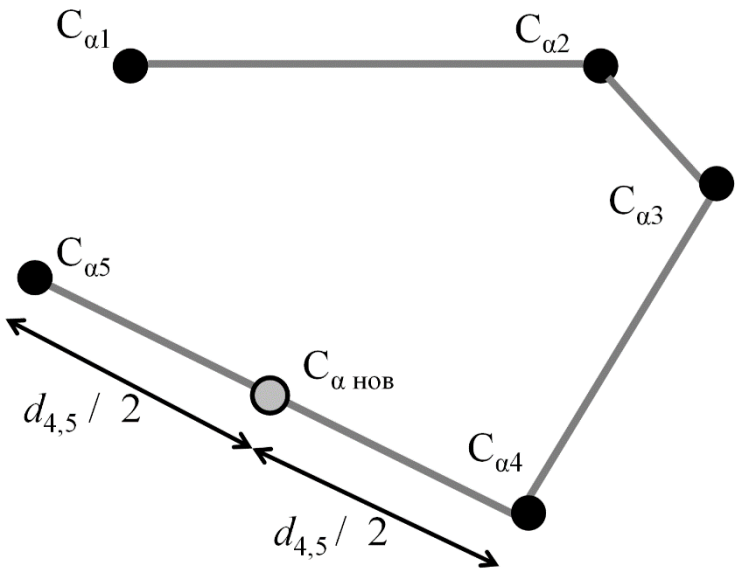


Слика 3.1 Рамномерна интерполација на протеинскиот скелет.

Нерамномерната интерполација се врши во три чекори:

1. Како иницијални интерполациски точки се земаат точките кои имаат исти координати како Cα атомите од протеинската структура чиј скелет треба да се интерполира. Се одредува $n_n=2^n$ како број кој е степен од 2 и кој е најблиску до бројот на Cα атоми $n_{C\alpha}$. Потоа се одредува бројот на интерполациски точки кои треба да се вметнат или отстранат како $\Delta n = (n_{final} - n_{C\alpha})$, каде што $n_{final} = 2^f$ и $f = max\ (n, k)$.
2. Доколку $\Delta n > 0$, тогаш се додаваат дополнителни Δn интерполациски точки на средината помеѓу Δn парови од соседни точки кои се на најголемо растојание вдолж скелетот во тековната итерација. На Слика 3.2 е илустриран процесот на додавање на нова интерполациска точка означена како $C_{\alpha\ нов}$.

   Доколку $\Delta n < 0$, тогаш се отстрануваат Δn интерполациски точки од оригиналните точки. Овие точки одговараат на првите Cα атоми од секој пар на соседни точки чие растојание е најмало вдолж скелетот во тековната итерација.
3. Бидејќи сакаме протеинскиот скелет да го претставиме преку $N=2^k$ интерполациски точки, затоа на крај предвид се зема секоја $(f-k+1)$-ва точка од претходно добиените точки.



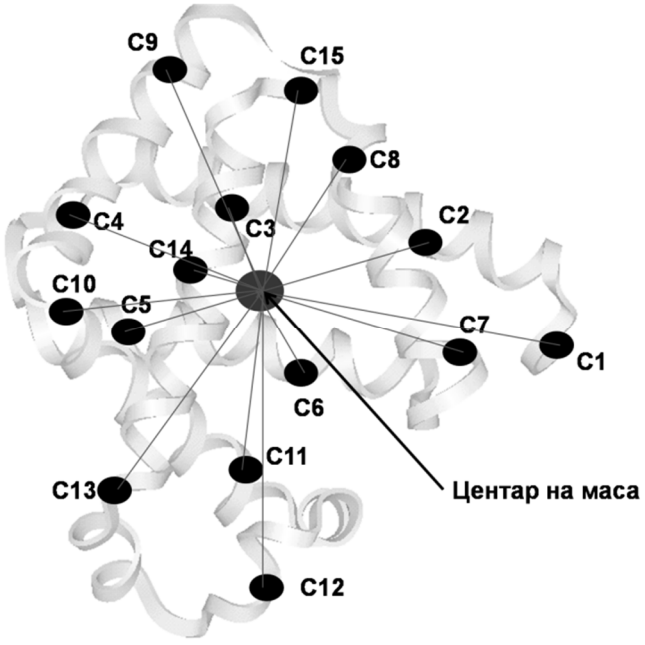
Слика 3.2 Додавање на нова интерполациска точка.

Така на пример ако се разгледува две протеински структури, така што првата има 1000, а втората 100 Cα атоми, тогаш ќе се добијат 1024 и 128 интерполациски точки соодветно. За првиот протеин ќе се направи додавање на 24 нови интерполациски точки, а за вториот протеин ќе се направи додавање на 28 нови интерполациски точки. Доколку треба да се извлечат

дескриптори со должина еднаква на 64, тогаш за првиот протеин предвид се зема секоја петта интерполациска точка, а за вториот протеин предвид се зема секоја втора интерполациска точка.

Со нерамномерната интерполација, различни делови од протеинскиот скелет се анализираат со различна резолуција. Иако растојанието помеѓу соседни Cα атоми е секогаш околу 3.7 Å (Ångstrom), сепак конкретната поставеност на протеинот во 3Д просторот ќе дефинира за кои парови од соседни Cα атоми растојанието ќе биде помало, а за кои парови ќе биде поголемо од ова просечно растојание (140).

Откако се обезбеди унифициран број на интерполациски точки, следува екстракција на дескрипторот. За таа цел се применува идејата на познатиот дескриптор базиран на зраци (англ. rays) за пребарување на 3Д објекти (138). Кај класичниот дескриптор, прво се прави триангулација на површината на 3Д објектот со што се добива мрежест (меш) модел, а потоа елементите на дескрипторот се пресметуваат како Евклидови растојанија помеѓу точките од меш моделот и центарот на маса. На сличен начин, елементите на протеинскиот дескриптор базиран на интерполација на скелетот се одредуваат како Евклидови растојанија помеѓу интерполациските точки и центарот на маса. На Слика 3.3 е илустриран процесот на екстракција на дескрипторот според интерполациските точки кои го апроксимираат скелетот на протеинот. Карактеристично за овој дескриптор е тоа што со одбирање на должината на дескрипторот може скелетот на протеинот да се анализира со различна резолуција согласно барањата за брзината и прецизноста кои треба да се постигнат во пребарувањето.



Слика 3.3 Илустрација на процесот на екстракција на дескрипторот од добиените интерполациски точки.

Идејата на овој дескриптор е да се опише како протеинскиот скелет се приближува и оддалечува во однос на центарот на маса. На Слика 3.4 се прикажани вредностите на елементите на дескрипторот за протеинот 102L ланец А. Прикажани се вредностите на елементите на дескрипторот со должина 512 користејќи рамномерна и нерамномерна интерполација на скелетот. Овој протеински ланец има 163 Cα атоми, па се додаваат 349 нови интерполациски точки при нерамномерна интерполација. Помеѓу соседните Cα атоми кои се на поголемо растојание се додаваат поголем број на нови точки. На овој начин со нерамномерна интерполација различни делови од скелетот се анализираат со различна резолуција во зависност од растојанијата помеѓу соседните Cα атоми кои се еднозначно дефинирани со конформацијата која ја формира протеинската структура. Од друга страна, со рамномерна интерполација се добиваат интерполациски точки кои се на еднакви меѓусебни растојанија вдолж скелетот на протеинот. Доколку за дадена протеинска структура бројот на Cα атоми е степен од 2, тогаш дескрипторите добиени со двете верзии на интерполација ќе бидат идентични.

Слика 3.4 Протеински дескриптор базиран на интерполација на скелетот на протеинот 102L ланец А.

Метрики за растојание

При споредба се користат L_1 и L_2 нормите како метрики за растојание помеѓу два вектори. Нека f_1 и f_2 се дескрипторите на протеините кои треба да се споредат, и нека должината на дескрипторите е N. Сумата од аритметичките растојанија D_1 помеѓу овие два вектори се пресметува со L_1 норма, а сумата од квадратните Евклидови растојанија D_2 се пресметува со L_2 норма

$$D_1 = \sum_{i=1}^{N} \left| f_1(i) - f_2(i) \right| , \qquad D_2 = \sqrt{\sum_{i=1}^{N} [f_1(i) - f_2(i)]^2} . \tag{3.11}$$

Споредбата на два дескриптори е со линеарна комплексност $O(N)$.

3.3. Протеински дескриптори базирани на бранчиња

Голем број на методи за пребарување на протеини се базираат на анализа на матрицата на растојанија која ги содржи Евклидовите растојанија помеѓу секој пар од Cα атоми. Оваа матрица е добар репрезент на протеинските структури бидејќи протеините со слична структура ќе имаат и слични матрици на растојанија. Преку матрицата на растојанија се обезбедува инваријантност на транслација и ротација. Со цел да се обезбеди инваријантност на скалирање, дополнително се врши нормализација на елементите така што да примаат вредности во даден предефиниран опсег. Некои пристапи за пребарување на протеински структури вршат порамнување на матриците на растојанија, како што се на пример DALI (19) и MatAlign (70) методите. Кај друга група на методи се врши анализа на матрицата на растојанија при што истата се разгледува како дискретен 2Д сигнал и врз неа се применуваат различни трансформации со цел да се генерира еднодимензионален карактеристичен вектор (дескриптор) (73).

Бидејќи матриците на растојанија зафаќаат многу мемориски простор, а пресметувањето на разликата помеѓу две такви матрици е со комплексност $O(N^2)$, каде што N е бројот на редици и колони во матрицата, затоа е подобро овие 2Д матрици да се трансформираат во еднодимензионални вектори со што ќе се овозможи побрзо пребарување. Матрицата на растојанија може да се разгледува како слика врз која може да се примени анализа на бранчиња за да се извлечат најрелевантните карактеристики на протеинската структура.

Трансформација на бранчиња

Во 19-ти век Jean Baptiste Joseph Fourier ја вовел Фуриевата трансформација со која се овозможува разложување на дадена функција на нејзините фреквентни компоненти. Преку Фуриевата трансформација може да се одреди кои фреквенции се присутни во сигналот, но не може да се утврди точното време кога се појавила дадена фреквенција. Затоа подоцна е воведена трансформацијата на бранчиња (англ. wavelet transformation) (141) која овозможува просторно-временска анализа на сигналот.

Со трансформацијата на бранчиња секоја функција може да се претстави како линеарна комбинација од базните функции

$$f(x) = \sum_k \alpha_k \varphi_k(x) \tag{3.12}$$

каде со $\varphi_k(x)$ се означени базните функции, а α_k се реални коефициенти. Од базната функција $\varphi(x)$, се формираа множество од функции $\{\varphi_{j,k}(x)\}$ кои се нејзини транслирани и скалирани верзии

$$\varphi_{j,k}(x) = 2^{j/2}\varphi(2^j x - k) \tag{3.13}$$

каде со k се дефинира позицијата на транслираната функција, а j е фактор за скалирање. Функцијата $\varphi(x)$ се нарекува скалирачка функција.

Нека V_n е множеството од функции кои можат да се развијат од $\{\varphi_{j_n,k}(x)\}$. Базната функција која припаѓа во множеството V_i може да се развие преку функциите кои припаѓаат во множеството V_{i+1}

$$\varphi_{j,k}(x) = \sum_n h_\varphi(n)\varphi_{j+1,n}(x), \tag{3.14}$$

каде $h_\varphi(n)$ е вектор со скалирачки коефициенти. Од равенствата (3.13) и (3.14) следи дека

$$\begin{aligned} \varphi_{j,k}(x) &= \sum_n h_\varphi(n) 2^{(j+1)/2}\varphi(2^{j+1}x - n) \\ \varphi(x) &= \sum_n h_\varphi(n)\sqrt{2}\varphi(2x - n). \end{aligned} \tag{3.15}$$

Нека е дадена функцијата $\psi(x)$ која формира множество од функции кои се транслирани и скалирани верзии од неа

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k). \tag{3.16}$$

Преку множеството $\{\varphi_{j,k}(x)\}$ со одредена апроксимација може да се развие функцијата $f(x)$, а со множеството $\{\Psi_{j,k}(x)\}$ може да се развие разликата помеѓу функција $f(x)$ и нејзината апроксимација. Функцијата $\Psi(x)$, која се нарекува функција на бранче, може да се развие како

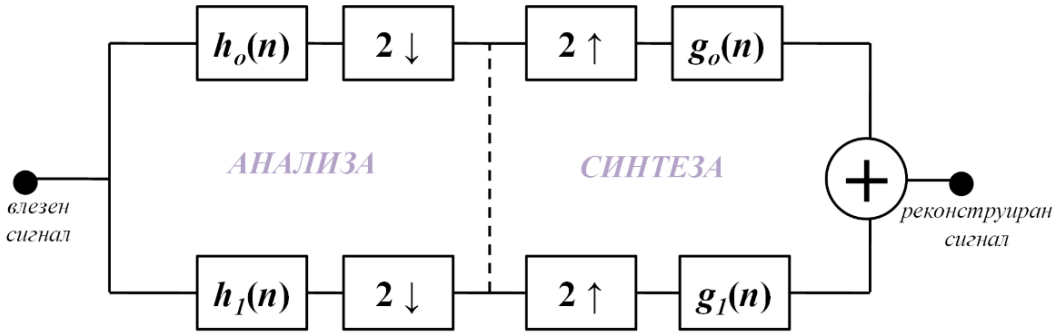
$$\psi(x) = \sum_n h_\psi(n)\sqrt{2}\varphi(2x - n) \tag{3.17}$$

каде $h_\Psi(n)$ е вектор на бранчето.

За побрзо пресметување на коефициентите на бранчето, воведена е брзата трансформација на бранчиња (Fast Wavelet Transform – FWT) која е базирана на кодирање на фреквентни опсези. Со оваа трансформација се врши декомпозиција на сигналот во подопсези, од кои потоа со примена на инверзна трансформација може да се добие оригиналниот сигнал без грешка. Секој опсег се добива така што сигналот се филтрира со нископропусен и високопропусен филтер. Опсегот на добиените сигнали е двојно помал од опсегот на оригиналниот сигнал, па според теоремата на Најквист нема да се загуби никаква информација доколку предвид се земе секој втор примерок. Потоа, во процесот на реконструкција на сигналот бројот на примероци се

дуплира, се врши филтрирање и на крај се прави сумирање на подопсезите, како што е прикажано на Слика 3.5.

Со $h_0(n)$ и $h_1(n)$ се означени импулсните одзиви на нископропусниот и високопропусниот филтер за анализа, додека за синтеза се користат филтри со импулсни одзиви $g_0(n)$ и $g_1(n)$. Со конволуција на $h_0(n)$ и влезниот сигнал се добива сигнал кој е груба апроксимација на влезниот сигнал, и истиот претставува збир од сигналите на ниски фреквенции. Со конволуција на $h_1(n)$ и влезниот сигнал се добива сигнал кој содржи информации за деталите на влезниот сигнал.



Слика 3.5 Анализа и синтеза на сигнал преку кодирање на опсег.

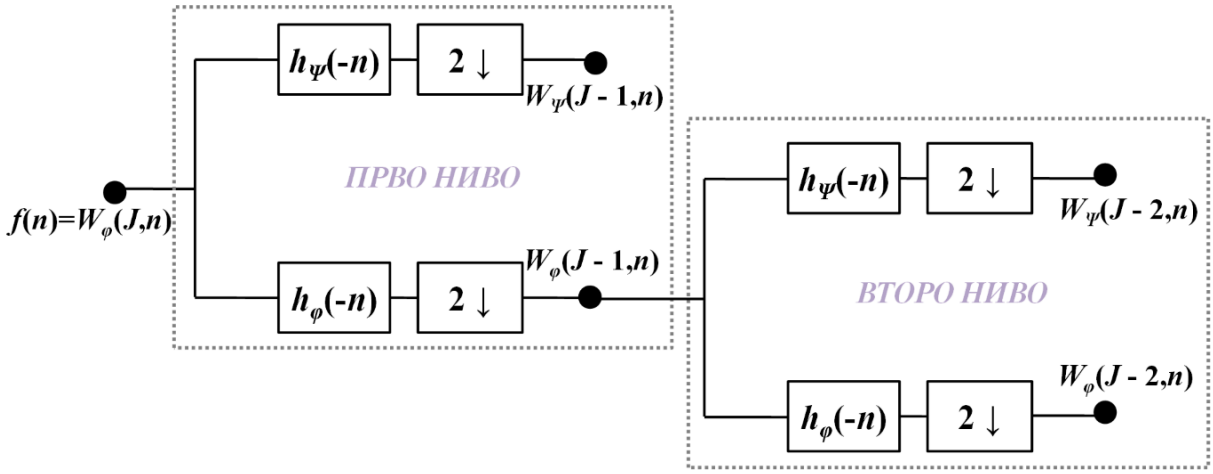
Импулсните одзиви на филтрите треба да се одредат така што по реконструкција на сигналот не треба да има никаква загуба на информација. Со математичка анализа може да се добие систем на равенки, па решенијата на овој систем равенки ќе соодветствуваат на формите на филтрите. Преку равенството (3.18) може да се дефинира прототип преку кој ако е позната формата на еден филтер, тогаш може да се одреди формата и на другиот филтер. Во литературата се среќаваат различни прототипови на филтри, па од тука произлегуваат и различните фамилии на бранчиња кои можат да се користат.

$$g_0(n) = (-1)^n h_1(n), \qquad g_1(n) = (-1)^{n+1} h_0(n) \tag{3.18}$$

Деталните коефициенти $W_\psi(j,k)$ кои ги опишуваат деталите на сигналот и апроксимативните коефициенти $W_\varphi(j,k)$ кои ги опишуваат најопштите својства на сигналот можат да се добијат со конволуција на апроксимативните коефициенти $W_\varphi(j+1,k)$ на претходното ниво

$$\begin{aligned} W_\varphi(j,k) &= h_\varphi(-n) * W_\varphi(j+1,n)\big|_{n=2k,k\geq 0} \\ W_\psi(j,k) &= h_\psi(-n) * W_\varphi(j+1,n)\big|_{n=2k,k\geq 0}, \end{aligned} \tag{3.19}$$

каде $h_\psi(-n)$ и $h_\varphi(-n)$ се огледално свртените верзии на бранчето и на скалирачкиот вектор, соодветно. Исто така бројот на примероци се намалува за 2 (n=2k). Ова може да се повторува на повеќе нивоа, со што апроксимираниот сигнал од претходното ниво ќе биде влезен сигнал во следното ниво. На Слика 3.6 е прикажан процесот на декомпозиција на сигналот на детални и апроксимативни коефициенти во две нивоа. На првото ниво сигналот се кодира во два опсега, а потоа истото се повторува на второ ниво при што на влез се дава сигналот добиен на ниските фреквенции. На повисоките нивоа, резолуцијата на ниските фреквенции се зголемува, со што се овозможува сигналот да се анализира со различна резолуција согласно потребите.



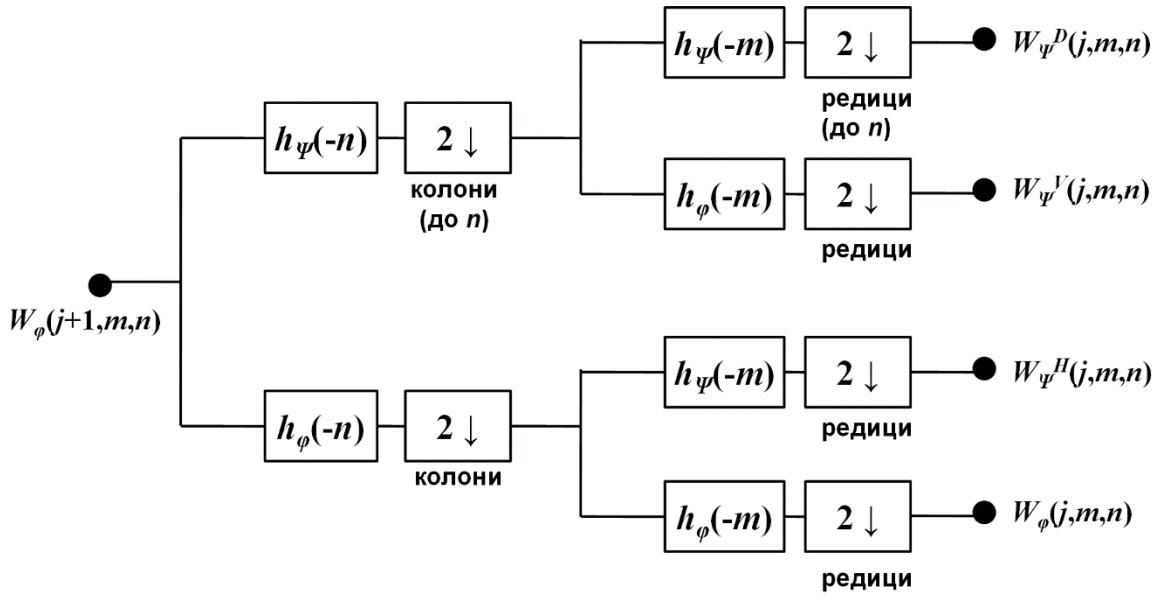
Слика 3.6 Декомпозиција на сигналот на две нивоа.

Во оваа докторска дисертација трансформацијата на бранчиња ќе биде направена врз матриците на растојанија кои претставуваат 2Д сигнали. За таа цел се користи дводимензионална скалирачка функција $\varphi(x,y)$ и три дводимензионални функции на бранчиња $\Psi^H(x,y)$, $\Psi^V(x,y)$ и $\Psi^D(x,y)$ кои се добиваат како производ на еднодимензионалните верзии на скалирачката и функцијата на бранче

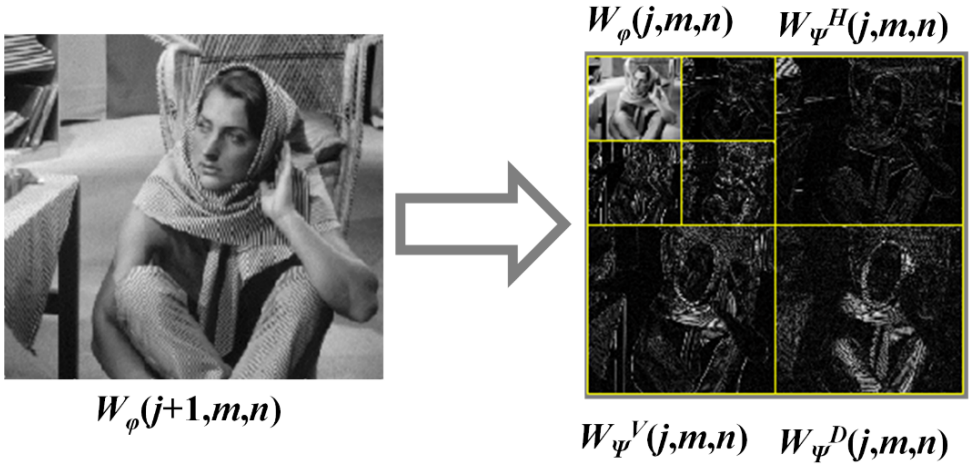
$$\begin{aligned}\varphi(x,y) &= \varphi(x)\,\varphi(y)\\ \psi^H(x,y) &= \psi(x)\,\varphi(y)\\ \psi^V(x,y) &= \varphi(x)\,\psi(y)\\ \psi^D(x,y) &= \psi(x)\,\psi(y).\end{aligned} \tag{3.20}$$

Со овие дводимензионални функции на бранче се анализираат промените во 2Д сигналот (во овој случај матрицата на растојанија) во три правци, и тоа: со $\Psi^H(x,y)$ се анализираат хоризонталните детали, со $\Psi^V(x,y)$ се анализираат вертикалните детали, а со $\Psi^D(x,y)$ се опишуваат дијагоналните детали. За да се добијат овие функции, прво се применува FWT врз редиците, а

потоа врз колоните. $\Psi^H(x,y)$ се добива така што прво се применува нископропусен филтер, а потоа високопропусен филтер. $\Psi^V(x,y)$ се добива така што прво се применува високопропусен филтер, па нископропусен филтер. Со примена на високопропусен филтер и врз редиците и врз колоните се добива $\Psi^D(x,y)$, а со примена на нископропусен филтер и на редиците и на колоните се добива апроксимацијата на сигналот $\varphi(x,y)$. На Слика 3.7 е илустрирана двонивовска FWT, додека на Слика 3.8 е даден пример од примената на двонивовска FWT врз конкретна 2Д слика.



Слика 3.7 Разложување на сигналот со 2Д брза трансформација на бранче.



Слика 3.8 Пример за двонивовска брза трансформација на бранче.

Со ова е илустрирана постапката на примена на FWT на две нивоа. Оваа постапка може да се повторува и на повеќе нивоа, а последно можно ниво е нивото $log_2 M$, каде M е минимум од ширината и висината на сликата. Апроксимацијата која ќе се добие на последното ниво е средниот интензитет на целата слика.

Фамилии на бранчиња

Како што беше споменато претходно, постојат различни фамилии на бранчиња кои имаат различни карактеристики. За различни типови на сигнали, можат да бидат погодни различни фамилии на бранчиња. Во истражувањата во овој докторски труд од интерес ни се само фамилиите на бранчиња кои можат да се користат за дискретна анализа.

Haar бранчето е најстарото и наједноставно бранче, и е предложено од Alfréd Haar. Скалирачката и функцијата на бранче на Haar бранчето се дефинирани како

$$\varphi(t)=\begin{cases}1 \;;\; \text{за}\;\; 0\le t<1\\ 0 \;;\; \text{останато}\end{cases} \qquad \psi(t)=\begin{cases}1 \;;\; \text{за}\;\; 0\le t<1/2\\ -1 \;;\; \text{за}\;\; 1/2\le t<1\\ 0 \;;\; \text{останато}\end{cases}. \tag{3.21}$$

Haar бранчето е симетрично, и должината на неговиот филтер е 2.

Daubechies бранчињата се несиметрични и формираат фамилија која е една од почесто користените фамилии на бранчиња. Карактеристично за нив е тоа што се самослични. Во оваа фамилија припаѓаат повеќе бранчиња кои се разликуваат според должината на филтрите која може да биде од 2 до 20 (Daubechies 2 - Daubechies 20). Должината на филтерот е двапати поголема од редот на бранчето. Во оваа докторска дисертација се користат Daubechies 2, Daubechies 3 и Daubechies 4 бранчињата.

Symlet бранчињата бележат многу сличности со Daubechies бранчињата. Овие бранчиња се погодни за анализа на симетрични или приближно симетрични сигнали. Тие се симетрични и должината на нивните филтри е двапати поголема од редот на бранчето. Во оваа докторска дисертација се користи Symlet 4 бранчето кое има филтер со должина 8.

Coiflet бранчето е дизајнирано така да се обезбеди поголема симетричност од Daubechies бранчињата. Должина на неговите филтри е шест пати поголема од редот на бранчињата. Во оваа докторска дисертација се користи Coiflet 1 бранчето чии филтри имаат по 6 коефициенти.

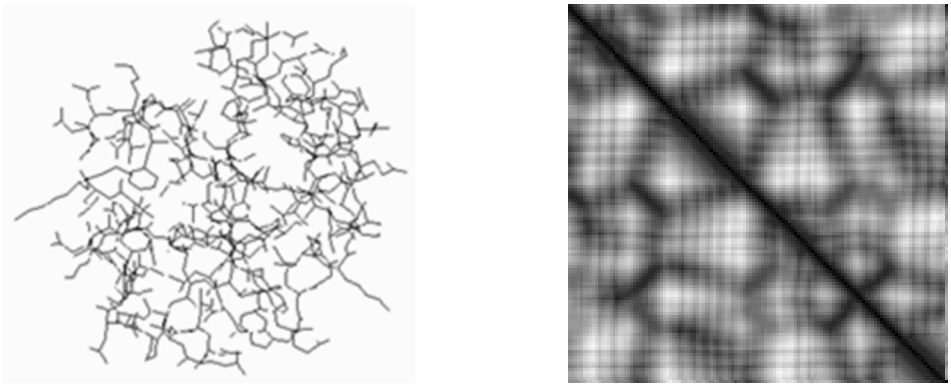
Постојат уште голем број други фамилии на бранчиња, но тие или не се погодни за дискретна трансформација на бранчиња, или пак имаат карактеристики кои не се релевантни за проблемот кој се решава во оваа докторска дисертација. Погоре накратко беа опишани најважните карактеристики на фамилиите на бранчиња чии претставници се користат во ова истражување.

При тоа избегнати се бранчиња со поширок филтри бидејќи нивното пресметување е значително временски позахтевно. Повеќе информации за трансформацијата на бранчиња и различните фамилии на бранчиња може да се најдат во (141), (142).

Екстракција на протеинските дескриптори базирани на бранчиња

Слично како и кај протеинскиот дескриптор кој е базиран на интерполација на скелетот на протеинот, така и кај протеинските дескриптори базирани на бранчиња предвид се земаат само Cα атомите кои го формираат скелетот на протеинот. На овој начин протеините не се разгледуваат како 3Д волуменски тела, туку како 3Д криви кои одговараат на нивните скелети. Потоа, од оваа крива може да се генерира дескриптор користејќи ги координатите на Cα атомите.

Нека скелетот на протеинот го претставиме како множество $S=\{C_1, C_2,\ldots, C_n\}$ кое се состои од сите Cα атоми подредени во редоследот кој е дефиниран со примарната структура на протеинот. Од ова множество може да се пресмета матрицата на растојанија D, така што елементот $D_{i,j}$ е Евклидовото растојание помеѓу i-тиот и j-тиот Cα атом. На Слика 3.9 е прикажан скелетот на протеинот 1tk5 ланец B и неговата матрица на растојанија. Матрицата на растојанија е прикажана како слика на која помалите растојанија се прикажани со потемна нијанса, додека поголемите растојанија се означени со посветла нијанса.



Слика 3.9 Скелетот на протеинот 1tk5 ланец B и неговата матрица на растојанија.

Матрицата на растојанија е квадратна матрица, односно има ист број на редици и колони. За матрицата на растојанија важи дека е симетрична ($D_{i,j} = D_{j,i}$), а елементите по главната дијагонала се еднакви на нула ($D_{i,i} = 0$). Во матрицата на растојанија α-хеликсите и паралелните ß-рамнини се појавуваат како темни региони кои се паралелни со главната дијагонала, а антипаралелните ß-рамнини се појавуваат како темни региони кои се нормални на главната дијагонала (73).

Дескрипторот кој ќе се генерира од матрицата на растојанија е инваријантен на транслација и ротација, бидејќи доколку даден протеин се транслира и/или ротира во просторот, тогаш Евклидовите растојанија помеѓу Cα атомите нема да се променат. Во продолжение ќе биде покажано дека дескрипторот генериран од матрицата на растојанија е инваријантен и на скалирање. Нека $C_i=(X_i, Y_i, Z_i)$ и $C_j=(X_j, Y_j, Z_j)$ се два Cα атоми од испитуваната протеинска структура. По скалирање на оваа протеинска структура за фактор на скалирање *fs*, C_i и C_j атомите ќе се најдат на позиција $C^t_i=(X^t_i, Y^t_i, Z^t_i)$ и $C^t_j=(X^t_j, Y^t_j, Z^t_j)$, а елементите на матрицата на растојанија за скалираната протеинска структура ќе бидат $D^t_{i,j}= fs\ D_{i,j}$. Од тука се забележува дека матрицата на растојанија на скалираниот протеин ги содржи истите детали како и матрицата на растојанија на нескалираниот протеин, само што истите детали се со помал или поголем интензитет во зависност од факторот на скалирање.

За да може врз оваа матрица на растојанија да се примени анализа на бранчиња, мора нејзините димензии да бидат број кој е степен од два. Димензиите на матрицата на растојанија се еднакви на бројот на Cα атоми *n*, а протеинските ланци имаат различен број на Cα атоми кој најчесто не е степен од два ($n \neq 2^k$), па затоа пред да се примени трансформацијата на бранчиња мора матрицата да се доведе до димензии кои се степен од два. Ова може да се реши со нерамномерна интерполација на скелетот на протеинот при што се врши додавање или бришење на интерполациски точки се додека бројот на интерполациски точки не биде степен од два, како во (73). Друг можен пристап е да се генерира матрицата на растојание директно од оригиналниот скелет на протеинот, а потоа врз добиената матрица на растојанија да се применат техники за скалирање на слики. На овој начин сликата се скалира до најблискиот број кој е степен од два и е поголем од бројот на Cα атоми, а при тоа се задржува оригиналната разместеност на Cα атомите, што не е случај доколку проблемот се решава преку нерамномерна интерполација на скелетот на протеинот. Во овој докторски труд се користи вториот пристап каде матрицата на растојанија добиена од оригиналниот протеински скелет се скалира така што да се добие матрица со димензии кои се степен од два. Матриците на растојанија се претставуваат како слики со нијанси на сиво. Вредностите на матрицата на растојанија се нормализираат, така што се доведуваат во опсегот [0, 255].

По скалирањето на матрицата на растојанија се применува анализа на бранчиња. Нека скалираната матрица на растојанија има димензии *N*x*N*, тогаш ќе се добијат N^2 коефициенти без оглед на тоа до кое ниво се врши декомпозицијата. Доколку протеинскиот дескриптор ги содржи сите овие коефициенти, тогаш не само што мемориските потреби се зголемуваат, туку и времето за споредба помеѓу два протеини ќе биде значително поголемо. За таа цел, прво се квантизираат коефициентите на 1 за голем позитивен коефициент, и на -1 за голем негативен коефициент со што се зголемува дискриминаторната моќ на дескрипторот. Потоа предвид се земаат само првите

k апроксимативни коефициенти кои се наоѓаат во горниот лев агол кои носат најгенерални информации за сигналот. Во овој докторски труд ќе бидат прикажани резултатите од пребарувањето на протеински структури користејќи различен број на апроксимативни коефициенти.

Во овој докторски труд се користат следните бранчиња: Haar, Daubechies2, Daubechies3, Daubechies4, Symlet4 и Coiflet1. Во Табела 3.2 се прикажани импулсните одзиви на бранчињата кои се користат во ова истражување.

Бранче	**Нископропусен филтер**	**Високопропусен филтер**
Haar	$h_{\varphi} = [1/\sqrt{2}, 1/\sqrt{2}]$	$h_{\psi} = [1/\sqrt{2}, -1/\sqrt{2}]$
Daubechies2	h_{φ} =[-0.1294, 0.2241, 0.8365, 0.4830]	h_{ψ} =[-0.4830, 0.8365, -0.2241, -0.1294]
Daubechies3	h_{φ} =[0.0352, -0.0854, -0.1350, 0.4599, 0.8069, 0.3327]	h_{ψ} =[-0.3327, 0.8069, -0.4599, -0.1350, 0.0854, 0.0352]
Daubechies4	h_{φ} =[-0.0106, 0.0329, 0.0308, -0.1870, -0.0280, 0.6309, 0.7148, 0.2304]	h_{ψ} =[-0.2304, 0.7148, -0.6309, -0.0280, 0.1870, 0.0308, -0.0329, -0.0106]
Symlet4	h_{φ} =[-0.0758, -0.0296, 0.4976, 0.8037, 0.2979, -0.0992, -0.0126, 0.0322]	h_{ψ} =[-0.0322, -0.0126, 0.0992, 0.2979, -0.8037, 0.4976, 0.0296, -0.0758]
Coiflet1	h_{φ} =[-0.0157, -0.0727, 0.3849, 0.8526, 0.3379, -0.0727]	h_{ψ} =[0.0727, 0.3379, -0.8526, 0.3849, 0.0727, -0.0157]

Табела 3.2 Нископропусни и високопропусни филтри на бранчињата.

Како што претходно беше кажано, трансформација на бранчиња може да се повтори до произволен број на нивоа се до најниското ниво. Во [А6] беше предложен протеински дескриптор кај кој се користи Haar бранчето при што трансформацијата на бранчиња се повторува до најниското ниво на кое се гледа најгрубата апроксимација на сигналот. Потоа, во [А31] покрај Haar бранчето беа применети и други бранчиња со цел да се овозможи попрецизно пребарување на протеинските структури. Во [А31] дополнително беше направена споредба со постоечкиот протеински дескриптор предложен во (73). Во (73), прво матрицата на растојанија се скалира во матрица 128x128, а потоа се применува Haar трансформација до четврто ниво, и на крај се земаат 36 апроксимативни коефициенти.

Метрика за растојание

Во продолжение ќе биде опишано како се одвива споредувањето на две протеински структури користејќи ги протеинските дескриптори базирани на бранчиња. Нека Q_1 и Q_2 се дескрипторите кои се добиваат после декомпозиција на матриците на растојанија на двете протеински структури кои се споредуваат. Растојанието помеѓу овие протеински дескриптори $d(Q_1, Q_2)$ се наоѓа како инверзна вредност од сумата на сличностите помеѓу коефициентите кои се на иста позиција (i,j) во двете матрици

$$d(Q_1, Q_2) = 1 \Big/ \sum_i \sum_j W(i,j)\,eval(i,j) \tag{3.22}$$

$$eval(i,j) = \begin{cases} 1; & Q_1(i,j) = Q_2(i,j) \\ 0; & Q_1(i,j) \neq Q_2(i,j) \end{cases},$$

каде што сличноста помеѓу два коефициенти кои се на иста позиција е 1 доколку коефициентите имаат иста вредност, а во спротивно сличноста е 0. Коефициентите кое се позиционирани поблиску до позицијата (0,0) носат погенерални информации за разлика од поодалечените коефициенти. Ова значи дека различни коефициенти имаат различна значајност, па затоа им се доделуваат различни тежини $W(i,j) = 1 / \min(\max(i,j), 5)$ така што коефициентите на помалите координати добиваат поголема тежина за разлика од коефициентите на поголемите координати. Со ова коефициентите кои се сместени во горниот лев агол добиваат тежини од 1 до 0.25, а останатите коефициенти кои се сместени на координати (i,j) каде i или j е поголемо или еднакво на 5 добиваат исти тежини од 0.2. Во процесот на екстракција на дескрипторите беа земени првите k апроксимативни коефициенти со што беа генерирани дескриптори со должина k. Во процесот на споредба треба секој коефициент од едниот дескриптор да се спореди со секој коефициент од другиот дескриптор, па од тука пресметувањето на ова растојание е со линеарна комплексност $O(k)$.

3.4. Метод за споредба на протеински структури преку порамнување на матрици на растојанија

Постојат многу методи за споредба на протеински структури кои вршат порамнување на матриците на растојанија, како на пример DALI (19) и MatAlign (70) методите. Нека $A[N_A][N_A]$ и $B[N_B][N_B]$ се матриците на растојанија на двете протеински структури кои сакаме да ги споредиме, така што $N_A<N_B$. MatAlign методот (70) користи пристап со примена на груба сила (brute force) со кој се обидува секоја редица од матрицата A да ја порамни со секоја редица од матрицата B. Од друга страна познатиот DALI метод (19) ги дели матриците на растојанија во

[6x6] подматрици, така што потоа секоја [6x6] подматрица од матрицата *A* се порамнува со секоја [6x6] подматрица од матрицата *B*. Двата методи имаат висока комплексност. Во [A31] беше предложен Matrix Alignment by Sequence Alignment within Sliding Window (MASASW) методот кој врши порамнување на матрици преку порамнување на секвенци (редици) во рамки на даден лизгачки прозорец. Во продолжение ќе биде опишан MASASW методот за споредување на две протеински структури преку порамнување на нивните матрици на растојанија.

Матрицата на растојанија се нормализира така да нејзините елементи примаат вредности во интервалот [0, 255]. Бидејќи протеинските ланци има различен број на Cα атоми, почнувајќи од десетици атоми, па се до неколку илјади Cα атоми, за различен ланец ќе се добијат матрици на растојанија со различни димензии. Па затоа првиот проблем кој треба да се реши е справување со споредба на два протеински ланци кои имаат различен број на Cα атоми. За таа цел се генерираат матрици на растојанија на различни размери [16x16], [32x32],…, [1024x1024] користејќи класични техники за скалирање на 2Д слики. При споредба на две протеински структури чии матрици *A* и *B* се со димензии N_AxN_A и N_BxN_B се порамнуваат матриците *A* и *B* на секој размер *s* така што $nA \le s \le nB$, каде $N_A \le 2^{nA}$ и $N_B \le 2^{nB}$. На овој начин двете матрици што се порамнуваат на даден размер *s* се со иста големина $n = 2^s$.

Во литературата постојат методи за одредување на хомологни протеини преку порамнување на секвенци користејќи лизгачки прозорец. Кај MASASW методот се користи слична идеа и се применува во 2Д простор за се порамнат 2Д матрици. За секоја редица се дефинира опсег во кој ќе се прави порамнување. Овој опсег се дефинира преку лизгачкиот прозорец. На овој начин се избегнува комплетното порамнување со сите редици кое се прави со MatAlign методот (70), каде што секоја редица од матрицата *A* се порамнува со секоја редица во матрицата *B*. Споредбата се одвива во два чекори. Во првиот чекор за секоја редица *I* од матрицата *A* се дефинира контекстен прозорец со големина $2W + 1$ кој ги содржи редиците $[I - W,\ldots, I + W]$. Редицата *I* се порамнува со редиците од матрицата *B* чии индекси се во контекстниот прозорец за редицата *I*. Во вториот чекор, со користење на алчен пристап се пресметува вкупната сличност на двете матрици *A* и *B*.

Во првиот чекор целта е да се одреди должината на најдолгата заедничка подсеквенца што одговара на сличноста на две редици $P=\{p_1,p_2,\ldots,p_n\}$ и $Q=\{q_1,q_2,\ldots,q_n\}$. Во (70) се користи Needleman-Wunsch алгоритамот за порамнување на две секвенци со динамичко програмирање. Временската и мемориската комплексност на овој алгоритам е $O(n^2)$. Сепак, со овој алгоритам може најдоброто порамнување кое ќе се најде да се однесува на многу оддалечени делови од секвенцата. Така на пример ако $P=\{p_1,p_2,..,p_{n/2},..,p_n\}$ и $Q=\{q_1,q_2,..,\ q_{n/2},p_1,p_2,..,p_{n/2}\}$, тогаш со Needleman-Wunsch алгоритамот ќе се најде порамнување со должина еднаква на $n/2$. Но, ако во редицата *Q* се сменат местата на првата и втората половина од секвенцата така што

$Q=\{p_1,p_2,..,p_{n/2},q_{n/2+1},..,q_n\}$, тогаш повторно ќе се добие истата сличност на секвенците иако во првиот случај порамнувањето е за различни делови во секвенците, додека во вториот случај порамнувањето е за исти делови од секвенците. Имено, во првиот случај овие матрици се репрезенти на протеини кои имаат одредени локални сличности, но глобално скелетите на протеините се комплетно различни. Со тоа овој алгоритам ќе даде погрешна информација дека скелетите на двата протеини се слични, што е неприфатливо.

За порамнување на две секвенци (во овој случај редици) P и Q се користи динамичко програмирање, при што се обидуваме секој елемент p_i од P да го порамниме со $\{q_{i-w},..,q_i,..,q_{i+w}\}$, каде w е големината на лизгачкиот прозорец. Емпириски беше одредено дека најсоодветна вредност за w е 8, па оваа вредност се користи во рамки на оваа докторска дисертација. На овој начин не само што методот ќе биде коректен, туку истовремено се намалува и временската комплексност на $O(nw)$, за разлика од DALI и MatAlign методите каде комплексноста за порамнување на две редици е $O(n^2)$. Алгоритамот за пресметување на должината на најдолгата заедничка подсеквенца (најдоброто порамнување) е презентиран со следниот псевдокод:

```
Function AlignRows (P[],Q[])
  int Last[2*w+1],Current[2*w+1]
  ZeroInitialize(Last)
  For I=1 to |P|
    For J=0 to 2*w+1
        int qIndex=calcAbsoluteIndex(I,J)
        if (abs(P[I]-Q[qIndex])<threshold)
              Current[J]=Last[J]+1
        else    Current[J]=max(Current[J-1],Last[J+1])
    EndFor
    Last=Current
  EndFor
Return max{Current}.
```

Емпириски се покажа дека 10 е најсоодветна вредност за прагот (*threshold*), па оваа вредност е користена во докторската дисертација. Алгоритамот е илустриран на Слика 3.10. Во примерот на сликата се користи лизгачки прозорец со големина еднаква на 2. Двете секвенци имаат иста должина еднаква на 7. Прагот е поставен на 0. Сивите полиња во четвртата редица се пополнети со пресметаните вредности на променливата *Last*. Нека моментално се обработува петтата редица. Кога ќе се споредуваат $P[5]$ и $Q[6]$ кои имаат иста вредност, тогаш вредноста на $Current[4]$ ќе биде $Last[4] + 1 = 3$.

P/Q	1	2	3	4	6	7	8
1	1	1	1	-	-	-	-
8	1	1	1	1	-	-	-
0	1	1	1	1	1	-	-
2	-	2	2	2	2	2	-
7	-	-	2	2	2	3	3
8	-	-	-	2	2	3	4
9	-	-	-	-	2	2	4

Слика 3.10 Одредување на најдолгата заедничка секвенца со динамичко програмирање.

Со предложениот MASASW метод, не само што е подобрена брзина на методот во однос на постоечките DALI и MatAlign методи, туку дополнително предвид се зема и просторната позиционираност на секвенците кои се порамнуваат. На овој начин се обезбедува порамнување на делови од секвенцата кои одговараат на региони чии позиции во секвенцата се блиску.

Во вториот чекор, целта е да се одреди која редица од првата матрица треба да се совпадне со која редица од втората матрица. За одредување на оптималното решение се користи динамичко програмирање. Слична техника се користи и во MatAlign методот (70), но со ова се добива комплексност $O(n^2)$. За да се намали оваа комплексност, се применува истата идеја која се користи и при порамнувањето на редиците (секвенци), односно дадена редица од првата матрица нема да се порамнува со сите редици од втората матрица туку само со редиците кои се во нејзиниот контекстен прозорец дефиниран со лизгачкиот прозорец со големина W. При тоа се користи алчен пристап за да се максимизира сличноста на матриците. Нека за редицата I од првата матрица A, се пресметани сличностите со редиците $[I\text{-}W,\ldots, I + W]$ од втората матрица B. Доколку за редицата J од матрицата B која е во контекстниот прозорец $[I\text{-}W,\ldots, I + W]$ на редицата I од матрицата A се пресметала најголема сличност, тогаш во следната итерација кога ќе се бара да се најде најдоброто совпаѓање за редицата I+1 од матрицата A, тогаш од втората матрица B предвид ќе се земат редиците почнувајќи од J+1 наместо од I+1-W. Во продолжение е даден псевдокод за алчниот пристап кој се користи за да се одреди оптималната вредност на сличноста помеѓу матриците на растојанија A и B:

```
Function AlignMatrices(A[][],B[][])
  int LastMatched=0
  int TotalSum=0
  For I=1 to |A|
    int MaxScore=0
    int MaxScoreIndex=0
    For J=max(I–W,LastMatched) to I+W
          int Score=AlignRows(A[I],B[J])
              if (Score>MaxScore)
                              MaxScore=Score
                              MaxScoreIndex=J
              EndIf
    EndFor
    TotalSum+=MaxScore
    LastMatched=MaxScoreIndex
  EndFor
Return TotalSum.
```

Вредноста која се враќа како резултат *TotalSum* е сличноста на матриците *A* и *B*. Повисока вредност означува поголема сличност помеѓу двете матрици на растојанија. Растојание помеѓу двете структури се наоѓа како 1/ *TotalSum*.

Со воведувањето на лизгачкиот прозорец, комплексноста на одредувањето на најдоброто совпаѓање на една редица од првата матрица со редиците во нејзиниот контекстен прозорец од другата матрица е намалено од $O(n^2)$ на $O(nW)$, каде што n е бројот на редици, а W е големината на лизгачкиот прозорец. Емпириски беше одредено дека W=5 е најсоодветната големина на лизгачкиот прозорец за одредување на најдоброто совпаѓање за дадена редица. Доколку предвид се земе и комплексноста за одредување на најдолгото совпаѓање помеѓу две редици, која изнесува $O(nw)$ каде што w е големината на лизгачкиот прозорец кој се користи при порамнување на две редици, тогаш вкупната комплексност на MASASW методот се добива дека е $O(n^2wW)$. Како што претходно беше кажано, емпириски беше одредено дека оптималните вредности за w и W се 8 и 5, соодветно. Параметарот n, кој е еднаков на бројот на Cα атоми, се движи од неколку Cα атоми, па сè до неколку илјади Cα атоми. Кај поголемиот број протеински ланци овој број е околу 100-200, од што може да се заклучи дека MASASW методот е повеќе од 300 пати побрз од DALI (19) и MatAlign (70) методите кои имаат комплексност од $O(n^4)$.

3.5. Евалуација на методите за пребарување на протеински структури

Во [A31] беше направена евалуација на методите за пребарување на протеински структури кои се презентирани во претходните секции. Исто така овие методи беа споредени со неколку постоечки методи за пребарување на протеински терциерни структури. Во оваа секција ќе бидат презентирани и продискутирани експерименталните резултати од споредбата помеѓу методите.

Опис на податочните множества

Податочните множества кои што ќе се користат во оваа анализа се формирани земајќи предвид дел од протеинските ланци во SCOP (Structural Classification Of Proteins) 1.75 базата на податоци (143), (144). SCOP содржи податоци за класификацијата на протеинските ланци според SCOP хиерархијата во која постојат неколку нивоа. Нивото домен е едно од поглавните нивоа во SCOP хиерархијата, па затоа најчесто методите за пребарување и класификација на протеински структури се евалуираат во однос на ова ниво.

Бидејќи постојат голем број на слични протеински ланци, при евалуација на методите за пребарување на протеински структури вообичаено предвид се зема одредено репрезентативно множество од ланци добиено со филтрирање на ланците според нивната секвентна сличност. Во оваа анализа предвид е земено множеството *PDB100* добиено со филтрирање на протеинските ланци во SCOP 1.75 кои имаат помалку од 100% секвентна сличност со користење на ASTRAL методот (145). При тоа предвид беа земени само домените кои имаат по барем два претставника. На овој начин се доби множество од 28460 протеински ланци кои припаѓаат во 5235 различни SCOP домени. При тоа дистрибуцијата на ланците во домени не е униформна.

По одбирање на репрезентативното множество *PDB100*, следно се прави поделба на протеинските ланци во множество за обука и множество за тестирање. За таа намена се користат два критериуми. Со првиот критериум во множеството за тестирање предвид се земаат протеинските ланци кои се новооткриени во SCOP 1.75, а кои не се класифицирани во претходната верзија на базата (SCOP 1.73), или пак се рекласифицирани во друг домен во верзијата SCOP 1.75. Сите останати репрезентативни протеински ланци го формираат множеството за обука. На овој начин се формира множеството *множество1*, и тоа содржи 26820 ланци за обука и 1640 ланци за тестирање. Со вториот критериум множеството *PDB100* се филтрира така што предвид се земаат протеинските ланци кои имаат помалку од 10% секвентна сличност користејќи го ASTRAL методот (145), со што се добива множеството за тестирање кое содржи 3314 протеински ланци. Преостанатите 25146 протеински ланци од множеството *PDB100* го формираат множеството за обука. Множествата за обука и тестирање добиени со вториот критериум за поделба го формираат множеството *множество2*. Бидејќи со вториот критериум како примероци за тестирање се земаат

протеинските ланци кои имаат помалку од 10% секвентна сличност, со ова се обезбедува множество за тестирање кое ги содржи најрепрезентативните протеински ланци. Во дополнителните материјали приложени со трудот [A31] дадени се детални информации за податочните множествата кои се користат во оваа анализа.

Анализите покажаа дека методите постигнуваат повисока прецизност на второто множество (*множество2*) наспрема првото множество (*множество1*). Ова и се очекуваше однапред бидејќи во второто множеството за тестирање предвид се земени најрепрезентативните протеински ланци кои помеѓу себе имаат помалку од 10% секвентна сличност. Од друга страна, кај првото множество за тестирање предвид се земени протеинските ланци кои не се класифицирани во SCOP 1.73, или во SCOP 1.75 се рекласифицирани во друг домен, што може да резултира со одбирање на протеин за тестирање кој нема доволна сличност со ниту еден протеин за обука. Затоа резултатите кои ќе бидат презентирани се добиени со второто множество (*множество2*), освен онаму каде што е кажано поинаку.

Евалуациски мерки

Евалуацијата на методите за пребарување на протеински терциерни структури е направена со користење на две евалуациски мерки кои најчесто се користат во областа на пребарување на информации, а тоа се прецизност и одзив. Нека N_q е вкупниот број на протеински ланци за тестирање. Одзивот (*recall*) покажува колкава фракција од релевантните протеински ланци за обука се пребарани (се наоѓаат помеѓу најсличните протеини). При тоа даден протеински ланец се смета дека е релевантен доколку припаѓа во истиот SCOP домен во кој припаѓа испитуваниот протеински ланец за тестирање. Во оваа анализа се одредува прецизноста на методите користејќи различна стапка на одзив од 10% па се до 100% движејќи се со чекор од 10%. Нека моментално се анализира q-тиот протеин за тестирање и нека моменталната вредност на стапката на одзив е *recall*, тогаш се одредуваат n_q**recall* најблиските ланци за обука кои припаѓаат во истиот домен d_q со ланецот за тестирање, каде што n_q е бројот на ланци за обука кои припаѓаат во SCOP доменот d_q. Нека r_q е векторот во кој протеините за обука се сортирани според нивната сличност со q-тиот протеин за тестирање врз основа на пребарувањето со методот кој се евалуира. Стапката на прецизност ($precision_{recall}$) за дадената стапка на одзив *recall* се пресметува како

$$precision_{recall} = \frac{\sum_{q=1}^{N_q} \frac{\sum_{i=1}^{n_q * recall} i / rank(r_q, i)}{n_q * recall}}{N_q}, \tag{3.23}$$

каде што функцијата $rank(r_q,i)$ го враќа рангот на i-тиот протеин за обука од SCOP доменот d_q во r_q. Рангот се одредува како редна позиција на која дадениот протеин за обука се наоѓа во векторот r_q добиен како резултат од пребарувањето. Во оваа докторска дисертација ќе бидат презентирани дијаграмите на прецизност-одзив кои ги покажуваат стапките на прецизност за различна стапка на одзив од 10% па се до 100% со чекор од 10%. Дијаграмите на прецизност-одзив кои ќе бидат презентирани се добиени за второто множество (*множество2*), освен ако не е поинаку наведено.

Бидејќи во SCOP базата постојат протеински ланци кои имаат висока сличност, а истите се класифицирани во различни SCOP домени, затоа не сите релевантни ланци за обука ќе бидат вратени на првите позиции во векторот r_q добиен како резултат од пребарувањето на q-тиот протеин за тестирање. Па затоа релевантните протеини за обука кои имаат помала сличност со испитуваниот протеин ќе имаат повисок ранг (што одговара на помала сличност) при пребарувањето, со што ќе се намали и стапката на прецизност. Покрај дијаграмите кои ги покажуваат прецизностите за различен одзив, ќе биде прикажана и просечната прецизност $p_{overall}$ која претставува средна вредност од стапките на прецизност добиени за различна стапка на одзив

$$p_{overall} = \frac{\sum_{i=1}^{10} precision_{recall=(10*i)\%}}{10}. \qquad (3.24)$$

Во продолжение ќе биде направена споредба на методите кои беа презентирани во претходните секции, и истите ќе бидат споредени со неколку постоечки методи: протеинскиот фрактален дескриптор (72), протеинскиот Haar дескриптор предложен во (73), MSVNS методот (71), MatAlign (70), DALI (19) и CE (20) методите.

Евалуација на протеинскиот воксел-базиран дескриптор

Прво беше анализирана предиктивната моќ на протеинскиот воксел-базиран дескриптор кој предвид ја зема целата протеинска структура (сите атоми). Дијаграмот на прецизност-одзив за протеинскиот воксел-базиран дескриптор е прикажан на Слика 3.16. Просечната прецизност $p_{overall}$ на *множество1* е 60.14%, додека на *множество2* се постигна прецизност од 64.84%.

Методите за пребарување на протеински структури презентирани во претходните три секции предвид го земаат само скелетот на протеинот (само Cα атомите), наместо предвид да ја земаат целата протеинска структура (сите атоми). Како што ќе биде покажано, доколку предвид се земат само Cα атомите тогаш се постигнува повисока прецизност бидејќи протеинскиот скелет ги содржи најважните карактеристики на протеинската структура преку кои се детерминира

припадноста во соодветниот SCOP домен, а останатите атоми носат информација која не е релевантна за класифицирање на протеинските структури во SCOP домени.

Евалуација на дескрипторот базиран на интерполација на протеинскиот скелет

Пред да се направи анализа на прецизноста на протеинскиот дескриптор базиран на интерполација на скелетот на протеинот, беше направена анализа на бројот на Cα атоми кај протеините во множеството *PDB100*. За таа цел матриците на растојанија беа скалирани до најблискиот поголем број кој е степен од два (види Слика 3.11), по што се утврди дека повеќето протеински ланци во множеството имаат матрици на растојанија со димензии 128x128.

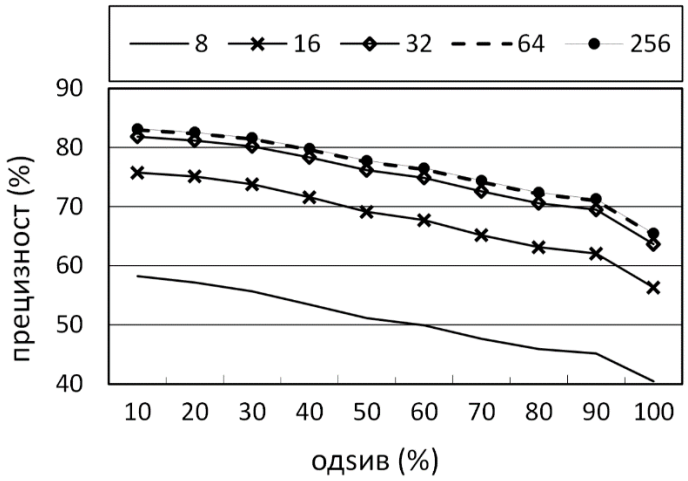
Слика 3.11 Дистрибуција на димензиите на скалираните матрици на растојанија.

Следно, беше испитана прецизноста при пребарување доколку се користи протеинскиот дескриптор базиран на интерполација на скелетот на протеинот. При тоа направена е споредба на двата пристапи за интерполација на протеинскиот скелет. Исто така предвид беа земени двете метрики на растојание, L_1 и L_2 нормите. Беше испитано влијанието на должината на дескрипторот N врз точноста на пребарувањето. Резултатите добиени за просечната прецизност $p_{overall}$ се прикажани во Табела 3.3. Резултатите дадени во Табела 3.3 покажуваат дека рамномерната интерполација е посоодветна за екстракција на најрелевантните карактеристики на протеинската структура. Вреди да се напомене дека разликата на просечната прецизност $p_{overall}$ на дескрипторите добиени со рамномерна интерполација на скелетот со 256 и 32 интерполациски точки е помала од 2% (со L_2 норма). Ова е многу важно бидејќи времето потребно за пребарување линеарно зависи од должината на дескрипторот, па така со намалување на должината на дескрипторот од 256 на само 32 карактеристики, времето може да се намали за осум пати, а за сметка на тоа прецизноста ќе се намали за помалку од 2%. Може да се забележи дека доколку се користи

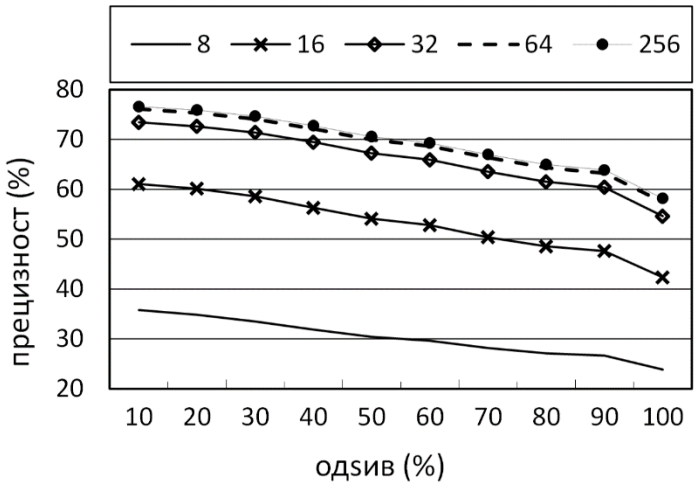
рамномерна интерполација, тогаш повисока прецизност се постигнува со користење на L_2 нормата, додека со нерамномерна интерполација повисока прецизност се добива кога се користи L_1 норма како мерка за растојание. На Слика 3.12 и Слика 3.13 дадени се дијаграмите на прецизност-одзив за протеинскиот дескриптор базиран на интерполација на скелетот со користење на рамномерна интерполација во комбинација со L_2 норма, и нерамномерна интерполација во комбинација со L_1 норма.

Метрика за растојание	**Должина на дескрипторот (*N*)**	**Рамномерна интерполација**	**Нерамномерна интерполација**
L_2 норма	256	76.49	68.34
	128	76.35	68.29
	64	76.26	67.81
	32	74.88	64.90
	16	67.96	52.23
	8	50.47	29.24
L_1 норма	256	76.43	69.36
	128	76.40	69.35
	64	76.10	68.75
	32	74.40	65.99
	16	67.32	53.17
	8	49.99	30.18

Табела 3.3 Просечна прецизност $p_{overall}$ (%) на протеинскиот дескриптор базиран на интерполација на скелет на протеинот.



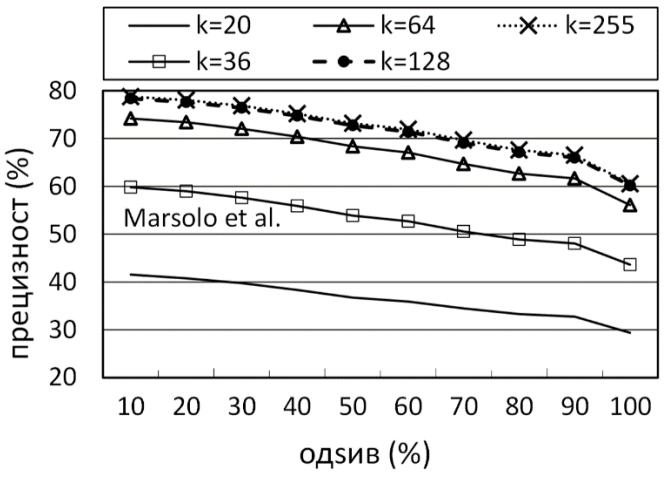
Слика 3.12 Дијаграм на прецизност-одзив за протеинскиот дескриптор базиран на рамномерна интерполација на скелетот на протеинот со користење на L_2 норма и дескриптори со различна должина *N*.



Слика 3.13 Дијаграм на прецизност-одзив за протеинскиот дескриптор базиран на нерамномерна интерполација на скелетот на протеинот со користење на L_1 норма и дескриптори со различна должина N.

Евалуација на протеинските дескриптори базирани на бранчиња

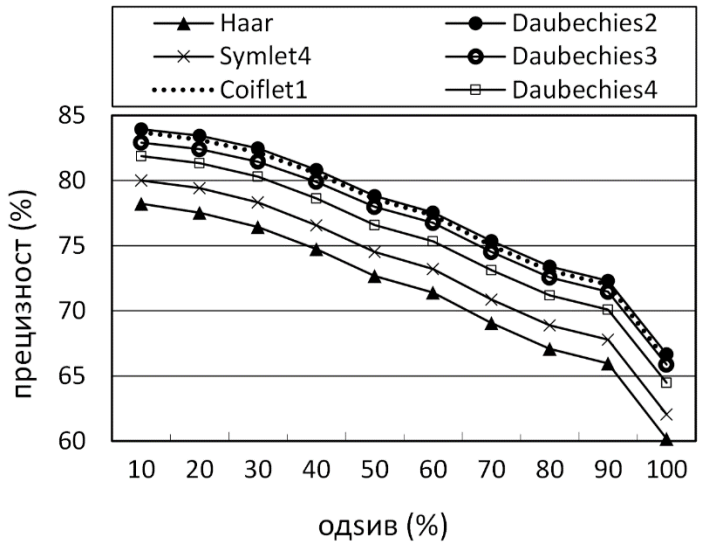
Следно, беше направена евалуација на протеинските дескриптори базирани на бранчиња. Прво беше испитано како бројот на апроксимативни коефициенти (k = 20, 65, 128, 255) влијае врз прецизноста на пребарување. При тоа беше користено Haar бранчето за да се направи декомпозиција до последното можно ниво [A6]. Резултатите од евалуацијата на Haar дескрипторот користејќи различен број на апроксимативни коефициенти k се дадени на Слика 3.14.



Слика 3.14 Дијаграм на прецизност-одзив за дескрипторот базиран на Haar бранчето со користење на различен број на апроксимативни коефициенти (k).

Резултатите покажуваат дека со зголемување на бројот на коефициенти од 20 до 64, прецизноста при пребарувањето се зголемува, но потоа со додавање на дополнителни коефициенти се забележува мало нараснување на прецизноста. Доколку предвид се земат премногу коефициенти (k>255), тогаш прецизноста опаѓа бидејќи покрај генералните карактеристики на протеинската структура дополнително предвид се земаат и коефициенти кои опишуваат некои локални детали за структурата. На тој начин две протеински структури кои имаат слични локални сегменти ќе се одреди дека имаат висока сличност иако нивните структури глобално не се слични. Дополнително предложениот Haar дескриптор кај кој декомпозицијата се прави до последно ниво [A6] беше спореден со Haar дескрипторот предложен од Marsolo et al. (73) кај кој декомпозицијата се прави до четвртото ниво и дескрипторот се формира од 36 апроксимативни коефициенти. Резултатите од оваа споредба се дадени на Слика 3.14, од каде може да се забележи дека 36-те апроксимативни коефициенти кои го формираат Haar дескрипторот предложен од Marsolo et al. (73) не се доволни за да ги претстават сите релевантни карактеристики на протеинската структура.

Следно, беше направена анализа за тоа која тип на бранче е најпогодно за екстракција на најрелевантните својства на протеинската терциерна структура. Анализите покажаа дека за Daubechies2, Daubechies3 и Symlet4 бранчињата најдобро е предвид да се земат 150 коефициенти, додека за Daubechies4 и Coiflet1 најдобро е да се земат 200 коефициенти. Во понатамошните анализи презентирани во оваа секција бројот на апроксимативни коефициенти кои го формираат дескрипторот е поставен на 150. На Слика 3.15 и Табела 3.4 се прикажани резултатите добиени за различните дескриптори базирани на бранчиња.



Слика 3.15 Дијаграм на прецизност-одзив за протеинските дескриптори базирани на бранчиња користејќи различен тип на бранче.

Бранче	***k***	$p_{overall}$**(%)**
постоечкиот Haar (Marsolo et al.)	36	52.99
предложениот Haar (декомпозиција до последно ниво)	150	71.33
Daubechies2	150	77.48
Daubechies3	150	76.59
Daubechies4	150	75.30
Symlet4	150	73.13
Coiflet1	150	77.18

Табела 3.4 Просечна прецизност $p_{overall}$ (%) на протеинските дескриптори базирани на бранчиња користејќи различен тип на бранче.

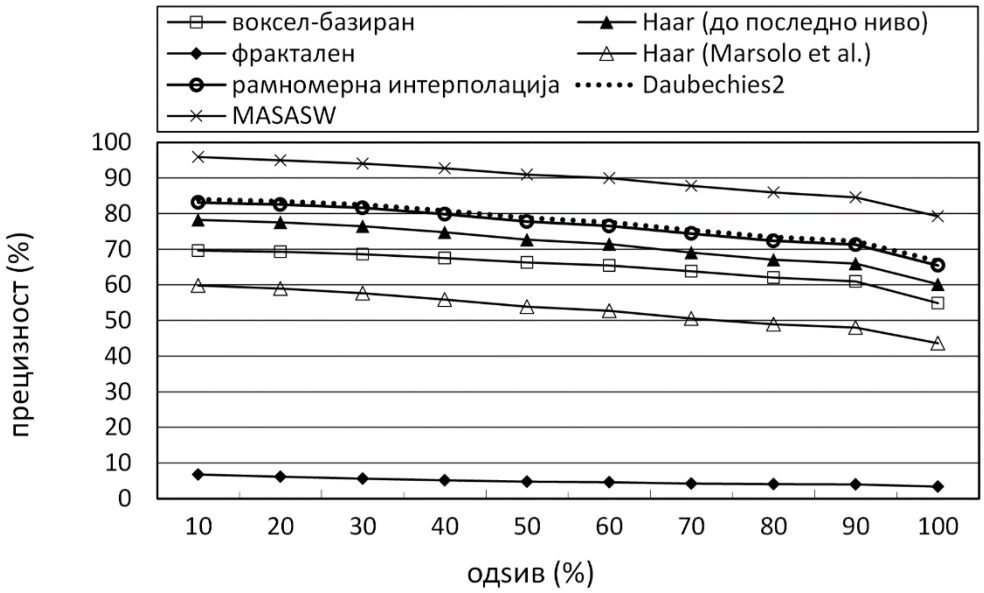
Од резултатите од Табела 3.4 може да се заклучи дека со дескрипторот базиран на Daubechies2 бранчето се овозможува најпрецизно пребарување на протеинските структури кои припаѓаат во ист SCOP домен, при што се добива просечна прецизност $p_{overall}$ од 77.48%. Исто така и другите бранчиња од Daubechies фамилијата, како и Coiflet1 бранчето постигнуваат $p_{overall}$ поголемо од 75.30%. Symlet4 дескрипторот има помала прецизност од овие дескриптори, но тој е попрецизен од Haar дескрипторот. Од ова може да се заклучи дека дескрипторите базирани на бранчиња предложени во [A31] постигнуваат поголема прецизност од Haar дескрипторот кој беше предложен во [A6]. Сепак, тука треба да се напомене дека прецизноста која се анализира во оваа секција се однесува на тоа дали протеинските структури кои припаѓаат во ист SCOP домен ќе имаат висока сличност или не. Но ако овие дескриптори се применат за предвидување на протеински функции наместо за класификација во SCOP домени, тогаш може да се добијат поинакви резултати.

Евалуација на предложениот метод базиран на порамнување на матрици на растојанија

Следно, беше направена евалуација на MASASW методот. Бидејќи овој метод е значително поспор од останатите методи кои се презентирани во ова поглавје, затоа оваа споредба е направена на првото множество (*множество1*). Резултатите од оваа анализа се дадени на Слика 3.16 и во Табела 3.5. Во оваа анализа не се земени предвид DALI, CE, и MatAlign методите бидејќи се премногу спори, па потребно е многу време за да се добијат резултати на големо податочно множество. Како што може да се забележи MASASW методот постигнува прецизност поголема од 79% за стапка на одзив 10% (Слика 3.16), а $p_{overall}$ е 89.61%.

Дополнително беше направена споредба на перформансите на MASASW со перформансите на DALI (19), CE (20) и MatAlign (70) методите. Бидејќи DALI, CE и MatAlign се многу спори, затоа оваа анализа беше направена на податочното множество од 200 протеински ланци кое е употребено во (70). Во Табела 3.6 е прикажано просечното време потребно за споредба на еден

протеински ланец со сите (200) протеински ланци во податочното множество. Може да се забележи дека MASASW е многу побрз од DALI, CE и MatAlign, што и е за очекување бидејќи MASASW се обидува да ја порамни секоја редица од првата матрица на растојанија само со редиците од втората матрица на растојанија кои се во нејзиниот контекстен прозорец. Подоцна ќе биде направена споредба на прецизноста на MASASW, DALI и CE методите на поголемо податочно множество.



Слика 3.16 Дијаграм на прецизност-одзив за методите за пребарување на протеински структури кои се земени во предвид во ова истражување (освен DALI, CE, MatAlign и MSVNS).

метод	***множество1***	***множество2***	***множество3***
фрактален	4.44	4.85	4.90
воксел-базиран	60.14	64.84	77.82
MSVNS	NA	NA	79.39
постоечкиот Haar (Marsolo et al.)	52.65	53.00	90.13
Daubechies2 (k=150)	75.29	77.48	91.25
интерполација на протеинскиот скелет (рамномерна интерполација, N=256, L_2 норма)	75.68	76.49	92.95
MASASW	89.61	NA	94.60

Табела 3.5 Просечна прецизност $p_{overall}$ (%) на методите за пребарување кои се земени во предвид во ова истражување (освен DALI, CE и MatAlign) добиени на различните множества кои се користени во ова истражување.

метод	време (секунди)
DALI	14688016
CE	7344020
MatAlign	18
MASASW	5

Табела 3.6 Просечно време (секунди) потребно за споредба на еден протеински ланец со сите протеински ланци во податочното множество.

Споредба со неколку постоечки методи за пребарување на протеински структури

Во [A31] беше направена споредба и со постоечкиот MultiStart Variable Neighbourhood Search (MSVNS) метод кој е предложен во (71). Овој метод се базира на порамнување на контактни мапи при што предвид се земаат динамичките промени во соседството дефинирано со лизгачки прозорец. За таа цел, прво беа креирани контактни мапи користејќи праг на растојание од 6 Å. Во оваа анализа беше користена верзијата 1 на MSVNS методот која е временски најмалку скапа бидејќи користи лизгачки прозорец со најмала големина. Сепак овој метод е многу спор, па затоа оваа анализа беше направена на помало податочно множество. За таа цел беа одбрани 6851 протеински ланец од SCOP 1.75 кои припаѓаат во 150-те SCOP домени кои имаат најголем број на претставници. При тоа приближно ист број на протеински ланци беа земени од секој домен. На овој начин беше формирано множеството *множество3*. Потоа ова множество случајно беше поделено на множества за обука и тестирање, така што 10% (684) од протеинските ланци припаѓаат во множеството за тестирање, додека останатите 90% (6167) припаѓаат во множеството за обука. При ова беше внимавано за униформната дистрибуција на протеинските ланци по домени да се задржи во множествата за обука и тестирање. Во трудот [A31] дадени се детални информации за ланците кои влегуваат во составот на множествата за обука и тестирање од *множество3*. Покрај тоа што MSVNS методот е временски и мемориски скап, од резултатите дадени во Табела 3.5 може да се заклучи дека просечната прецизност на MSVNS методот е 79.39% што е помало од прецизноста на поголем дел од останатите методи.

Во (72) протеинските структури се споредуваат врз основа на нивната волуменска фрактална димензија и протеинскиот радиус. Фракталната димензија го покажува степенот на самосличност на протеинскиот скелет. Бидејќи фракталната димензија не е доволно дискриминаторна карактеристика, затоа предвид се зема и протеинскиот радиус кој се одредува како радиус на најмалата сфера во која може да се смести протеинската структура. На Слика 3.16 е прикажан дијаграмот на прецизност-одзив на протеинскиот фрактален дескриптор каде предвид е земен и протеинскиот радиус, а како мерка за растојание се користи L_1 нормата. Просечната прецизност $p_{overall}$ доколку предвид се земе и протеинскиот радиус е 4.44% на *множество1* и 4.85% на

множество2, а доколку предвид се земе само фракталната димензија како единствена карактеристика во дескрипторот $p_{overall}$ изнесува 1.8% на *множество1* и 1.54% на *множество2*. При пресметување на самосличноста на протеинската структура, предвид може да се земат сите атоми кои влегуваат во составот на протеинот, но на тој начин прецизноста е помала (4.22% на *множество1* и 4.36% на *множество2*).

На Слика 3.5 се дадени резултатите од споредбата на методите користејќи го *множество1*, додека во Табела 3.5 се прикажани резултатите добиени користејќи ги *множество1*, *множество2* и *множество3*. Со NA е означено дека со дадениот метод не се добиени резултати врз соодветното множество. Може да се забележи дека MASASW постигнува највисока прецизност. Потоа следат дескрипторот базиран на рамномерна интерполација на скелетот и предложените дескриптори базирани на бранчиња кај кои се врши декомпозиција до последното ниво. Потоа следат воксел-базираниот дескриптор и MSVNS методот, додека фракталниот дескриптор бележи најлоши резултати ($p_{overall}$<5%).

Споредба со DALI и CE методите

Во [A31] дополнително беше направена споредба на предложените методи со DALI (19) и CE (20) методите. Бидејќи DALI и CE се пресметковно скапи, затоа беа употребени готовите резултати, кои се достапни на нивните веб страни, добиени од споредбата на протеински структури. Податочното множество *множество4* кое се користи во оваа анализа се формира како подмножество од ланците кои ги има во *множесто2* и истовремено ги има во достапните резултати од пребарувањето со DALI и CE методите. Добиеното множество има 484 ланци за тестирање и 8523 ланци за обука. Во дополнителните материјали на трудот [A31] дадена е листа на протеински ланци кои се содржат во множествата за обука и тестирање. Бидејќи во резултатите добиени со DALI и CE методите има информации само за најблиските соседи кои имаат сличност поголема од одреден праг, затоа во оваа анализа се користи модифицирана верзија на равенството (3.23) за да се одреди прецизноста на методите

$$precision = \frac{\sum_{q=1}^{N_q} \frac{\sum_{i=1}^{\min(DALI_q * CE_q)} i / rank(r_q, i)}{\min(DALI_q * CE_q)}}{N_q}, \tag{3.25}$$

каде $DALI_q$ и CE_q се бројот на соседи за кои со DALI и CE методите е добиено дека со q-тиот протеин за тестирање имаат сличност поголема од 2 (за DALI) и 3.8 (за CE), соодветно. Во Табела 3.7 дадени се резултатите кои се добиени за $p_{overall}$ користејќи го *множество4*.

Метод	$p_{overall}$ (%)
фрактален	1.42
воксел-базиран	33.68
рамномерна интерполација на протеинскиот скелет	56.81
нерамномерна интерполација на протеинскиот скелет	47.35
предложениот Haar (декомпозиција до последно ниво)	55.60
постоечкиот Haar (Marsolo et al.)	36.68
Daubechies2	60.58
MASASW	71.87
DALI	82.55
CE	55.41

Табела 3.7 Просечна прецизност $p_{overall}$ (%) на методите за пребарување на протеински структури кои се земени во предвид во ова истражување.

Како што беше покажано во Табела 3.6, MASASW е значително побрз од DALI и CE методите, а MASASW постигнува значително повисока прецизност од CE методот. Дури и некои од останатите методи кои се значително побрзи од MASASW, постигнуваат повисока прецизност од CE методот. DALI методот постигна највисока прецизност во пребарувањето, но тој е пресметковно многу поскап.

3.6. Проширување на протеинскиот дескриптор базиран на рамномерна интерполација на скелетот на протеинот со дополнителни карактеристики

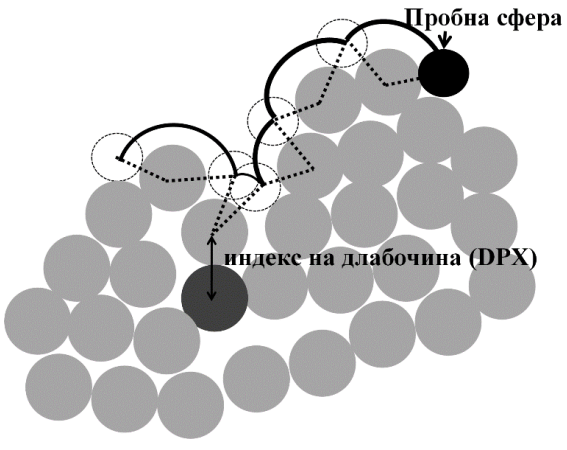
Во [A6] беше воведен протеинскиот дескриптор базиран на рамномерна интерполација на протеинскиот скелет. Потоа, во [A17] и [A25] овој дескриптор беше проширен со дополнителни карактеристики со цел да се обезбеди повисока прецизност при пребарувањето. За таа цел, прво се врши рамномерна интерполација на протеинскиот скелет, како што беше опишано во секција 3.2. Потоа за секоја интерполациска точка покрај тоа што предвид се зема оддалеченоста од таа точка до центарот на маса, дополнително се додаваат и неколку карактеристики на амино-киселинскиот остаток до кој е најблиску дадената интерполациска точка. За таа цел се идентификува амино-киселинскиот остаток во кој припаѓа Cα атомот кој е најблиску до интерполацискатa точка, а потоа на интерполацискатa точка и се придружуваат неколку дополнителни карактеристики на тој амино-киселински остаток. Во [A17] дескрипторот се проширува со карактеристиките дофатлива површина (Accessible Surface Area - ASA) (146), (147), релативна дофатлива површина (Relative ASA - RASA) (148) и хидрофобичност (149), а потоа во [A25] дополнително дескрипторот беше проширен и со карактеристиките индекс на длабочина (depth index - DPX) (150) и индекс на испакнатост (protrusion index - CX) (151). Со додавањето на овие дополнителни карактеристики целта е да се извлече 3Д дескриптор кој покрај тоа што ќе опишува како скелетот на протеинот се приближува и оддалечува во однос на

центарот на маса, туку дополнително ќе опише како скелетот се приближува и оддалечува во однос на површината, како и да се опише густината на регионот во кој се наоѓа даден сегмент од протеинскиот скелет. Во продолжение ќе биде опишан процесот на екстракцијата на овие карактеристики за секој амино-киселински остаток чии својства треба да се вклучат во дескрипторот.

3.6.1. Екстракција на карактеристиките на амино-киселинските остатоци

Бидејќи еден амино-киселински остаток се состои од неколку атоми, затоа прво се врши екстракција на карактеристиките за секој атом, а потоа се пресметуваат карактеристиките за целиот амино-киселинскиот остаток врз основа на карактеристиките на поединечните атоми кои влегуваат во неговиот состав.

Една од карактеристиките на атомите која често се користи за предвидување на регионите каде што може да настане интеракција со друга протеинска структура е дофатливата површина на атомот (Accessible Surface Area - ASA) (146), (147). ASA е воведена од Lee и Richards (146) и се пресметува со користење на алгоритамот со тркалечка топка ("rolling ball") (147), каде со користење на пробна сфера со предефиниран радиус се одредува колкав дел од површината на даденият атом може да биде дофатен од пробната сфера. Пресметаната дофатлива површина вообичаено се изразува во единица мерка $Å^2$. Радиусот на пробната сфера најчесто се поставува на 1.4 Å, што е еднакво на радиусот на молекулата на вода. Во "rolling ball" алгоритамот пробната сфера се врти околу протеинската структура при што се прават мали дискретни поместувања, види Слика 3.17.



Слика 3.17 Екстракција на дофатливата површина (ASA) и индексот на длабочина (DPX).

При движењето на пробната сфера околу протеинската површина се формираат ланци кои го претставуваат делот од површината на протеинот кој може да се дофати. За секое единечно

поместување се пресметува колкава површина од атомот може да се дофати во тоа поместување. За таа цел се користи апроксимацијата воведена во (146) според која ASA_i во *i*-тото поместување се пресметува како

$$ASA_i = \frac{R}{\sqrt{R^2 - Z_i^{\,2}}} * (\Delta Z/2 + \Delta' Z) * L_i\,, \qquad \Delta' Z = \min(\Delta Z, R - Z_i)\,, \tag{3.26}$$

каде што R е радиусот на испитуваниот атом, L_i е должината на лакот направен во *i*-тото поместување, Z_i е нормалното растојание од поставеноста после *i*-тото поместување до центарот на сферата, а ΔZ е должината на завртувањето кое се прави со единечно поместување. Потоа дофатливата површина на испитуваниот атом се добива со сумирање на дофатливите површини добиени од сите поместувања. Во ова истражување се користи NACCESS (148) програмата за пресметување на дофатливата површина на атомите.

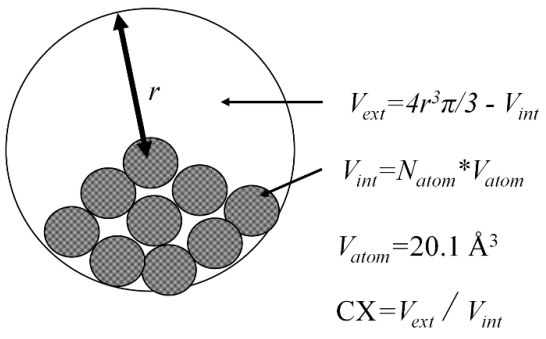
Бидејќи еден амино-киселински остаток содржи повеќе атоми, затоа неговата вкупна дофатлива површина се пресметува со сумирање на дофатливите површини на сите негови атоми. При сумирањето на дофатливите површини наместо предвид да се земат сите атоми, може предвид да се земат само атомите кои го формираат скелетот на протеинот (backbone ASA), или сите преостанати атоми (side-chain ASA). Слично, предвид може да се земат сите поларни атоми (кислород, азот, фосфор), или сите неполарни атоми (јаглерод). На овој начин може да се добијат пет различни карактеристики кои се однесуваат на дофатливата површина, а тоа се: totalASA, main-chainASA, side-chainASA, polarASA и non-polarASA, како во (152).

Различни амино-киселини се состојат од различен број на атоми, па затоа дофатливата површина ќе биде поголема за амино-киселините во кои влегуваат повеќе атоми. Затоа често се користи релативната дофатлива површина (Relative ASA - RASA), која се пресметува како однос помеѓу дофатливата површина на амино-киселинскиот остаток и стандардната дофатлива површина на тој тип на амино-киселина (148). Во ова истражување се користат стандардните дофатливи површини на амино-киселините кои се користат во (148). Слично како и за дофатливата површина, така и за релативната дофатлива површина се пресметуваат карактеристиките totalRASA, main-chainRASA, side-chainRASA, polarRASA и non-polarRASA, така што во односот предвид се зема соодветната карактеристика за дофатливата површина.

Друга важна карактеристика на амино-киселинските остатоци е индексот на длабочина (depth index - DPX) (150) која покажува колку е оддалечен даден атом во однос на најблискиот атом кој може да биде дофатен од пробната сфера (тоа се атомите кои имаат ASA поголема од нула). Длабочината на *i*-тиот атом (DPX_i) е растојанието од тој атом до најблискиот атом кој може да биде дофатен од пробната сфера, односно $DPX_i = \min(d_1, d_2, d_3, ..., d_n)$, каде што $d_1, d_2, d_3, ..., d_n$ се

растојанијата од *i*-тиот атом до атомите кои се дофатени од пробната сфера. Според ова, атомите кои можат да бидат дофатени од пробната сфера имаат $DPX_i=0$, додека за останатите атоми се добива вредност поголема од нула. Поголема вредност укажува дека атомот е подлабоко во протеинската структура, односно е подалеку од површината на протеинот. На Слика 3.17 означен е индексот на длабочина на потемнетиот атом. Оваа карактеристика носи обратна информација од карактеристиката која се зема предвид во протеинскиот дескриптор базиран на интерполација на скелетот на протеинот. Имено, Евклидовото растојание помеѓу интерполациската точка и центарот на маса покажува колку дадената точка е далеку во однос на центарот, додека DPX покажува колку таа точка е оддалечена од површината на протеинската молекула.

Друга важна карактеристика на амино-киселинските остатоци е индексот на испакнатост (protrusion index - CX). Пресметувањето на оваа карактеристика се прави користејќи ја процедурата предложена во (151). Во продолжение е дадено кусо објаснување на оваа процедура. Прво, се пресметува бројот на неводородни атоми N_{atom} кои се на растојание помало од даден предефиниран радиус r во однос на испитуваниот атом за кого се пресметува карактеристиката CX. На овој начин всушност околу атомот кој се разгледува се поставува сфера со радиус r и се одредува колку неводородни атоми има во тој дел од просторот. Во ова истражување радиусот на сферата е поставен на r=10 Å согласно препораките дадени во (151). Потоа се пресметува колкав волумен од таа сфера е зафатен од протеинската структура со користење на апроксимацијата $V_{int}=N_{atom}*V_{atom}$, така што бројот на неводородни атоми се множи со просечниот волумен на еден атом V_{atom}. Во ова истражување просечниот волумен на еден атом се поставува на V_{atom}=20.1 Å^3, како што е препорачано во (151). Потоа се одредува разликата помеѓу вкупниот волумен на сферата чиј радиус е r и зафатениот волуменот V_{int}, со што се добива незафатениот волумен $V_{ext}=4r^3\pi/3 - V_{int}$. На крај, индексот на испакнатост се пресметува како однос помеѓу незафатениот и зафатениот волумен, односно $CX=V_{ext} / V_{int}$. На Слика 3.18 е илустрирана процедурата за пресметување на оваа карактеристика.



Слика 3.18 Екстракција на индексот на испакнатост (CX).

Оваа карактеристика дава информација за густината на пополнетост на регионот околу испитуваниот атом. Имено, атомите кои се опкружени од многу неводородни атоми ќе имаат мал CX, додека за атомите кои се лоцирани во региони со мала густина ќе се добие висока вредност за оваа карактеристика.

Погоре беше опишано на кој начин се пресметуваат карактеристиките DPX и CX за секој поединечен атом. Потоа се пресметува просечната вредност на овие карактеристики земајќи ги предвид сите атоми кои влегуваат во составот на испитуваниот амино-киселински остаток, со што се добиваат карактеристиките avgDPX и avgCX. Исто така се пресметуваат минималниот DPX (minDPX) и CX (minCX), како и максималниот DPX (maxDPX) и CX (maxCX) земајќи ги предвид сите атоми кои го формираат амино-киселинскиот остаток. На овој начин, minDPX=0 за амино-киселинските остатоци за кои важи дека пробната сфера може да дофати барем еден од нивните атоми.

Последната карактеристика која е земена предвид е хидрофобичноста (149), и таа дава евиденција за хидрофобичните својства на амино-киселините. Оваа карактеристика е многу важна бидејќи хидрофобичните амино-киселини почесто се лоцирани во внатрешноста на протеинската структура, додека хидрофиличните амино-киселини вообичаено се лоцирани во близината или на самата површина. Хидрофобичноста на амино-киселинските остатоци беше земена предвид и во формирањето на протеинскиот воксел-базиран дескриптор опишан во секција 3.1. Во литературата може да се најдат повеќе скали со кои се дефинираа хидрофобичноста на амино-киселините. Во оваа докторска дисертација се користи скалата предложена од Kyte и Doolittle (149).

Погореопишаните карактеристики вообичаено се користат за одредување на регионите од протеинската структура каде што настанува интеракција со друга протеинска структура. Во оваа анализа, идејата е да се одреди кои од овие карактеристики се релевантни за пребарување на слични протеински терциерни структури. За секоја интерполациска точка покрај Евклидовото растојание до центарот на маса предвид се земаат и овие карактеристики, па така со индексот на длабочина DPX дополнително може да се опише како скелетот на протеинот се приближува и оддалечува во однос на површината на молекулата. Хидрофобичноста носи информација како скелетот на протеинот се приближува или оддалечува во однос на центарот на маса, додека индексот на испакнатост (CX) покажува како скелетот поминува низ погусти или поретки региони, со што на некој начин се опишуваат испакнатините на површината.

3.6.2. Евалуација на проширениот дескриптор базиран на рамномерна интерполација на скелетот на протеинот

Во секција 3.5 беше направена евалуација на методите за пребарување на протеински терциерни структури, и истите беа споредени со неколку постоечки методи за пребарување на протеински терциерни структури. Резултатите покажаа дека иако екстракцијата на дескрипторот базиран на рамномерна интерполација на скелетот на протеинот е наједноставна, сепак овој дескриптор има споредлива прецизност со дескрипторите базирани на бранчиња. Во [A17], [A25] беше предложено проширување на дескрипторот така што за секоја интерполациска точка добиена со рамномерна интерполација на протеинскиот скелет, покрај Евклидовото растојание до центарот на маса дополнително предвид се земаат и карактеристиките на најблискиот аминокиселински остаток кои беа опишани во претходната секција.

Експериментални резултати од пребарувањето на протеински структури

Во продолжение ќе бидат презентирани резултатите за дискриминаторната моќ на проширениот дескриптор. Во оваа анализа скелетот на протеинот рамномерно се интерполира со 512 интерполациски точки. За евалуација на проширениот дескриптор базиран на рамномерна интерполација на протеинскиот скелет беше користен дел од SCOP 1.73 базата на податоци (143), (144). Податочното множество се состои од 6825 случајно одбрани протеински ланци од 147 SCOP домени кои имаат најголем број на претставници. Приближно ист број на протеински ланци беа земени од секој домен. Податочното множество беше поделено така што 90% од ланците беа земени во множеството за обука, а останатите 10% го формираат множеството за тестирање. Со оваа поделба се доби множество со 6137 ланци за обука и 688 ланци за тестирање. При поделбата во множество за обука и тестирање беше запазено да се задржи униформната дистрибуција по домени во двете множествата. Евалуацијата на проширениот дескриптор се прави во однос на домен нивото во SCOP хиерархијата. Дискриминаторната моќ на дескрипторот ќе се мери користејќи ја евалуациската мерка просечна прецизност $p_{overall}$ која е дефинирана во секција 3.5 со равенството (3.24).

Прво, беше испитана дискриминаторната моќ на секоја од карактеристиките доколку се користат како единствена карактеристика во дескрипторот. Во Табела 3.8 прикажана е просечната прецизност $p_{overall}$ со користење на секоја од карактеристиките како индивидуална карактеристика. Генерално, може да се заклучи дека со користење на L_1 норма се постигнува повисока прецизност. Евклидовото растојание се покажа како најдобра карактеристика која има највисока дискриминаторна моќ за да ги разликува протеинските ланци од различни SCOP домени. RASA се покажа како подобра карактеристика од ASA, што и се очекуваше бидејќи RASA претставува

однос на ASA и стандардната ASA на амино-киселините, па со користење на RASA сите амино-киселини добиваат подеднаква тежина. Карактеристиките ASA и RASA кои носат информации само за скелетот, или само за преостанатиот дел од протеинската структура носат подетални информации од вкупната ASA и RASA. Па така, со користење на L_2 норма и backboneASA, backboneRASA, side-chainASA и side-chainRASA се постигнува повисока прецизност отколку ако се користи totalASA или totalRASA. Сепак, ова не е случај доколку L_1 нормата се користи како мерка за растојание. Со карактеристиките кои носат податоци за дофатливата површина на неполарните атоми се добива послаба прецизност отколку ако предвид се земе целата дофатлива површина, додека со дофатливата површина на поларните атоми се постигнува повисока прецизност отколку ако предвид се земе дофатливата површина на сите атоми. Со карактеристиката DPX се постигна послаба прецизност. Од трите карактеристики кои се однесуваат на DPX, maxDPX дава најдобри резултати што и беше за очекување бидејќи преку maxDPX се опишуваат вдлабнатините на површината на протеинската молекула кои не можат да бидат дофатени од пробната сфера. Карактеристиката индекс на испакнатост (CX) покажа дека има најголема дискриминаторна моќ во однос на останатите карактеристики кои се додадени во дескрипторот. Хидрофобичноста се покажа дека доколку сама се користи како единствена карактеристика, тогаш се добиваат полоши резултати отколку со останатите карактеристики. Од Табела 3.8 може да се заклучи дека од сите новододадени карактеристики со avgCX се постигнува најголема прецизност ($p_{overall}$=88.32%).

Следно, беше испитана прецизноста при пребарување доколку се користи пар од две карактеристики, при што Евклидовото растојание се комбинира со новододадените карактеристики. Експерименталните резултати од оваа споредба се дадени во четвртата и петтата колона од Табела 3.8. Генерално, со комбинирање на Евклидовото растојание со уште една карактеристика се постигна повисока прецизност отколку ако се користи Евклидовото растојание како единствена карактеристика (како во дескрипторот кој е предложен во [A6]). Единствено парот формиран од карактеристиките Евклидово растојание и хидрофобичност даде полоши резултати со користење на L_2 норма. Од оваа анализа може да се заклучи дека генерално воведувањето на дополнителните карактеристики придонесе до зголемување на дискриминаторната моќ на дескрипторот базиран на рамномерна интерполација на протеинскиот скелет. Највисока прецизност ($p_{overall}$=91.07%) се постигна со парот формиран од Евклидовото растојание и totalASA при користење на L_1 норма.

карактеристика	поединечна карактеристика		во комбинација со Евклидовото растојание	
	L_2 норма	**L_1 норма**	**L_2 норма**	**L_1 норма**
Евклидово растојание	**88.49**	**89.60**	/	/
totalASA	77.32	85.28	90.26	**91.07**
backboneASA	82.32	84.74	89.62	90.57
side-chainASA	81.84	84.13	90.51	90.97
non-polarASA	76.94	79.75	90.00	90.62
polarASA	81.10	85.90	89.97	90.81
totalRASA	83.57	87.90	89.94	90.77
backboneRASA	84.12	85.79	89.64	90.64
side-chainRASA	83.89	85.63	90.37	90.96
non-polarRASA	83.56	84.26	90.12	90.84
polarRASA	84.89	87.23	89.37	90.31
avgDPX	80.30	82.29	89.45	90.16
maxDPX	83.09	84.96	90.02	90.63
minDPX	61.68	60.58	88.81	89.61
avgCX	85.15	**88.32**	88.63	90.16
maxCX	83.28	87.44	88.64	90.17
minCX	84.26	87.75	88.62	90.14
хидрофобичност	80.15	85.09	87.62	90.66

Табела 3.8 Просечна прецизност $p_{overall}$ (%) со користење на дополнителните карактеристики.

Покрај овие анализи, во [A25] исто така беше направена споредба на прецизноста што се обезбедува со користење на дескрипторите доколку во нив се вклучат повеќе карактеристики од карактеристиките кои беа опишани во секцијата 3.6.1. Резултатите од оваа анализа се прикажани во Табела 3.9.

дополнителни карактеристики	со хидрофобичност		без хидрофобичност	
	L_2 норма	L_1 норма	L_2 норма	L_1 норма
totalASA + totalRASA	88.64	91.59	90.49	91.34
backboneASA + backboneRASA	89.36	91.76	89.94	90.80
side-chainASA + side-chainRASA	88.82	91.49	90.68	90.97
avgDPX + avgCX	88.14	91.22	89.68	90.71
maxDPX + avgCX	88.81	91.53	90.08	90.70
maxDPX + maxCX	88.80	91.53	90.09	90.72
totalASA + totalRASA + avgDPX + avgCX	90.83	91.43	90.83	91.43
totalASA + totalRASA + maxDPX + avgCX	90.69	91.48	90.84	91.48
totalASA + totalRASA + avgCX	88.89	91.74	90.76	91.40
totalASA + totalRASA + maxDPX	89.52	91.67	90.81	91.45
backboneASA + backboneRASA + avgCX	89.35	91.71	89.95	90.86
side-chainASA + side-chainRASA + avgCX	89.02	91.62	90.71	91.14

Табела 3.9 Просечна прецизност $p_{overall}$ (%) со вклучување на неколку карактеристики.

Генерално, со додавање на повеќе карактеристики се постигна повисока прецизност, но сепак мора да се внимава на тоа дека времето на пребарување линеарно расте со бројот на карактеристики. Повисока прецизност се постигна со користење на L_1 норма. Како што може да се забележи, со користење на L_2 нормата повисока прецизност се постигнува доколку хидрофобичноста не се земе предвид, додека ако се користи L_1 норма подобро е во дескрипторот да се вклучи и хидрофобичноста. Со двете мерки за растојание се постигна зголемување на прецизноста за повеќе од 2% во однос на случајот кога Евклидовото растојание се користи како единствена карактеристика.

Експериментални резултати од класификацијата на протеински структури

Во [A2] беа презентирани неколку методи за класификација на протеински структури во SCOP домени користејќи ги нивните воксел-базиран дескриптори и дескрипторите базирани на рамномерна интерполација на протеинскиот скелет. Резултатите од класификацијата на протеинските структури со различните методи за класификација се презентирани во [A4], [A5], [A7], [A9], [A11], [A14], [A15], [A16], [A18], [A19], [A20], [A22] и [A23], при што е направена детална анализа на влијанието на параметрите кои се користат кај различните класификатори. Во анализите направени во [A17], [A25] кои беа презентирани погоре се покажа дека со воведувањето на дополнителни карактеристики на амино-киселинските остатоци кои се најблиску до интерполациските точки се постигнува повисока прецизност во одредувањето на слични протеини кои припаѓаат во ист SCOP домен. Во [A26] проширениот дескриптор базиран на рамномерна интерполација на скелетот на протеинот беше применет за класификација на протеински структури во SCOP домени. Во продолжение ќе бидат презентирани резултатите од оваа анализа, и истите се објавени во [A26].

Во оваа анализа се користат следниве познати класификатори: C4.5 дрвата на одлука (153), Наивниот Баесов класификатори (Naïve Bayes) (154) и методот на *k*-најблиски соседи (*k*-nearest neighbours – *k*-nn) (155). Кај *k*-nn класификаторот се користи тежинско гласање каде што гласовите на најблиските соседи се со тежина $w = 1/distance$, каде што *distance* е растојанието помеѓу испитуваниот примерок и соодветниот најблизок сосед кој моментално гласа. За евалуација на класификациските модели беше користено истото податочно множество од 6825 протеински ланци кое се користеше во претходните анализи [A17], [A25]. Во оваа анализа, за разлика од претходната, не се прави поделба на множеството во множества за обука и тестирање, туку се користи вкрстена валидација со користење на 10 превои (10-fold cross validation). Во ова истражување се користат имплементациите на класификаторите кои се достапни во Weka софтверот (156). Евалуацијата на класификациските модели се прави во однос на нивото SCOP домен. Во оваа анализа протеинскиот скелет се интерполира со 64 интерполациски точки. При тоа беа

изградени класификациски модели користејќи ја секоја од карактеристиките индивидуално, како и во пар со Евклидовото растојание помеѓу интерполациските точки и центарот на маса. Резултатите за класификациската точност на класификациските модели добиени во оваа анализа се прикажани во Табела 3.10. Анализите покажаа дека со користење на Евклидовото растојание како индивидуална карактеристика се постигнува највисока точност. Со користење на C4.5 не се доби подобрување со ниту една од дополнително воведените карактеристики. Интересно е да се забележи дека со користење на само една карактеристика, моделот добиен со земање во предвид на хидрофобичноста покажа најдобри резултати после моделот добиен со користење на Евклидовото растојание. За Naïve Bayes се постигна класификациска точност од 94.30%, при што со воведувањето на карактеристиките за DPX и CX се постигна повисока точност отколку ако се користи само Евклидовото растојание. За *k*-nn беа направени анализи со користење на 1 и 3 најблиски соседи (*k*=1 и *k*=3), при што се доби подобрување со воведувањето на ASA, RASA и CX. Точностите кои се задебелени во Табела 3.10 се точностите на моделите кај кои со воведување на дополнителна карактеристика се постигна зголемување на класификациската точност. Со ова се покажа дека кај Naïve Bayes и *k*-nn класификаторите со воведувањето на дополнителни карактеристики на амино-киселинските остатоци се зголемува класификациската точност на предиктивните модели.

карактеристика	**поединечна карактеристика**				**во комбинација со Евклидовото растојание**			
	C4.5	**Naïve Bayes**	***k*-nn (*k*=1)**	***k*-nn (*k*=3)**	**C4.5**	**Naïve Bayes**	***k*-nn (*k*=1)**	***k*-nn (*k*=3)**
Евклидово растојание	**93.83**	92.32	98.36	98.18	/	/	/	/
totalASA	85.76	89.86	96.37	96.15	92.86	91.97	**98.49**	**98.34**
totalRASA	85.66	90.01	96.95	96.54	92.84	91.56	**98.42**	**98.27**
avgDPX	79.53	91.03	96.22	95.56	91.91	**94.12**	98.31	98.10
maxDPX	79.44	89.04	97.00	96.40	91.74	**93.05**	**98.39**	**98.37**
minDPX	72.76	69.76	80.47	79.63	93.30	**93.27**	97.95	97.67
avgCX	88.85	89.11	97.67	97.26	92.97	**93.88**	**98.39**	**98.34**
maxCX	88.53	89.32	96.98	96.51	92.62	**94.08**	**98.43**	**98.31**
minCX	88.67	90.62	97.74	97.22	92.66	**94.30**	**98.39**	**98.36**
хидрофобичност	90.84	86.87	96.37	96.28	93.44	91.05	97.89	97.73

Табела 3.10 Класификациска точност (%) на предиктивните модели.

4

ДЕТЕКЦИЈА НА СВРЗНИТЕ ДЕЛОВИ ОД ПРОТЕИНСКАТА СТРУКТУРА

Како што беше опишано во првото поглавје, предвидувањето на функцијата на протеинските структури наместо преку одредување на хомологни протеини може да се прави и преку детекција и анализа на сврзните делови од протеинската структура. Во ова поглавје ќе бидат презентирани неколку методи за детекција на сврзните делови од протеинската структура.

При детекција на сврзните делови од протеинските структури предвид може да се земат различни карактеристики на амино-киселинските остатоци. Амино-киселинските остатоци кои стапуваат во интеракција со остатоци од други протеински структури имаат висока конзервација која се должи на хидрогенските врски, Ван дер Валсовите сили, електростатските интеракции и хидрофобичните интеракции (25). Некои методи за одредување на сврзните делови од протеинската структура предвид ја земаат големината на сврзните региони (26), (27), (28), додека друга група на методи предвидувањето на сврзните региони го базира на растојанието помеѓу амино-киселинските остатоци кои стапуваат во интеракција (29), (30), (31). Постојат методи кои користат техники за пребарување на комплементарни површини за да ја одредат комплементарноста на структурите кои се испитуваат со цел да се одредат регионите каде што може да настане интеракција (32), додека во (33), (34), (35) се испитува комплементарноста на електростатските и хемиските карактеристики на протеинските структури. Во (36) и (37) е даден широк преглед

на алатките за одредување на сврзните региони од протеинските структури, како и на базите на податоци кои содржат информации за предвидените интерфејси на протеинските структури.

При одредување на сврзните делови од протеинската структура, во оваа докторска дисертација предвид се земени следниве карактеристики на амино-киселинските остатоци: дофатлива површина (Accessible Surface Area - ASA) (146), (147), релативна дофатлива површина (Relative ASA - RASA) (148), индекс на длабочина (depth index - DPX) (150), индекс на испакнатост (protrusion index - CX) (151) и хидрофобичност (149). Во секција 3.6.1 беше опишана постапката на екстракција на овие карактеристики. Во секција 3.6 овие карактеристики беа вклучени во протеинскиот дескриптор базиран на рамномерна интерполација на протеинскиот скелет со цел да се обезбеди попрецизно пребарување и класификација на протеинските структури. Во ова поглавје овие карактеристики се користат за да се идентификуваат амино-киселинските остатоци кои формираат сврзен регион каде што настанува интеракција со друга протеинска структура. Подоцна, врз основа на карактеристиките на сврзните региони може да се прави функционално анотирање на протеинските структури.

Голем дел од атомите кои влегуваат во составот на испитуваната протеинска структура се позиционирани во внатрешниот дел на структурата, и истите не можат да бидат дофатени од пробната сфера која се користи за одредување на дофатливата површина на атомите. Овие недофатливи атоми не можат да бидат дел од сврзен регион. Па затоа со цел да се избегне непотребното предвидување за неповршинските остатоци, како и да се намали бројот на примероци кои ќе се користат за обука на предиктивните модели и за тестирање на моделите, предвид се земаат само површинските амино-киселински остатоци. При тоа даден амино-киселински остаток се смета дека се наоѓа на површината доколку барем 5% од неговата вкупна површина може да биде дофатен од пробната сфера, како што е препорачано во (157).

По филтрирањето на амино-киселинските остатоци кои се наоѓаат на површината од протеинската структура, следи процесот на градење на предиктивен модел користејќи одреден класификациски метод. Во ова истражување се изградени предиктивни модели кои користат неколку класични методи за класификација, како и методи за класификација кои се базираат на непрецизираната логика. Во [A29], [A32], [A34], [A35] класичните методи за класификација беа применети за одредување на сврзните региони од протеинската структура, додека методите кои се базираат на непрецизираната логика беа применети во [A27], [A28], [A29], [A32], [A33], [A35].

4.1. Класични методи за одредување на сврзните региони од протеините

Во литературата се среќаваат разни класификатори со кои може да се изградат индивидуални модели за одлучување при класификација. Исто така постојат и техники со кои индивидуалните модели може да се соединат и да формираат ансамбл, со цел да се зголеми предиктивната моќ на моделот. Во оваа докторска дисертација се изградени повеќе индивидуални модели, како и ансамбли, и резултатите од истите ќе бидат презентирани во следните две секции.

4.1.1. Индивидуални модели за одредување на сврзните делови од протеинската структура

Во оваа докторска дисертација предвид се земени неколку класични методи за класификација како што се C4.5 дрвата на одлука (153), Дрва со поткастрување со намалена грешка (Reduced Error Pruning Tree - REPTree) (158), Наизменично дрво за одлучување (Alternating Decision Tree - ADTree) (159) и неговата верзија за решавање повеќекласни проблеми LADTree (160), Функционално дрво (Functional Tree - FTree) (161), Случајни шуми (Random Forest - RF) (162), Наивниот Баесов класификатор (Naïve Bayes) (154), Наивно Баесово дрво (Naïve Bayes Tree - NBTree) (163), Баесова мрежа (Bayesian Network) (164) и Повеќеслоен перцептрон (Multilayer perceptron) (165). Бидејќи овие методи се добро познати, истите нема да бидат презентирани. Во оваа докторска дисертација се користат имплементациите на овие класификатори кои се достапни во Weka софтверот (156). Во продолжение ќе биде направена евалуација на моделите за одредување на сврзните региони од протеинската структура користејќи ги класичните методи наведени погоре. Во анализите на предиктивната моќ на моделите добиени со класичните класификатори, кои ќе бидат презентирани во оваа и следната секција, предвид се земени следниве карактеристики на амино-киселинските остатоци: totalASA, avgDPX, avgCX и хидрофобичноста.

Опис на податочното множество

Како стандард на вистина се користи Biomolecular Interaction Network Database (BIND) (48) базата на податоци во која е сместено знаење во однос на сврзните делови од протеинските структури, при што ова знаење е добиено по експериментален пат. Бидејќи постојат голем број на протеински структури чија структура е веќе позната и сместена во PDB базата на податоци, и уште повеќе постојат бројни протеински структури кои се многу слични бидејќи претставуваат хомологни протеини, затоа најчесто при градење на моделите не се земаат во предвид сите познати структури туку само одредено репрезентативно множество. Репрезентативните протеински ланци се одбираат така што да имаат многу мала секвентна сличност помеѓу себе. Податочното множество кое се користи во оваа анализа е формирано со филтрирање на

протеинските ланци кои имаат помалку од 20% секвентна сличност користејќи го ASTRAL методот (145). Потоа, множеството за тестирање се формира од протеинските ланци кои имаат помалку од 10% секвентна сличност, додека останатите ланци го формираат множеството за обука. Со ова множеството за обука е формирано од амино-киселинските остатоци од 633 протеински ланци, додека множеството за тестирање се формира од остатоците на 3530-те ланци. Како што беше споменато погоре, голем број на амино-киселински остатоци се наоѓаат длабоко во внатрешноста на протеинската структура, и истите не можат да бидат дофатени од пробната сфера. Па затоа истите не можат да бидат дел од сврзен регион. Со цел да се намали димензионалноста на проблемот, се филтрираат само површинските остатоци, односно тоа се остатоците за кои важи дека барем 5% од нивната површина може да биде дофатена од пробната сфера. Со ова се добиваат 115579 амино-киселински остатоци во множеството за обука и 625939 остатоци во множеството за тестирање. При одредување на сврзните делови од протеинската структура, се градат модели кои одлучуваат за класниот атрибут, кој е бинарна променлива и покажува дали дадениот амино-киселински остаток е дел од сврзен регион или не. Од амино-киселинските остатоци кои се дел од множеството за обука само 15696 (околу 13.58%) се дел од сврзен регион според знаењето од BIND базата. Од ова може да се утврди дека класата која ги опфаќа амино-киселинските остатоци кои не се дел од сврзен регион е доминантна наспроти другата класа. Со цел да се спречи градење на предиктивни модели кои се наклонети кон доминантната класа, се прави балансирање на множеството за обука со избирање на примероци од множеството се до 27% од неговата големина без замена на примероците (даден примерок може само еднаш да се земе предвид) и следејќи рамномерна дистрибуција на класниот атрибут. Треба да се напомене дека ова балансирање е применето само над множеството за обука, додека множеството за тестирање останува небалансирано, па затоа во процесот на евалуација мора да се користи евалуациска мерка која е соодветна за небалансирани податочни множества. По балансирањето на множеството за обука, потоа сите влезни карактеристики се нормализираат во интервалот [0, 1] со цел сите карактеристики да имаат иста тежина во одлучувањето.

Евалуациска мерка

Постојат различни евалуациски мерки кои можат да се користат за евалуација на моделите кои предвидуваат даден класен атрибут. Сепак, класификациската точност не е соодветна мерка за евалуација во случај кога множеството за тестирање не е балансирано, како што е во оваа анализа. За таа цел во оваа анализа се користи Плоштината под ROC кривата (Area Under the ROC Curve) (AUC-ROC) мерката за евалуација за да се одреди предиктивната моќ на моделите. AUC-ROC се пресметува како

$$\text{AUC-ROC} = TPR*TNR+TPR*(1-TNR)/2+TNR*(1-TPR)/2= (TPR+TNR)/2, \quad (4.1)$$

каде *TPR* и *TNR* се стапките на точни позитивни и точни негативни предвидувања, соодветно. Стапката на точни позитивни предвидувања се пресметува како *TP*/(*TP*+*FN*), додека стапката на точни негативни предвидувања се одредува како *TN*/(*TN*+*FP*), каде што со *TP* и *TN* се означува бројот на точни позитивни предвидувања и бројот на точни негативни предвидувања, додека со *FP* и *FN* се означува бројот на грешни позитивни и бројот на грешни негативни предвидувања, соодветно. На овој начин AUC-ROC мерката прима вредности во интервалот [0, 1], каде вредноста 1 означува идеално предвидување на класниот атрибут, вредност од 0.5 претставува бескорисен класификатор, а вредноста 0 означува спротивно предвидување.

Експериментални резултати

Во Табела 4.1 се дадени резултатите добиени за AUC-ROC за моделите кои се генерирани со класичните класификациски методи кои беа наведени. Од резултатите може да се забележи дека со FTree се постигнува највисока предиктивна моќ, потоа следат C4.5 и NBTree класификаторите, додека со Multilayer perceptron се добиени најслаби резултати, пред сè затоа што се фаворизира негативната класа која ги содржи остатоците кои не се дел од сврзен регион. Потоа следат Случајните шуми (Random Forest), ADTree и LADTree класификаторите. Со ADTree и LADTree се добиваат исти резултати, бидејќи LADTree е проширување на ADTree за решавање на повеќекласни проблеми, но во овој случај проблемот кој се решава е бинарен (има 2 класи), па затоа се добива ист модел со двата методи.

Метод	**AUC-ROC**
C4.5	0.587
REPTree	0.568
ADTree/LADTree	0.546
FTree	0.589
Random Forest	0.543
Naïve Bayes	0.567
NBTree	0.586
Bayesian Network	0.576
Multilayer perceptron	0.526

Табела 4.1 AUC-ROC на моделите добиени со класичните методи за класификација.

4.1.2. Ансамбли за одредување на сврзните делови од протеинската структура

Со цел да се зголеми предиктивната моќ на моделите, во [A34] беа применети ансамблите со кои неколку класификациски модели се комбинираат во еден ансамбл. Постојат повеќе техники за градење на ансамбли, како што ***bagging*** (166) и ***boosting*** (167). При градење на моделите, примероците (во овој случај амино-киселинските остатоци) случајно се одбираат со замена, што

значи дека ако даден примерок случајно бил одбран, потоа истиот може повторно да биде одбран. На почетокот секој примерок им поседнаква веројатност да биде одбран. Кај bagging, при градење на следните модели кои влегуваат во ансамблот, веројатноста даден примерок да биде одбран останува униформна. Од друга страна, кај boosting во следните итерации при градење на модели се форсира учењето на примероците кои се погрешно класифицирани од претходните модели. На овој начин примероците кои се потешки за учење почесто се презентираат при градење на моделот. Со ова се форсира учење на примероците кои во просторот се наоѓаат во региони во кои има поголем број на грешки при класификацијата. Во продолжение накратко ќе биде опишана постапката на градење на ансамблите.

Техники за градење на ансамбли

Кај двете техники случајно се одбираат примероците кои ќе се земат предвид при градење на моделите кои го формираат ансамблот. Нека множеството за обука D има $|D|$ примероци, и нека сакаме да изградиме ансамбл кој се состои од m модели добиени со некој класификациски метод. Прво се генерираат m податочни множества за обука D_i, i=1, 2,…, m, со големина $|D_i| \leq |D|$ добиени со одбирање на примероци од податочното множество D со замена на примероците (даден примерок може да биде избран повеќе пати), при тоа следејќи ја дистрибуцијата на класниот атрибут во целокупното множество за обука. Бидејќи примероците се обираат со замена, тоа значи дека даден примерок може да биде одбран повеќе пати во исто множество за обука. Со користење на множествата за обука D_i, i=1, 2,…, m се градат m-те модели кои го формираат ансамблот. Во процесот на тестирање, даден примерок се доведува на влез на сите m модели, и конечната одлука за класниот атрибут се донесува со мнозинско гласање. Во ова истражување при гласањето секој модел кој е дел од ансамблот добива иста тежина при гласањето. Кај bagging техниката, примероците имаат рамномерна веројатност да бидат случајно одбрани во текот на целиот процес (во сите итерации). Од друга страна, кај boosting техниката примероците кои се потешки за учење во подоцнежните итерации добиваат поголема веројатност да бидат одбрани. На почетокот примероците имаат рамномерна веројатност да бидат одбрани. Случајно се одбираат примероци, и се формира множеството за обука D_1. Потоа моделот M_1 се гради користејќи ги примероците од множеството D_1, по што истите му се презентираат на моделот за тестирање. За секој примерок од D_1 кој е погрешно класифициран од моделот M_1 му се зголемува веројатноста да биде одбран во следните итерации при градење на следните модели. Од друга страна, на секој примерок од D_1 кој е точно класифициран од моделот M_1 му се намалува веројатноста да биде случајно одбран во следните итерации. Врз основа на новодобиените веројатности, потоа случајно се одбираат примероците кои ќе го формираат множеството D_2. По тоа следи градење на моделот M_2 користејќи ги примероците од D_2, и потоа врз основа на

предвидувањата добиени за примероците од D_2 со новогенерираниот модел M_2, се одредуваат новите веројатности за одбирање на примероците. Оваа постапка се повторува сè додека да се генерираат m модели, или сè додека подобрувањето во последните итерации да е под некој предефиниран праг. Со ова, доколку даден примерок е погрешно класифициран од претходните модели, тогаш истиот има поголема веројатност да биде одбран при градење на следните модели.

Експериментални резултати

При градење на ансамблите предвид се земени следниве класификатори: C4.5 дрвата на одлука (153), Наизменично дрво за одлучување (Alternating Decision Tree - ADTree) (159), и неговата верзија за решавање повеќекласни проблеми LADTree (160), Наивниот Баесов класификатор (Naïve Bayes) (154), Наивно Баесово дрво (Naïve Bayes Tree - NBTree) (163) и Баесова мрежа (Bayesian Network) (164). Методот за градење Функционални дрва (FTree) не е земен превид бидејќи времето на тестирање доколку се користи овој класификатор е значително подолго, па при користење на овој класификатор за градење на ансамбл се добива модел со кој потребно е многу повеќе време за тестирање. Во оваа докторска дисертација се користат имплементациите на bagging и boosting техниките кои се достапни во Weka софтверот (156). За boosting се користи Adaptive Boosting (AdaBoost) методот (167). За секој метод се користат иницијалните поставки, освен ако не е кажано поинаку во текстот.

Прво ќе бидат претставени резултатите од евалуацијата на моделите добиени со bagging техниката. При тоа направено е испитување на влијанието на бројот на итерации m, што одговара на бројот на модели кои се комбинираат во ансамблот. Исто така ќе бидат изградени повеќе ансамбли користејќи множества D_i, i=1,2,…,m со различна големина $|D|*k/100$, каде k=5, 10, 20, 50 и 100, а со $|D|$ е означен бројот на примероци во целокупното множество за обука. Со параметарот k се дефинира колкав дел од целокупното податочно множество се зема предвид при градење на поединечните модели. Резултатите од оваа анализа се прикажани во Табела 4.2. Слично, генерирани се и ансамбли користејќи ја boosting техниката. При тоа генерирани се повеќе модели користејќи различна вредност за максимално дозволениот број на итерации (m=10, 20 и 50). Резултатите за предиктивната моќ на ансамблите добиени со boosting техниката се прикажани во Табела 4.3. Од резултатите презентирани во Табела 4.2 и Табела 4.3 може да се утврди дека ансамблот добиен со C4.5 класификаторот користејќи ја bagging техниката дава најдобри резултати (AUC-ROC=0.588) со користење на m=20 и k=50. Со користење на bagging техниката, предиктивната моќ на моделите добиени со C4.5 класификаторот е зголемена од 0.587 на 0.588. Сепак, со користење на boosting техниката не се постигна зголемување на предиктивната моќ доколку се користи C4.5 класификаторот. Во однос на ADTree методот за класификација, со користење на поединечен модел се постигна AUC-ROC од 0.546. Со користење на

ансамбли, вредноста на AUC-ROC се зголеми на 0,580 со bagging и на 0,584 со boosting техниката. Слично, со користење на Naïve Bayes, вредноста на AUC-ROC се зголеми од 0,567 на 0,574. Слично, и со останатите класификатори со користење на ансамбли се подобри предиктивната моќ на моделите. Генерално, boosting техниката се покажа како подобра техника за градење на ансамбли, освен во случајот кога се гради ансамбл од C4.5 дрва на одлука. Во однос на C4.5 ансамблот добиен со boosting, интересно е да се спомене дека за m=10 се добива попрецизен модел отколку ако m=20 или 50. Со boosting, во подоцнежните итерации примероците кои почесто се грешени од претходните модели почесто се одбираат при тренирањето, но тоа во случајов резултира со преобучување на моделот.

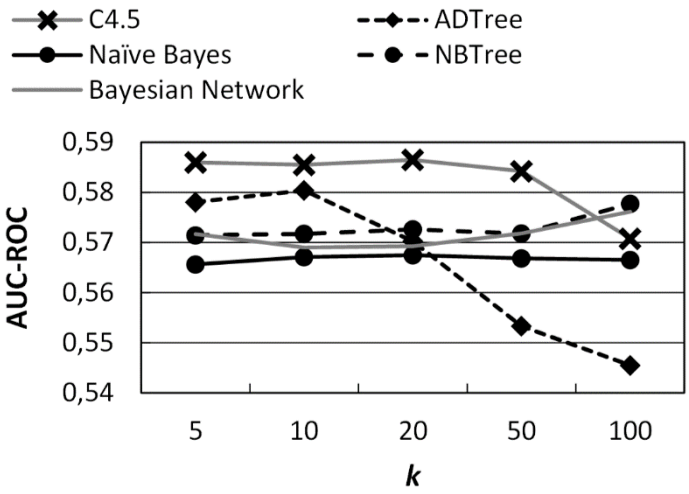
m	*k*	C4.5	ADTree/ LADTree	Naïve Bayes	NBTree	Bayesian Network
10	5	0.586	0.578	0.566	0.571	0.572
10	10	0.586	0.580	0.567	0.572	0.569
10	20	0.587	0.570	0.568	0.573	0.569
10	50	0.584	0.553	0.567	0.572	0.572
10	100	0.571	0.546	0.567	0.578	0.576
20	5	0.587	0.570	0.567	0.573	0.573
20	10	0.587	0.569	0.567	0.571	0.570
20	20	0.588	0.575	0.567	0.573	0.569
20	50	**0.588**	0.561	0.567	0.571	0.571
20	100	0.573	0.546	0.567	0.578	0.575

Табела 4.2 AUC-ROC на ансамблите добиени со bagging техниката.

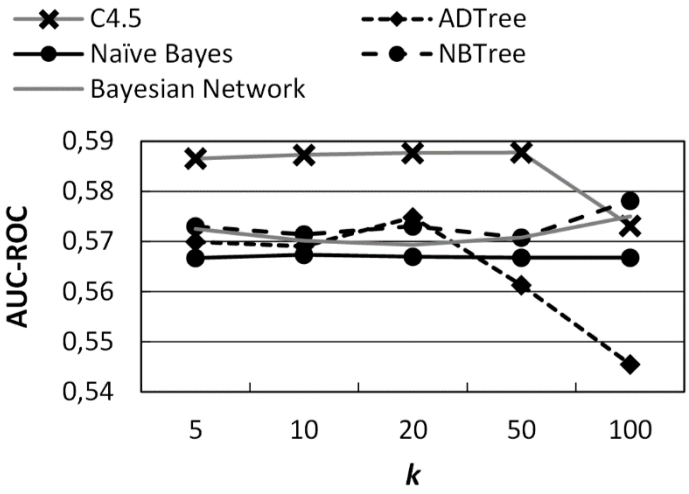
m	C4.5	ADTree/ LADTree	Naïve Bayes	NBTree	Bayesian Network
10	0.583	0.584	0.574	0.587	0.586
20	0.572	0.584	0.574	0.587	0.586
50	0.572	0.584	0.574	0.587	0.586

Табела 4.3 AUC-ROC на ансамблите добиени со boosting техниката.

Во однос на параметарот k, од Табела 4.2 може да се забележи дека попрецизни модели се добиваат доколку моделите кои го формираат ансамблот не се генерирани врз основа на целото податочно множество (за k<100). На овој начин користејќи помала вредност на параметарот k се избегнува преобучување на моделот. Сепак, користејќи премногу мали подмножества за градење на моделите (за k=5), предиктивната моќ е помала. За да се добие подобра слика за влијанието на параметарот k, на Слика 4.1 и Слика 4.2 презентирани се резултатите користејќи различни вредности за овој параметар.



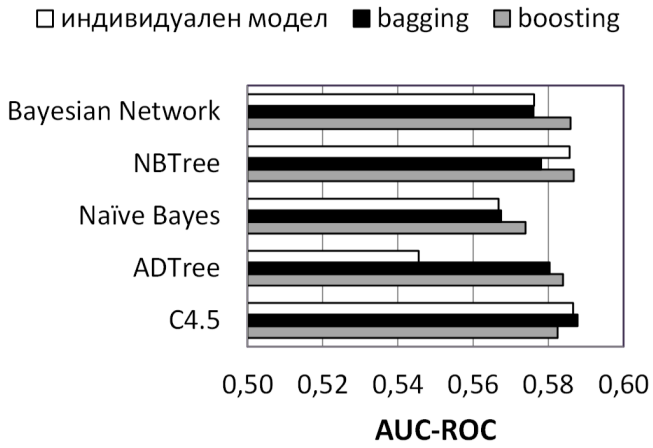
Слика 4.1 AUC-ROC за ансамблите добиени со bagging техниката за m=10.



Слика 4.2 AUC-ROC за ансамблите добиени со bagging техниката за m=20.

Како што може да се забележи, општо земено, оптимални вредности за k се 10, 20 и 50. Од резултатите очигледно е дека за k=100 (користејќи податочни множества D_i чија големина е еднаква на големината на целокупното множество D), предиктивната моќ на моделите е полоша при користење на C4.5 и ADTree класификаторите бидејќи моделите се преобучени. Сепак, со користење на методите кои се базираат на веројатности (Naïve Bayes, NBTree и Bayesian Network) ова не е случај. Кај овие класификатори кои се базираат на Баесовата теорема, користејќи поголем дел од податоците ќе се добие порелевантна информација за распределбата на веројатноста на карактеристиките кои се користат, па затоа доколку се користат овие класификатори вредноста на k треба да биде што поголема.

На Слика 4.3 се прикажани сумарните резултати од оваа анализа каде што се споредени вредностите за AUC-ROC за индивидуалниот модел, како и за ансамблите добиени со bagging и boosting техниките. Резултатите покажаа дека со користење на ансамбли се подобри предиктивната моќ на моделите. Генерално, со користење на boosting техниката се добија модели со повисока предиктивна моќ, освен кај ансамблот добиен со C4.5 класификаторот.



Слика 4.3 AUC-ROC за индивидуалните модел и за ансамблите добиени со bagging и boosting техниките користејќи различни методи за класификација.

4.2. Методи за одредување на сврзните региони од протеинските структури базирани на непрецизирана логика

Сепак, класичните методи се осетливи на мали промени во податочното множество. Од друга страна, при еволуција карактеристиките на амино-киселинските остатоци бележат мали промени, па затоа со користење на класичните методи за класификација може за два слични протеини да се добијат различни предвидувања како резултат на малите разлики во карактеристиките на амино-киселинските остатоци. Токму затоа беше воведена непрецизираната логика за откривање на сврзните делови од протеинската структура. Во оваа секција ќе бидат презентирани неколку методи базирани на непрецизираната логика со кои може да се врши детекција на сврзните региони од протеинската структура.

За да се надмине проблемот со голема осетливост на малите промени во вредностите на карактеристиките врз основа на кои се донесуваат заклучоци при класификацијата, се воведени Непрецизираните дрва на одлука (англ. Fuzzy Decision Trees) (НДО) (168). Во литературата може да се најдат бројни истражувања со НДО, пред се заради робустноста на промени на податоците,

како и на отпорноста на преобучување. Во (169) се презентирани алгоритами за градење на правила преку учење на примероци базирајќи се на непрецизираната логика. Класичните дрва на одлука (153) инспирирале многу истражувања во областа на непрецизираните дрва на одлука. Во (168), (170) воведени се различни методи за градење на непрецизирани дрва на одлука. Во (171) и (172) се презентирани различни оптимизации на непрецизираните одлучувачки стебла, а во (173) е даден преглед на предностите и недостатоците на дрвата на одлука базирани на непрецизираната логика. Во [A27] беа воведени непрецизираните дрва на одлучување за одредување на сврзните делови од протеинската структура. Сепак, кај непрецизираните одлучувачки стебла може да се користи само еден оператор за агрегација, додека кај Непрецизираните дрва на припадност (англ. Fuzzy Pattern Trees) (НДП) (174) има можност за комбинирање на различни оператори за агрегација во процесот на градење на предиктивниот модел. Во (174) е воведен метод од-дното-нагоре (bottom-up) за градење на непрецизирани дрва на припадност при што стеблото се гради од листовите кон коренот, додека во (175) е предложен метод од-врвот-надолу (top-down) со кој непрецизираното дрво на припадност се гради од коренот кон листовите. Во продолжение ќе бидат опишани постапките на градење на непрецизираните дрва на припадност.

4.2.1. Непрецизирана (fuzzy) логика

Методите за градење на непрецизирани дрва на припадност се базираат на теоријата на непрецизираните (матни, англ. fuzzy) множества (176), (177), која се обидува да ги реши проблемите со кои се соочува класичната теорија на множества при градење на предиктивни модели. Имено во класичната теорија на множества даден елемент или припаѓа или не припаѓа во дадено множество, додека кај непрецизираните множества припадноста на даден елемент во одредено множество се дефинира преку одреден степен на припадност. Нека U е множеството на непрецизирани променливи кои се означени како x. Во непрецизираната логика може да се дефинира функцијата на припадност $\mu_A(x)$ со која се претставува степенот со кој дадената променлива x припаѓа во непрецизираното множество A. Непрецизираната променлива x претставува променлива која има лингвистички вредности наречени непрецизирани (fuzzy) термини преку кои се дефинираат непрецизираните множества. Множеството од непрецизирани променливи во овој случај ќе ги содржи сите непрецизирани термини кои се користат за карактеристиките на амино-киселинските остатоци кои се земаат предвид при одредувањето на сврзните региони на протеинската структура. Преку непрецизираните термини може да се опише дали даден амино-киселински остаток има мала или голема дофатлива површина, дали има мал или голем индекс на длабочина со што се опишува дали остатокот е блиску или далеку во однос на површината на протеинската структура.

Фунцкии на припадност

Во литературата дефинирани се повеќе функции на припадност (ФП) кои се применуваат за различни распределби на променливите кои се користат за градење на класификациски модел. Триаголната и трапезоидната функција на припадност се основните функции во непрецизираната логика и тие се воведени во (178). Триаголната функција на припадност се дефинира преку три параметри a, b и c како

$$\mu_A(x,a,b,c)=\begin{cases}0, & x\le a\\ \dfrac{x-a}{b-a}, & a\le x\le b\\ \dfrac{c-x}{c-b}, & b\le x\le c\\ 0, & c\le x\end{cases},\quad x,a,b,c>0, \tag{4.2}$$

додека трапезоидната функција на припадност се дефинира преку четири параметри a, b, c и d

$$\mu_A(x,a,b,c,d)=\begin{cases}0, & x\le a\\ \dfrac{x-a}{b-a}, & a\le x\le b\\ 1, & b\le x\le c\\ \dfrac{d-x}{d-c}, & c\le x\le d\\ 0, & d\le x\end{cases},\quad x,a,b,c,d>0. \tag{4.3}$$

Гаусовата функција на припадност може подобро да ги претстави променливите кои имаат нормална распределба. Ова функција на припадност е дефинирана со параметрите a и b како

$$\mu_A(x,a,b)=\frac{1}{b\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2b^2}},\quad x,a,b>0, \tag{4.4}$$

каде параметрите a и b се средната вредност и стандардната девијација за дадената функција на припадност.

Сигмоидалната (Sigmoidal) функција на припадност е дефинирана со параметрите a и b како

$$\mu_A(x,a,b)=\frac{1}{1+e^{-a(x-b)}},\quad x,b>0, \tag{4.5}$$

каде b е средната вредност, додека константата a прима вредности +1 и -1. Функцијата на припадност добиена за a = 1 ја означуваме како сигмоидална(+1), додека функцијата добиена за a = -1 ја означуваме како сигмоидална(-1).

Покрај погореопишаните функции на припадност, постојат и други функции на припадност, но во ова истражување се користат само четирите функции кои беа опишани.

Непрецизирани оператори за агрегација

Во класичната теорија на множества постојат основните операции унија, пресек и комплемент, па слично и во теоријата на непрецизираната логика дефинирани се операции кои може да се применат за агрегација на функциите на припадност. Ова се постигнува со користење на непрецизирани оператори за агрегација, кои може да се поделат на: Т-норми, Т-конорми и оператори за усреднување. Т-нормата и Т-конормата се воведени во (179) и претставуваат триаголни норми кои имаат свој пандан во класичната теорија на множества. Имено, пресек може да се претстави преку Т-норма, додека унија може да се претстави со Т-конорма. Така на пример MIN и MAX непрецизираните оператори за агрегација се пандан на пресек и унија кај класичните множества. Во продолжение ќе бидат дадени дефиниции на Т-нормите и Т-конормите кои се користат во оваа докторска дисертација. Нека со a и b ги означиме функциите на припадност $\mu_A(x)$ и $\mu_B(x)$, соодветно. Минимум Т-нормата, која уште се нарекува и Годел Т-норма, се пресметува како Т-норма$_{min}(a, b)$ = min(a, b) и претставува стандардна семантика за слаба конјугација. Производ Т-нормата има значење на силна конјугација и се пресметува како Т-норма$_{производ}(a, b) = a * b$. Т-нормата на Лукасиевич (Lukasiewicz) претставува силна конјугација во непрецизираната логика на Лукасиевич и се пресметува како Т-норма$_{Лукасиевич}(a, b)$ = max($a + b$ - 1, 0). Ајншајновата Т-норма е дефинирана преку воопштената Хамашерова Т-норма

$$\text{Т-норма}_{\text{Хамашерова}}(a,b,p)=\frac{ab}{p+(1-p)(a+b-ab)}, \tag{4.6}$$

при што за $p = 2$ се добива Ајншајновата Т-норма

$$\text{Т-норма}_{\text{Ајнштајнова}}=\text{Т-норма}_{\text{Хамашерова}}(a,b,2)=\frac{ab}{2-(a+b-ab)}. \tag{4.7}$$

Покрај Т-нормата може да се користи и комплементарната конорма, уште наречена како Т-конорма. Со Т-конормата се претставуваат спротивните операции во однос на Т-нормите, односно Т-нормата претставува пресек, додека Т-конормата претставува унија од множествата. Нека е дадена Т-нормата Т-норма(a, b), тогаш нејзината комплементарна конорма (Т-конорма) се пресметува како

$$\text{Т-конорма}(a, b) = 1 - \text{Т-норма}(1 - a, 1 - b). \tag{4.8}$$

Максимум Т-нормата е комплементарна конорма на минимум Т-нормата и е дефинирана како Т-конорма$_{max}(a, b)$ = max(a, b). Оваа норма е најслаба од сите Т-конорми. Сума Т-конормата е

комплементарна конорма на производ Т-нормата, и истата ја претставува веројатноста на унија од независни настани. Оваа Т-конорма е дефинирана како Т-конорма$_{сума}(a, b) = a + b – a * b$. Врзаната сума претставува комплементарна норма на Лукасиевич Т-нормата, и истата се пресметува како Т-конорма$_{Лукасиевич}(a, b) = \min(a + b, 1)$. Ајншајновата Т-конорма се пресметува како

$$\text{Т-конорма}_{\text{Ајнштајнова}} = \frac{a+b}{1+ab}. \tag{4.9}$$

Во Табела 4.4 е даден преглед на Т-нормите и Т-конормите кои се користат во оваа докторска дисертација.

Име	**Т-норма**	**Т-конорма**
MIN / MAX	$\min\{a, b\}$	$\max\{a, b\}$
Производ (AND) / Сума (OR)	$a * b$	$a + b – a * b$
Лукасиевич	$\max(a + b - 1, 0)$	$\min(a + b, 1)$
Ајнштајнова	$\frac{ab}{2-(a+b-ab)}$	$\frac{a+b}{1+ab}$

Табела 4.4 Т-норми и Т-конорми кои се користат во оваа докторска дисертација.

Како што беше споменато претходно, покрај Т-нормите и Т-конормите постојат и непрецизирани оператори за агрегација кои вршат усреднување. Во продолжение ќе бидат опишани четири непрецизирани оператори за агрегација кои вршат усреднување.

Со Weighted Averaging (WA) операторот (180) од ред n се прави мапирање $R^n \rightarrow R$ со

$$\text{WA}(a_1, a_2, ..., a_n) = \sum_{i=1}^{n} w_i * a_i, \qquad \sum_{i=1}^{n} w_i = 1 \tag{4.10}$$

каде $W = (w_1, w_2, \ldots, w_n)^T$, $w_i \in [0, 1]$, $1 \leq i \leq n$, е n-димензионален вектор со тежини, а $a_1, a_2, \ldots, a_n$ се вредностите на елементите кои се агрегираат со непрецизираниот оператор за агрегација. За одредување на тежинските вектори може да се користат различни методи. Во (180) предложени се два методи за одредување на тежинските вектори. Во оваа докторска дисертација се користи методот предложен во (180), кој служи за одредување на тежинскиот вектор при што користи даден непрецизиран квантификатор Q. Со користење на овој метод, тежините се одредуваат како

$$w_i = Q(i/n) - Q((i-1)/n), \; i = 1, ..., n, \tag{4.11}$$

каде што Q е непрецизираниот квантификатор кој се користи.

Кај Ordered Weighted Averaging (OWA) операторот (180) се врши подредено тежинско усреднување. Кај овој оператор мапирањето се прави со

$$\mathrm{OWA}(a_1, a_2, ..., a_n) = \sum_{i=1}^{n} w_i * f_i(a_1, a_2, ..., a_n), \qquad \sum_{i=1}^{n} w_i = 1 \tag{4.12}$$

каде $f_i(a_1, a_2, \ldots, a_n)$ го враќа i-тиот најголем елемент во множеството $\{a_1, a_2, \ldots, a_n\}$. Главната разлика помеѓу WA и OWA операторите е тоа што кај OWA нема специфични тежини за елементите, туку тежините се асоцирани со подредените позиции на елементите. WA и OWA непрецизираните оператори за агрегација кои користат непрецизиран квантификатор Q при пресметување на тежинскиот вектор W се означуваат со WA_Q и OWA_Q.

Weighted Geometric (WG) и Ordered Weighted Geometric (OWG) операторите се е дефинирани во (181) и вршат тежинско геометриско усреднување. Овие оператори ги комбинираат одлуките за податоците кои имаат соодносен размер. Во ова истражување податоците ќе бидат нормализирани така што сите карактеристики на амино-киселинските остатоци ќе бидат во ист опсег, па со тоа и ќе имаат соодносен размер. Со WG операторот од ред n се прави мапирање $R^n \rightarrow R$ со

$$\mathrm{WG}(a_1, a_2, ..., a_n) = \prod_{i=1}^{n} w_i a_i, \tag{4.13}$$

каде $W = (w_1, w_2, \ldots, w_n)^{\mathrm{T}}$, $w_i \in [0, 1]$, $1 \leq i \leq n$, е n-димензионален вектор со тежини. OWG операторот врши агрегирање на листа на вредности $\{a_1, a_2, \ldots, a_n\}$ според следново равенство

$$\mathrm{OWG}(a_1, a_2, ..., a_n) = \prod_{i=1}^{n} c_i^{w_i} \tag{4.14}$$

каде $C = (c_1, c_2, \ldots, c_n)^{\mathrm{T}}$ е вектор во кој елементите од множеството $\{a_1, a_2, \ldots, a_n\}$ се подредени во опаѓачки редослед, односно c_i е i-тиот најголем елемент од множеството на вредности. Кај геометриските оператори за тежинско усреднување се користи истиот метод за одредување на тежинскиот вектор W кој беше користен и кај WA и OWA операторите. Слично како и кај WA и OWA операторите, така и кај WG и OWG операторите доколку се користи даден непрецизиран квантификатор Q, тогаш операторите се бележат со WG_Q и OWG_Q.

Метрики за сличност

Во продолжение ќе бидат презентирани метриките за сличност кои во ова истражување се користат при генерирање на класификациски модел за одредување на сврзните делови од протеинската структура. Нека A и B се две непрецизирани множества кои се однесуваат на множествата на две променливи во множеството на непрецизирани променливи U. Функциите

на припадност за двете непрецизирани множества A и B се означени со $\mu_A(x)$ и $\mu_B(x)$. Во (182) воведена е RMSE (Root Mean Squared Error) метриката со која се пресметува средната квадратна грешка на двете непрецизирани множества кои се споредуваат A и B. RMSE се пресметува како

$$\mathrm{RMSE}(A,B)=1-\sqrt{\frac{\sum_{i=1}^{n}\left(\mu_A(x_i)-\mu_B(x_i)\right)^2}{n}}. \tag{4.15}$$

RMSE прима вредности во интервалот [0,1], при што поголема вредност означува поголема сличност помеѓу непрецизираните множества кои се споредуваат. Бидејќи непрецизираните множества кои се споредуваат се однесуваат на две карактеристики кои се земаат предвид како влезни или излезни атрибути при класификацијата, затоа со метриката на растојание се врши споредување на тие две карактеристики.

Во [A30] беа предложени две нови метрики за сличност, и истите беа применети за градење на класификациски модели за одредување на индикаторските карактеристики на дијатомеите во водните екосистеми. Во оваа докторска дисертација овие метрики за сличност се применети и за градење на класификациски модели за детекција на сврзните региони од протеинската структура. RMSE^2 метриката за сличност е дефинирана како

$$\mathrm{RMSE}^2(A,B)=1-\sqrt{\frac{\sum_{i=1}^{n}\left(\mu_A(x_i)^2-\mu_B(x_i)^2\right)^2}{n}} \tag{4.16}$$

Како што може да се забележи од равенствата (4.14) и (4.15), кај RMSE^2 метриката за сличност поголемо значење им се дава на карактеристиките кои имаат поголема вредност наспрема останатите карактеристики. Во [A30] оваа метрика за сличност беше воведена со цел да дијатомеите кои се позастапени, да добијат поголема тежина при одлучувањето. Затоа кај оваа метрика за сличност степенот на припадност се зема со степен два. Во [A30] се покажа дека со воведувањето на оваа метрика се постигна зголемување на предиктивната моќ на моделите. Во оваа докторска дисертација ќе биде испитано дали оваа метрика за сличност е соодветна за градење на модели кои ќе се користат за одредување на сврзните региони од протеинските структури.

Другата метрика за сличност која беше воведена во [A30] е G_MAX метриката, и таа е инспирирана од истата причина како и RMSE^2 метриката. Имено, кај двете метрики целта е да се фаворизираат дијатомеите кои во водниот екосистем се позастапени во рамки на дадената класа на вода. G_MAX метриката за сличност е дефинирана како

$$\mathrm{G\_MAX}(A,B)=1-\left[\frac{\sum_{i=1}^{n}\left|\mu_A(x_i)-\mu_B(x_i)\right|*\max\{\mu_A(x_i),\mu_B(x_i)\}}{n}\right] \tag{4.17}$$

Во продолжение ќе биде даден пример за одредувањето на трите метрики за сличност кои беа презентирани. Нека A и C се две непрецизирани множества кои се однесуваат на еден од термините на некој влезен атрибут и на класниот атрибут кој треба да се предвиди со класификацискиот модел. Нека A={0.8; 0.9; 0.5; 0.2; 0.4; 0.3} и C={0.9; 0.8; 0.7; 0.9; 0.3; 0.1} се двете непрецизирани множества кои ги содржат вредностите за степенот на припадност за непрецизираните термини кои се споредуваат. Сличноста помеѓу двете непрецизирани множества A и C добиена со трите метрики за сличност е

$$\mathrm{RMSE}(A,C)=$$
$$1-\sqrt{\frac{(0.8-0.9)^2+(0.9-0.8)^2+(0.5-0.7)^2+(0.2-0.9)^2+(0.4-0.3)^2+(0.3-0.1)^2}{6}}=0.6838$$

$$\mathrm{RMSE}^2(A,C)=$$
$$1-\sqrt{\frac{(0.8^2-0.9^2)^2+(0.9^2-0.8^2)^2+(0.5^2-0.7)^2+(0.2^2-0.9^2)^2+(0.4^2-0.3^2)^2+(0.3^2-0.1^2)^2}{6}}=0.6537$$

$$\mathrm{G\_MAX}(A,C)=$$
$$1-\left[\frac{|0.8-0.9|*0.9+|0.9-0.8|*0.9+|0.5-0.7|*0.7+|0.2-0.9|*0.9+|0.4-0.3|*0.4+|0.3-0.1|*0.3}{6}\right]=0.825.$$

4.2.2. Методи за градење на Непрецизирани Дрва на Припадност (НДП)

Во претходните секции беа презентирани основните поими кои се користат во непрецизираната логика. Исто така беа презентирани функциите на припадност, операторите за агрегација и метриките за сличност кои се земени предвид во ова истражување за градење на предиктивни модели за детекција на сврзните региони од протеинската структура. Користејќи ги функциите на припадност, прво се прави омекнување (англ. fuzzyfication) на карактеристиките користејќи одреден број на непрецизирани термини, а потоа со користење на операторите за агрегација и

метриките за сличност се гради класификациски модел. Во продолжение ќе биде опишан процесот на градење на предиктивен модел, како и процесот на одлучување при класификација.

Нека имаме четири влезни атрибути a_1, a_2, a_3 и a_4 кои се однесуваат на карактеристиките на амино-киселинските остатоци. Нека направиме омекнување така што за секоја непрецизирана променлива дефинираме по T непрецизирани термини. Од тука доменот на непрецизираната променлива a_i е домен(a_i) = $\{t_{i1}, t_{i2},\ldots, t_{iT}\}$. Пример, ако за дадена променлива се дефинираат три непрецизирани термини, тогаш доменот за таа променлива ќе биде {мало, средно, големо}. Бидејќи не може точно да се дефинираат границите за вредностите на карактеристиките според кои може да се прави прецизно одредување на сврзните региони од протеинската структура, затоа подобро е да се користат опсези преку кои се дефинираат непрецизираните термини. Нека c_j е j-тата класа за која постои непрецизирано множество во множеството U. Преку функцијата $\mu_{c_j}(x)$ може да се дефинира степенот со кој даденият објект x припаѓа во j-тата класа c_j. Секој атрибут a_i е дефиниран како лингвистички атрибут, и тој може да прима лингвистички вредности до доменот домен(a_i) = $\{t_{i1}, t_{i2},\ldots, t_{iT}\}$. Секој непрецизиран термин (лингвистичка вредност) е член во непрецизираното множество U. Преку $\mu_{t_{i,k}}(x)$ се претставува степенот на припадност за објектот x за непрецизираниот термин $t_{i,k}$ дефиниран за атрибутот a_i.

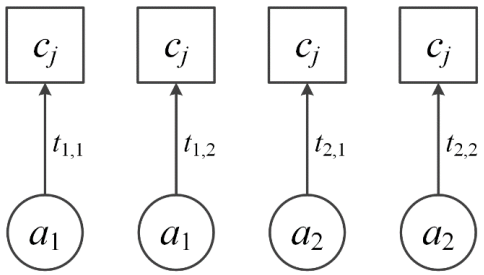
Како што беше споменато претходно, прво се воведени Непрецизираните дрва на одлука (НДО) (168), а потоа се воведени Непрецизираните дрва на припадност (НДП) (174). Прво ќе биде направена споредба на овие два методи. Со НДП се прави посебен модел за секоја излезна класа, па во процесот на одлучување, даден примерок за тестирање се доведува на влез на сите модели и се класифицира во класата за која е добиена најголема вредност за сличноста. Од друга страна кај НДО се прави еден модел со кој се врши класификација на даден непознат примерок во некоја од излезните класи. Даден модел добиен со методот за градење на НДО може да се претстави како множество од НДП, каде бројот на НДП ќе биде еднаков со бројот на излезни класи кои се предвидуваат со НДО моделот. Односно даден НДО модел може да се претстави како множество од семи-бинарни НДП модели. Методот за градење на НДО се обидува да ги раздели примероците кои припаѓаат во различни класи, додека НДП се обидува да ги идентификува заедничките карактеристики на примероците кои припаѓаат во дадената класа за која се гради моделот. Кај методот за градење на НДО предвид се земаат само Т-норма$_{производ}$ и Т-норма$_{сума}$ операторите за агрегација (кои уште се нарекуваат и алгебарски AND и OR, соодветно), додека кај НДП може да се користат различни оператори за агрегација.

Следно, ќе биде направена споредба помеѓу класичните дрва на одлука, како што е на пример C4.5 методот (153), и НДП. Класичните методи се базираат на евристична функција, која е базирана на ентропија, за да одредат кој атрибут носи најголема информациона добивка. Од

друга страна кај НДП се користи одредена метрика за сличност за да се одреди со која агрегација ќе се постигне моделот подобро (со поголема сличност) да ги претстави примероците кои припаѓаат на класата за која се гради моделот. Кај методот за градење на НДП предложен во (174) стеблото на одлука се гради од листовите кон коренот (од-дното-нагоре), додека кај класичните дрва на одлука стеблото се гради од коренот кон листовите. Сепак, подоцна во (175) предложен е метод за градење на НДП каде стеблото се гради од коренот кон листовите (од-врвот-надолу). Подоцна во оваа секција ќе бидат презентирани разликите помеѓу двата методи за градење на НДП кои се земени предвид во ова истражување, а тоа се методите од-дното-нагоре (174) и од-врвот-надолу (175).

Метод од-дното-нагоре за градење Непрецизирани Дрва на Припадност (НДП)

Како што беше кажано, со НДП се опишуваат карактеристиките на примероците кои припаѓаат во иста класа. Значи целта на методот не е да ги најде различните карактеристики помеѓу примероците кои припаѓаат во различни класи, туку да ги најде сличностите помеѓу примероците од иста класа. Се гради посебен модел за секоја излезна класа која се предвидува. Градењето на НДП започнува со одредување на основните (примитивните) НДП кои постојат на нулто ниво. За таа цел се прави посебно примитивно стебло за секој непрецизиран термин кој се користи од сите влезни карактеристики според кои треба да се одлучува. На Слика 4.4 е даден пример за примитивни НДП за класа c_j добиени доколку се користат две карактеристики a_1 и a_2, и за секоја карактеристика се користат по два непрецизирани термини (T = 2). Доменот на непрецизираните променливи е домен(a_1)={$t_{1,1}$, $t_{1,2}$} и домен(a_2)={$t_{2,1}$ и $t_{2,2}$}.



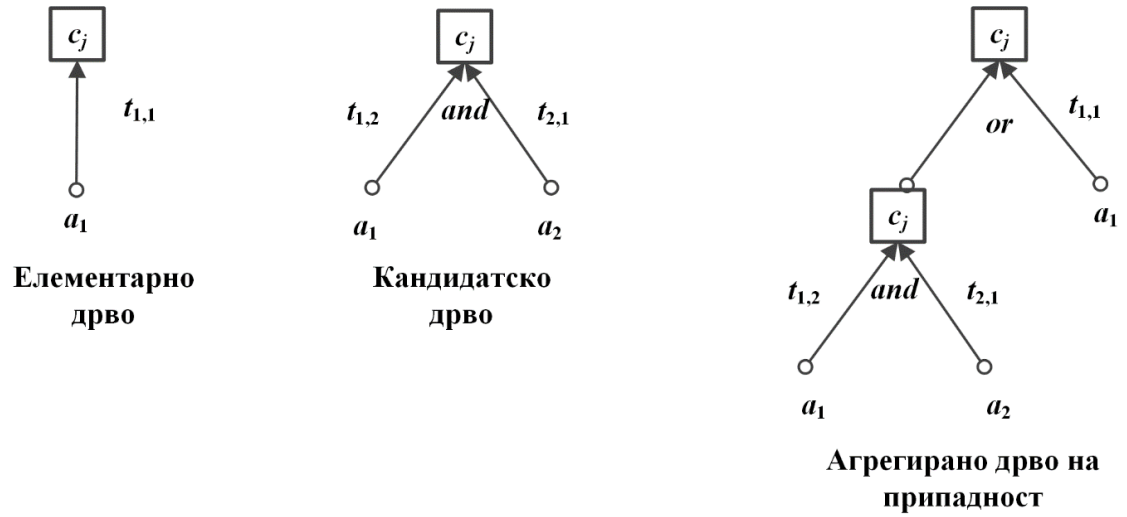
Слика 4.4 Примитивни непрецизирани дрва на припадност.

Во ова истражување при градење на НДП моделите предвид се земаат четири карактеристики на амино-киселинските остатоци, па доколку бројот на непрецизирани термини по карактеристика T се постави на 5, тоа значи дека ќе се добијат 20 примитивни дрва. За секое примитивно

дрво користејќи одредена метрика за сличност може да се одреди колкава е сличноста за дадената класа за која се гради НДП моделот. Имено, се пресметува сличноста помеѓу непрецизираните множества кои се однесуваат на непрецизираниот термин кој е асоциран за соодветното примитивно дрво, со непрецизираниот термин за класниот атрибут.

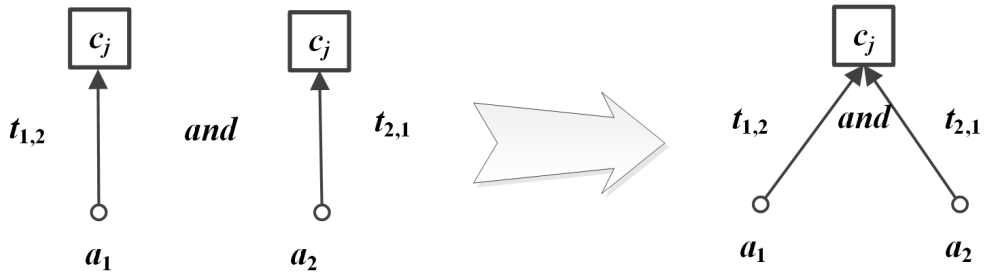
Сепак со користење само на едно примитивно стебло не може да се обезбеди модел кој ќе има задоволителна предиктивна моќ. За таа цел примитивното дрво од нулто ниво за кое е добиена највисока сличност се агрегира со останатите примитивни стебла, и со тоа се формираат кандидатските стебла на прво ниво. Со ова, доколку на нулто ниво има 20 примитивни дрва, тогаш на прво ниво има 19 кандидатски дрва. Потоа се одредува сличноста за секое од кандидатските дрва на прво ниво. Кандидатското дрво на прво ниво за кое е добиена највисока сличност, потоа се агрегира со дрвата од нулто и прво ниво кои не се земени при агрегација, и со тоа се формираат кандидатските дрва на второ ниво. Со користење на 4 влезни карактеристики и 5 непрецизирани термини по карактеристика, се добиваат 37 кандидати стебла на второ ниво. При агрегација може да се агрегираат стебла кои опфаќаат два или повеќе непрецизирани термини кои се однесуваат на иста карактеристика, со цел да се надмине проблемот на осетливост на мала промена во вредностите на карактеристиките настанати како резултат на промените кои настануваат кај протеинските структури во текот на еволуцијата. При агрегација стеблото кое има највисока сличност на последното ниво се зема како лево дете во стеблото кое се формира со спојувањето, а другото стебло кое се зема при агрегирањето се поставува како десно дете. Оваа постапка на агрегација на стеблата се повторува и на следните нивоа, се додека да биде исполнет некој критериум за конвергенција. Бидејќи при агрегацијата стеблото кое има највисока сличност на последното ниво се спојува со останатите стебла, така што тоа досега најдобро стебло станува подстебло во новодобиеното кандидатско стебло, затоа градењето на стеблото е од листовите кон коренот. На Слика 4.5 е даден пример за елементарно и кандидатско дрво, како и агрегираното стебло кое е добиено со нивно спојување со OR операторот за агрегација.

При градење на моделот, кандидатското стебло за кое е добиена највисока сличност на последното ниво може да се спојува со сите примитивни и кандидатски стебла кои до сега не се земени при агрегација, или може да се ограничи да тоа стебло да може да се спојува само со примитивните стебла од нулто ниво кои не се земени при агрегација. На овој начин може да се добијат два типа на модели: едноставен модел и генерален модел. Основната разлика помеѓу едноставниот и генералниот модел е тоа што кај едноставниот модел најдоброто (најсличното) кандидатско стебло од последното ниво може да се спојува само со примитивните стебла, додека кај генералниот модел спојувањето може да се прави и со сите останати кандидатски стебла од претходните нивоа кои не се земени при агрегација.

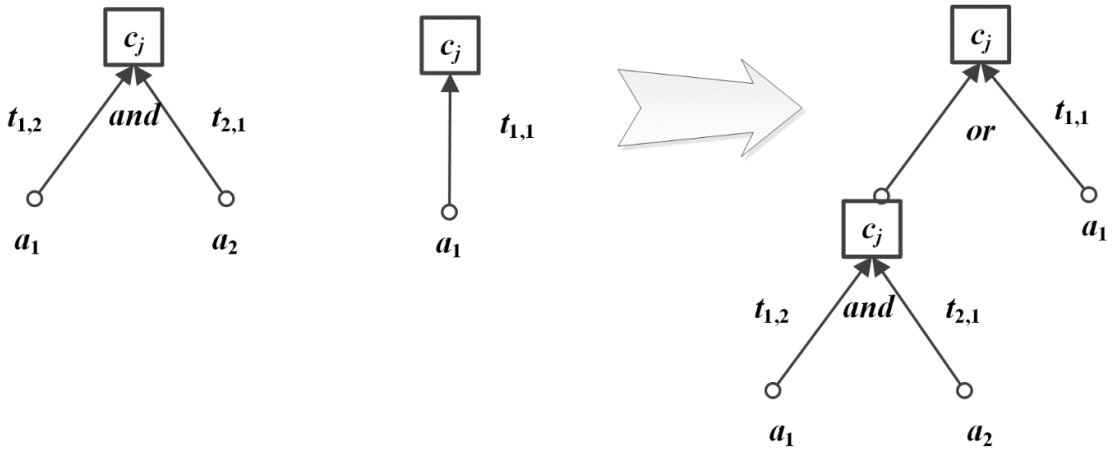


Слика 4.5 Типови на непрецизирани дрва на припадност.

Во [А30] генерирани се едноставни и генерални НДП модели за одредување на индикаторските карактеристики на дијатомеите. Во оваа докторска дисертација моделите кои се генерираат работат со значително поголем број на примероци. Со тоа градењето на генералните НДП модели е со големи мемориски побарувања, па затоа беа генерирани само едноставни НДП модели. На Слика 4.6 е прикажана постапката на агрегација на две примитивни стебла, додека на Слика 4.7 е прикажана постапката на агрегација на најдоброто кандидатско стеблото од последното ниво со примитивно стебло кое не е земено предвид при агрегацијата.



Слика 4.6 Агрегација на две примитивни НДП.



Слика 4.7 Агрегација на кандидатско НДП од прво ниво со примитивно НДП.

Како што беше споменато погоре, постапката на агрегација и формирање на кандидатски стебла на повисоките нивоа се повторува се додека да биде исполнет некој критериум за конвергенција. Еден можен критериум е да се ограничи гранењето на стеблото да оди до некое предефинирано ниво. На овој начин директно се ограничува и комплексноста на моделот. Овој критериум за конвергенција се користи кај методот од-дното-нагоре (174). Во оваа докторска дисертација нивото е поставено на 5, односно тоа значи дека покрај примитивните стебла на нулто ниво, дополнително се генерираат и кандидатските стебла на пет следни нивоа. На крај кандидатското стебло добиено на петто ниво за кое е добиена највисока сличност ќе го претставува конечниот модел за класата за која се гради модел. Сепак со ваквото ограничување да стеблото расте до одредена предефинирана длабочина, не се овозможува прилагодување на комплексноста на стеблото согласно комплексноста на проблемот кој се решава. Па затоа подоцна во методот од-врвот-надолу (175) за градење НДП е воведен друг критериум за конвергенција.

Метод од-врвот-надолу за градење Непрецизирани Дрва на Припадност (НДП)

Во методот од-врвот-надолу предложен во (175), авторите предлагаат неколку подобрувања на методот од-дното-нагоре (174) за градење на непрецизирани дрва на припадност. Со првото подобрување се менува насоката во која се гради стеблото, додека второто подобрување се однесува на воведување на критериум за конвергенција кој зависи од комплексноста на проблемот кој се решава со моделот. Првото подобрување се однесува на насоката во кој се гради моделот. Имено, кај методот од-дното-нагоре (174) стеблото се гради од листовите кон коренот, додека кај методот од-врвот-надолу (175) градењето на стеблото е од коренот кон листовите. Кај методот од-дното-нагоре спојувањето се одвива така што во агрегираното стебло, кандидатското стебло за кое е добиена највисока сличност на последното ниво се поставува како

лево дете, па со тоа новодобиениот модел значително се разликува од моделот во претходната итерација. Со промената на насоката на градење на моделот, врз моделот добиен на тековното ниво се прават мали модификации на следното ниво со цел да истиот пофино ги претставува податоците. Тоа се прави така што даден јазел лист се заменува со подстебло со кое се агрегирани два непрецизирани термини. Во процесот на одлучување кај методот од-врвот-надолу, операторите за агрегација кои се на подолните нивоа имаат помало влијание од операторите кои се поблиску до коренот на стеблото. Второто подобрување кое е воведено кај методот од-врвот-надолу се однесува на критериумот за конвергенција. За таа намена, авторите на трудот (175) направиле анализа на четириесет податочни множества и утврдиле дека во 90% од случаите процесот на градење на стеблото завршува со постигнување на максимално дозволената длабочина на стеблото, а кај останатите 10% од случаите процесот на градење на стеблото завршува доколку нема понатамошно подобрување во стеблото со воведување на дополнителни нивоа. Исто така во податочните множества, одредена класа може да биде претставена со многу едноставен модел кој има мала комплексност (мал број јазли), додека за некои класи има потреба од градење на покомплексни модели со поголема длабочина. Од тука очигледно е дека критериумот за конвергенција дефиниран преку максимално дозволената длабочина на стеблото не е соодветен, бидејќи со овој критериум градењето на моделот може да се прекине или премногу рано или премногу касно бидејќи не се зема предвид комплексноста на проблемот. Со цел да се обезбеди адаптибилен критериум за конвергенција, кај методот од-врвот-надолу се анализира релативното подобрување во две последователни итерации. Имено, градењето на моделот прекинува кога $sim_{max}(t) < (1 + \epsilon)\ sim_{max}(t - 1)$, каде ϵ е вредност во интервалот [0, 1] која е предефинирана од корисникот, а $sim_{max}(t)$ и $sim_{max}(t - 1)$ се сличностите добиени за најдоброто кандидатско стебло во t-тата и $(t - 1)$-вата итерација. Во оваа докторска дисертација градењето на моделот прекинува кога подобрувањето не е поголемо од 25%, односно ϵ е поставено на 0.25 како што е препорачано во (175). Покрај овие две подобрувања, во (175) воведени се и други подобрувања, од кои позначајно е тоа што корисникот не треба да го специфицира типот на функција на припадност, туку истата се одредува од самиот метод така што најдобро да одговара за податочното множество. Кај методот од-врвот-надолу омекнувањето (англ. fuzzification) се прави така што се користат по три непрецизирани термини за секоја карактеристика, а омекнувањето се прави со користење на пристапот опишан во (175). Исто така авторите овозможуваат справување со вредности кои недостасуваат (missing values) во податочното множество.

4.2.3. Експериментални резултати

Во [A27] и [A29] беа воведени НДО за одредување на сврзните делови од протеинската структура. Потоа, во [A28] беше воведен методот од-дното-нагоре за градење на НДП, а во [A33] беа воведени непрецизираните оператори кои вршат усреднување. Во [A32] беше воведен методот од-врвот-надолу за градење на НДП за детекција на сврзните делови од протеинот. Во оваа секција ќе бидат презентирани резултатите од евалуацијата на методите за одредување на сврзните делови од протеините кои се базираат на непрецизираната логика, кои беа опишани во претходната секција. Во оваа анализа се користи истото податочно множество кое се користеше во секција 4.1. Предвид се земаат истите карактеристики кои се користеа во секција 4.1.

Евалуација на методот од-дното-нагоре за градење НДП

Прво ќе биде направена евалуација на моделите добиени со методот од-дното-нагоре за градење на НДП кој е предложен во (174). Беше испитано влијанието на функцијата на припадност користејќи различен број на лингвистички вредности (непрецизирани термини) по карактеристика *T*. Исто така во оваа анализа изградени се модели користејќи различни оператори при агрегација. Во анализите направени во оваа докторска дисертација длабочината на стеблата кои се градат со методот од-дното-нагоре е ограничена на пет. Со ова бројот на нивоа на кои се прави агрегација е пет, односно има вкупно 6 нивоа во стеблото. Во Табела 4.5 прикажани се резултатите за вредностите на AUC-ROC добиени со моделите користејќи ја RMSE метриката за сличност. Може да се забележи дека најдобри резултати се добиваат доколку се користат T=10 непрецизирани термини по карактеристика. Генерално, најдобри резултати се добиваат со користење на триаголната и Гаусовата функција на припадност. Покрај AND и OR операторите за агрегација, доколку дополнително се додадат WA и OWA операторите се зголемува предиктивната моќ на моделите. Но, од друга страна се покажа дека со воведувањето на WG и OWG операторите не се зголеми точноста на моделите. Со цел да се испита влијанието на метриката за сличност, дополнително беше направена анализа користејќи ги RMSE, $RMSE^2$ и G_MAX метриките за сличност. При тоа бројот на функции на припадност по карактеристика е поставен на најдобрата вредност согласно Табела 4.5, а тоа е T=10. Резултатите од оваа анализа се прикажани во Табела 4.6. Од резултатите може да се забележи дека генерално со $RMSE^2$ се постигнуваат полоши резултати отколку ако се користи RMSE метриката за сличност, додека со G_MAX во некои случаи има влошување, а во некои случаи има подобрување на моделите. Моделот добиен со триаголната функција на припадност, користејќи ги AND, OR, WA и OWA операторите за агрегација во комбинација со RMSE метриката за сличност има највисока вредност за AUC-ROC од 0.580, што е најдобриот модел кој се доби со користење на методот од-дното-нагоре за градење НДП.

ОПЕРАТОРИ ЗА АГРЕГАЦИЈА: AND, OR				
Функција на припадност	***T* = 3**	***T* = 4**	***T* = 5**	***T* = 10**
Триаголна	0.557	0.564	0.549	0.571
Трапезоидна	0.534	0.557	0.563	0.552
Гаусова	0.556	0.564	0.550	0.571
Сигмоидална(+1)	0.539	0.530	0.549	0.551
Сигмоидална(-1)	0.542	0.526	0.550	0.551
ОПЕРАТОРИ ЗА АГРЕГАЦИЈА: AND, OR, WA, OWA				
Функција на припадност	***T* = 3**	***T* = 4**	***T* = 5**	***T* = 10**
Триаголна	0.568	0.575	0.573	**0.580**
Трапезоидна	0.551	0.565	0.570	0.567
Гаусова	0.568	0.572	0.572	0.576
Сигмоидална(+1)	0.550	0.553	0.553	0.553
Сигмоидална(-1)	0.550	0.553	0.553	0.553
ОПЕРАТОРИ ЗА АГРЕГАЦИЈА: AND, OR, WG, OWG				
Функција на припадност	***T* = 3**	***T* = 4**	***T* = 5**	***T* = 10**
Триаголна	0.558	0.568	0.538	0.561
Трапезоидна	0.538	0.558	0.563	0.552
Гаусова	0.521	0.564	0.521	0.571
Сигмоидална(+1)	0.549	0.551	0.551	0.551
Сигмоидална(-1)	0.548	0.567	0.550	0.551

Табела 4.5 AUC-ROC за моделите добиени со методот од-дното-нагоре за градење НДП користејќи RMSE метрика за сличност.

ОПЕРАТОРИ ЗА АГРЕГАЦИЈА: AND, OR			
Функција на припадност	**RMSE**	**$RMSE^2$**	**G_MAX**
Триаголна	0.571	0.512	0.571
Трапезоидна	0.552	0.558	0.559
Гаусова	0.571	0.539	0.571
Сигмоидална(+1)	0.551	0.500	0.550
Сигмоидална(-1)	0.551	0.550	0.550
ОПЕРАТОРИ ЗА АГРЕГАЦИЈА: AND, OR, WA, OWA			
Функција на припадност	**RMSE**	**$RMSE^2$**	**G_MAX**
Триаголна	**0.580**	0.512	0.573
Трапезоидна	0.567	0.552	0.570
Гаусова	0.576	0.566	0.573
Сигмоидална(+1)	0.553	0.500	0.551
Сигмоидална(-1)	0.553	0.550	0.550
ОПЕРАТОРИ ЗА АГРЕГАЦИЈА: AND, OR, WG, OWG			
Функција на припадност	**RMSE**	**$RMSE^2$**	**G_MAX**
Триаголна	0.561	0.512	0.571
Трапезоидна	0.552	0.558	0.559
Гаусова	0.571	0.539	0.571
Сигмоидална(+1)	0.551	0.500	0.550
Сигмоидална(-1)	0.551	0.550	0.550

Табела 4.6 AUC-ROC за моделите добиени со методот од-дното-нагоре за градење НДП користејќи T=10 и различни метрики за сличност (RMSE, $RMSE^2$ и G_MAX).

Евалуација на методот од-врвот-надолу за градење НДП

Потоа беа направени анализи за дискриминаторната моќ на моделите добиени со методот од-врвот-надолу за градење НДП кој е предложен во (175). Во оваа анализа се користи RMSE метриката за сличност. Како што беше кажано во претходната секција, кај методот од-врвот-надолу за градење на НДП не се дефинира функцијата на припадност од страна на корисникот, туку самиот модел одбира функција на припадност која најдобро соодветствува за податочното множество. При тоа беше испитана точноста на моделите користејќи различни оператори за агрегација. Резултатите од оваа анализа се прикажани во Табела 4.7. Од резултатите може да се забележи дека генерално моделите во кои AND и OR операторите се земени предвид се подобри од моделите во кои се користат MIN и MAX операторите. Може да се забележи дека со користење на MIN, MAX, AND и OR операторите за агрегација се добиваат подобри резултати отколку ако се користат само MIN и MAX или само AND и OR операторите. Исто така беа направени анализи каде што предвид беа земени и операторите кои вршат усреднување (WA и OWA операторите). Со овие оператори се генерира попрецизен модел отколку ако се користат само MIN и MAX. При тоа доколку на овие оператори за усреднување (WA и OWA) им се придружат MIN и MAX операторите се добива подобра прецизност отколку ако се придружат AND и OR операторите, што не е случај кај другите оператори. Беа направени експерименти и со користење на Ајншајновата норма и конорма, како и на Лукасиевичевата норма и конорма. Ајнштајновите оператори се покажаа како подобри во однос на Лукасиевичевите операторите. Кај Ајнштајновите оператори, со додавањето само на MIN и MAX или само AND и OR се генерираат модели со полоша предиктивна моќ, но со додавање на сите четири оператори се доби модел кој е попрецизен од моделот добиен само со Ајнштајновите оператори. Во однос на операторите на Лукасиевич, со додавање на MIN и MAX или AND и OR, како и со додавање на сите четири оператори се постигна зголемување на дискриминаторната моќ на моделите. Беа направени експерименти каде Ајнштајновите и Лукасиевичевите операторите се земаат заедно со MIN, MAX, AND и OR операторите, и при тоа се покажа дека моделот кај кој се користат сите овие оператори постигнува највисока вредност за AUC-ROC од 0.587. Од оваа анализа може да се заклучи дека генерално со додавањето на дополнителни оператори за агрегација, предиктивната моќ на моделите се зголемува. Сепак, тука треба да се напомене дека времето на тренирање линеарно расте со бројот на оператори, бидејќи бројот на кандидатски дрва линеарно се зголемува со додавање на дополнителни оператори за агрегација.

Со методот од-врвот-надолу се добива малку попрецизен модел (AUC-ROC=0.587) отколку со методот од-дното-нагоре (AUC-ROC=0.580). Времето на тренирање е подолго со методот од-врвот-надолу, но од друга страна со методот од-дното-нагоре тестирањето трае подолго.

Непрецизирани оператори за агрегација	AUC-ROC
MIN, MAX	0.564
AND,OR	0.584
MIN, MAX, AND, OR	**0.586**
WA, OWA	0.581
MIN, MAX, WA, OWA	0.581
AND, OR, WA, OWA	0.573
MIN, MAX, AND, OR, WA, OWA	0.573
Ајнштајнова норма и конорма	0.574
Ајнштајнова норма и конорма, MIN, MAX	0.566
Ајнштајнова норма и конорма, AND, OR	0.558
Ајнштајнова норма и конорма, MIN, MAX, AND, OR	**0.586**
Лукасиевичева норма и конорма	0.544
Лукасиевичева норма и конорма, MIN, MAX	0.569
Лукасиевичева норма и конорма, AND, OR	0.575
Лукасиевичева норма и конорма, MIN, MAX, AND, OR	**0.585**
Ајнштајнова и Лукасиевичева норма и конорма, MIN, MAX	0.566
Ајнштајнова и Лукасиевичева норма и конорма, AND, OR	0.577
Ајнштајнова и Лукасиевичева норма и конорма, MIN, MAX, AND, OR	**0.587**

Табела 4.7 AUC-ROC за моделите добиени со методот од-врвот-надолу за градење на НДП користејќи RMSE метрика за сличност и различни непрецизирани оператори за агрегација.

4.3. Подобрување на моделите преку избор и трансформација на карактеристики

Во претходните секции од ова поглавје, моделите за детекција на сврзните делови од протеинската структура беа генерирани користејќи ги следниве карактеристики на амино-киселинските остатоци: totalASA, avgDPX, avgCX и хидрофобичноста [A27], [A28], [A29], [A32], [A33], [A34]. Сепак, покрај овие четири карактеристики, предвид можат да се земат и останатите карактеристики на амино-киселинските остатоци кои беа опишани во секција 3.6.1. Со користење на техники за избор на карактеристики може да се одредат најрелевантните карактеристики. Иако бројот на карактеристики на амино-киселинските остатоци кои се земаат предвид во оваа дисертација не е премногу голем, сепак анализите се прават со користење на големо податочно множество кое содржи стотици илјади примероци, па затоа намалувањето на димензионалноста е едно од клучните барања кои треба да ги исполниме. Со изборот на карактеристиките, или со трансформација на карактеристиките во помал број на карактеристики, не само што ќе се намали димензионалноста на проблемот и комплексноста на моделите, туку ќе се намалат и времињата на тренирање и тестирање. Исто така може да се постигне и зголемување на дискриминаторната моќ на моделите како резултат на елиминација на нерелевантните карактеристики, а може и да се избегне евентуалното преобучување кое може да настане. Со намалување на комплексноста на моделите, дополнително се обезбедува и

полесна можност за толкување на истите, што е од особена важност доколку сакаме да провериме дали знаењето кои е извлечено со моделите соодветствува на начинот на кој експертите рачно ги донесуваат одлуките за сврзните региони на протеинските структури.

Од седумнаесетте карактеристики на амино-киселинските остатоци кои беа извлечени во секција 3.6.1, се отфрла карактеристиката minDPX бидејќи кај сите површински остатоци кои се земаат предвид во податочното множество оваа карактеристика има вредност 0. Потоа со користење на техники за избор и трансформација на карактеристики се одредува множеството на карактеристики кои ќе се користат при градење на предиктивните модели. Во [A35] беа применети различни техники за избор и трансформација на карактеристики, а потоа беа изградени модели за идентификација на сврзните региони од протеинската структура користејќи класични методи, како и методи базирани на непрецизираната логика. Во продолжение ќе биде даден преглед на најпознатите методи за избор и трансформација на карактеристики кои се користени во ова истражување.

4.3.1. Техники за избор и трансформација на карактеристики

Во литературата постојат различни техники за трансформација на карактеристики. Во оваа докторска дисертација се користи методот за анализа на главните состојки (Principal Components Analysis - PCA) (183), (184) за да се намали бројот на карактеристики. PCA ги трансформира оригиналните корелирани карактеристики во нови карактеристики кои не се корелирани. Оваа трансформација се прави со мапирање на примероците од оригиналниот координатен систем во нов систем чии координатни оски се однесуваат на новодобиените карактеристики. Новите карактеристики претставуваат линеарна комбинација од оригиналните карактеристики и се добиваат како сопствени вектори на коваријансната матрица. Најпрво се пресметува коваријансната матрица чии елементи соодветствуваат на коваријансата помеѓу секој пар од карактеристики. На овој начин се зема предвид корелацијата помеѓу карактеристиките. Користејќи ја коваријансната матрица се пресметуваат сопствените вектори и нивните соодветни сопствени вредности. Сопствената вредност ја покажува варијансата добиена за соодветниот сопствен вектор. Редукцијата на карактеристиките се прави така што се земаат само одреден број на сопствени вектори со најголема сопствена вредност. Во ова истражување беа направени експерименти земајќи доволно сопствени вектори за да се покрие 85% и 97% од варијансата на оригиналните примероци. На овој начин, се добија четири и осум карактеристики, соодветно.

Техниките за избор на карактеристики генерално може да се поделат на техники за филтрирање и техники за обвиткување (англ. wrapper) (185). Главната разлика помеѓу овие две категории е тоа што техниките за обвиткување користат одреден метод за индукција на моделот

за да се идентификува најсоодветното множество на карактеристики. На овој начин, тие целат кон максимизација на финалната објективна функција (на пример класификациската точност при класификација). Од друга страна, техниките за филтрирање оптимизираат некоја друга објективна функција (на пример корелацијата помеѓу карактеристиките и класниот атрибут), која е различна од финалната објективна функција. Покрај овие два типа на техники, постојат и вгнездени шеми каде изборот на карактеристики се прави со индукциски метод. Во ова истражување предвид се земени различни техники за филтрирање и обвиткување при изборот на најрелевантните карактеристики.

Техниките на филтрирање може да се поделат во две категории: техники кои независно ги евалуираат карактеристиките и техники кои евалуираат подмножества од карактеристики. Во ова истражување се користат неколку техники за индивидуално рангирање на карактеристиките преку евалуација на нивната значајност користејќи различни мерки. Прво, се користи хи-квадратниот (χ^2- Chi-square) тест (186) за да се измери зависноста помеѓу испитуваните карактеристики и класниот атрибут. Бидејќи χ^2 тестот може да се користи за одредување на зависноста помеѓу дискретни карактеристики, додека карактеристиките на амино-киселинските остатоци кои се извлечени се континуални, затоа претходно се прави дискретизација користејќи го критериумот предложен од Fayyad и Irani (187). Дискретизираните податоци се користат и за другите техники за избор на карактеристики кои работат со дискретни атрибути. Во χ^2 тестот, очекуваната фреквенција дека испитуваната карактеристика ја има i-тата вредност и примерокот припаѓа на j-тата класа се пресметува како $E_{ij}=N_i*C_j/N$, каде што. N_i е бројот на примероци кои ја имаат i-тата вредност за испитуваната карактеристика, C_j означува број на примероци кои припаѓаат на j-тата класа и N е вкупниот број на примероци. Хи-квадратната статистика χ^2 (186) се пресметува како

$$\chi^2 = \sum_i \sum_j \frac{(A_{ij} - E_{ij})^2}{E_{ij}^2}, \qquad (4.18)$$

каде сумирањето се прави за сите дискретни вредности на испитуваната карактеристика и класниот атрибут. A_{ij} ја означува фактичката (емпириска) фреквенција дека еден примерок ја има i-тата вредност за карактеристиката и припаѓа на j-тата класа. Поголема вредност на χ^2 статистиката покажува поголема зависност со класниот атрибут.

Покрај Хи-квадратната статистика, ја користиме и информационата добивка (InfoGain) (153), (188) и стапката на добивка (GainRatio) (153) како мерки за рангирање на карактеристиките во однос на класниот атрибут. Овие мерки се базираат на ентропија, која за дадено податочно множество S се пресметува како $H(S)= - \Sigma_{i=1,\dots k}\ p(i)*\log_2 p(i)$, каде $p(i)$ е веројатноста дека даден

примерок припаѓа на *i*-тата класа, а со *k* е означен бројот на класи. Во овој случај има две класи (остатоци кои припаѓаат или кои не припаѓаат во сврзен регион). Информационата добивка добиена со избор на дадена карактеристика *f* во податочното множество *S* се пресметува како

$$InfoGain(S, f) = H(S) - \sum_{i \in \text{вредности на } f} \frac{|S_i|}{|S|} H(S_i), \tag{4.19}$$

каде S_i е подмножеството на примероци во кое испитуваната карактеристика *f* ја има својата *i*-та вредност, додека $|S_i|$ и $|S|$ ги означуваат бројот на примероци во множествата S_i и *S* соодветно. На овој начин информационата добивка го покажува намалувањето на ентропијата по изборот на карактеристиката. Сепак, информационата добивка ги фаворизира карактеристиките кои имаат голем број на можни вредности. Поради тоа, во стапката на информациона добивка, информационата добивка се нормализира преку делење со ентропијата на карактеристиката.

Во ова истражување се користи и Relief техниката (189) за рангирање на карактеристиките, каде се користи учење базирано на примероци за да се одредат тежините на карактеристиките. Овие тежини соодветствуваат на рангот на карактеристиката. Процедурата претставена во (189) може да се користи за бинарни проблеми и предвид зема само еден најблизок сосед од секоја класа. Во (190), Relief техниката е проширена за избор на карактеристики кај повеќекласни проблеми, при што предвид се земаат *k*-најблиски соседи од секоја класа. Во ова истражување, пресметката на тежините се прави според процедурата претставена во (190). Најпрво, тежините на карактеристиките W_i се иницијализираат на нула. Потоа секој примерок се одбира и тежините на карактеристиките се ажурираат врз основа на растојанието помеѓу моментно испитуваниот примерок и неговите најблиски соседи во рамки на секоја класа. Нека моментниот примерок биде означен со *X*, и неговиот *j*-ти најблизок сосед од истата и спротивната класа се означуваат како $Hit_{X,j}$ и $Miss_{X,j}$, соодветно. Промената на тежините се прави како во (190) користејќи го равенството

$$W_i = W_i - \frac{d(X, Hit_{X,j}, i)}{N} + \frac{d(X, Miss_{X,j}, i)}{N}, \tag{4.20}$$

каде *d(X,Y,i)* означува апсолутно растојание помеѓу вредностите на *i*-тата карактеристика на примероците *X* и *Y*, а *N* е бројот на примероци. Делењето со *N* се прави за да се задржат тежините во интервалот [-1, 1]. Промената на тежините се прави за сите *k*-најблиски соседи од двете класи, при тоа се изминуваат сите примероци во податочното множество. На овој начин, се пенализира растојанието помеѓу соседите во рамки на една класа, и се фаворизира растојанието со соседите кои припаѓаат на спротивната класа. Дополнително, на разликите за дадена карактеристика помеѓу испитуваниот примерок и неговите соседи може да им бидат доделени тежини врз основа

на рангот *j* на соседите, користејќи тежина exp(- (*j*/ σ)^2), каде што σ е предефиниран параметар. Во ова истражување се користи σ=2. При одредување на тежините предвид се земаат десет најблиски соседи од истата класа и десет најблиски соседи од спротивната класа за секој од испитуваните примероци. При тоа направени се анализи со користење на Relief техниката со и без задавање на тежини на растојанијата помеѓу соседите.

Претходните техники ги рангираат карактеристиките независно. Сепак, овие техники се стремат само да ја максимизираат релевантноста на карактеристиките, но не ја испитуваат нивната редундантност. Постојат различни техники за избор на карактеристики кои имаат за цел да идентификуваат оптимално подмножество на карактеристики за кое се добива оптимална вредност за некоја мерка. Во техниките објаснети во понатамошниот текст, евалуацијата на подмножествата се прави со користење на Пирсоновиот корелациски коефициент (Pearson correlation coefficient - PCC) (191). За дадено множество на карактеристики *SF* со N_F карактеристики, PCC се пресметува како во (191) според

$$\mathrm{PCC} = \frac{N_F \,\mathrm{avg}(\mathrm{correlation}_{FC})}{\sqrt{N_F + N_F (N_F - 1)\,\mathrm{avg}(\mathrm{correlation}_{FF})}}, \tag{4.21}$$

каде avg(correlation$_{FC}$) означува средна вредност од корелацијата помеѓу карактеристиките во *SF* и класната карактеристика, додека avg(correlation$_{FF}$) означува средна вредност од корелацијата помеѓу секој пар на карактеристики во *SF*. Со користење на PCC се цели кон идентификација на подмножество на карактеристики кое има висока корелација со класниот атрибут, и ниски корелации помеѓу самите парови од карактеристики.

Може да се идентификуваат неколку категории на техники на филтрирање кои евалуираат подмножество од карактеристики, како експоненцијални, секвенцијални и рандомизирачки. Во ова истражување се користи исцрпното пребарување (англ. Exhaustive search), каде се испитува секое можно подмножество на карактеристики и се избира оптималното подмножество. Ова пребарување припаѓа во првата категорија бидејќи бројот на можни подмножества расте експоненцијално со бројот на карактеристики.

За да се избегне испитување со примена на груба сила (brute-force) на сите подмножества, може да се користи некое евристично пребарување (192). Во ова истражување се користи алчен пристап чекор-по-чекор кој припаѓа во секвецијалните техники бидејќи секвенцијално додава или отстранува карактеристики. Ова пребарување може да се прави нанапред или наназад. При избирање нанапред (англ. Forward selection), првичното подмножество се иницијализира на празно множество. Потоа, сите можни надмножества кои се добиваат со додавање на една од преостанатите карактеристики се земаат како кандидати надмножества. Надмножеството кое

дава најголем раст на PCC се избира како моментно оптимално множество. Оваа процедура се повторува рекурзивно сè додека не се задоволи некој услов за терминација. Во ова истражување, процедурата запира кога множеството ги содржи сите карактеристики, а потоа се одредува рангот на секоја карактеристика врз основа на вредностите на PCC добиени за сите испитувани подмножества. Елиминација наназад (англ. Backward elimination) се прави во спротивна насока, при што иницијалното множество ги содржи сите карактеристики, а потоа се прави елиминација на карактеристики. Секвенцијалната елиминација на карактеристики се прави со отстранување на најнерелевантните карактеристики. Во оваа докторска дисертација исто така се користи и пребарување прво-по-најдобриот (англ. Best-first search), кое почнува од празно множество или полно множество од карактеристики, а потоа секвенцијално се додаваат или отстрануваат карактеристики. Разликата помеѓу пребарувањето прво-по-најдобриот и претходните две пребарувања (селекција нанапред и елиминација наназад) е тоа што пребарувањето прво-по-најдобриот дозволува двонасочно пребарување, односно додавање или отстранување на една карактеристика во секој чекор, додека останатите го пребаруваат просторот само во една насока. Сепак, за да не се прави целосно пребарување како кај исцрпното пребарување, интензитетот на пребарување наназад во пребарувањето прво-по-најдобриот се ограничува според некој критериум. Во ова истражување се ограничува пребарувањето наназад со прекинување кога во пет последователни чекори нема да се најде подобро подмножество од карактеристики. При тоа, во ова истражување, кај пребарувањето прво-по-најдобриот се почнува со празно почетно множество, и се пребарува нанапред.

Користејќи ги овие алчни пристапи, може да се избегне евалуирање на сите можни подмножества на карактеристики. Сепак, овие техники може да најдат локални минимуми, но тоа може да се избегне со користење на техники на рандомизација. Од оваа категорија на техники во ова истражување се користи генетскиот алгоритам - ГА (англ. Genetic algorithm) (193), кој се базира на природната еволуција. ГА зема предвид генерација од неколку решенија (подмножества од карактеристики во овој случај), и тие решенија еволуираат сè до оптималните решенија. Во контекст на избор на карактеристики, решенијата се бинарни низи каде вредноста на *i*-тата позиција означува дали *i*-тата карактеристика е присутна или отсутна во решението. Нова генерација се формира со комбинирање на решенијата од претходните генерации (се прави вкрстување). Решенијата со поголема прилагоденост (PCC во овој случај) имаат повисока веројатност да бидат избрани како родители при креирањето на нови решенија во следната генерација. По вкрстување може да се појават мутации. Во контекст на избор на карактеристики, мутациите значат дека некои карактеристики може произволно да се додадат или отстранат од множеството добиено со вкрстување. Користејќи случајни мутации, ГА може да избегне локални минимуми. Во ова истражување се генерираат дваесет генерации со дваесет решенија

(подмножества од карактеристики). Веројатноста за мутација е поставена на 0,033 и веројатноста за вкрстување *c* на 0,6 и 1. Со експериментите направени во ова истражување, за двете вредности на *c* се селектираат истите подмножества од карактеристики.

Претходните техники вршат оптимизација на PCC за да ја максимизираат средната корелација помеѓу карактеристиките и класниот атрибут, и да ја минимизираат средната корелација помеѓу карактеристиките. Во поновата литература, сè почесто се користи minimum-Redundancy-Maximum-Relevance (mRMR) техниката (194) за избор на карактеристики. Оваа техника ја минимизира редундантноста и ја максимизира релевантноста во исто време. Во mRMR, наместо со корелација, зависноста (взаемната информација *I*) помеѓу две дискретни карактеристики f_A и f_B се одредува како

$$I(f_A, f_B) = \sum_{i,j} p(f_{A,i}, f_{B,j}) \log_2 \frac{p(f_{A,i}, f_{B,j})}{p(f_{A,i}) p(f_{B,j})}, \tag{4.22}$$

каде $f_{A,i}$ и $f_{B,j}$ се однесуваат на *i*-тата вредност на карактеристиката f_A и *j*-тата вредност на карактеристиката f_B, додека $p(f_{A,i},f_{B,j})$, се однесува на здружената веројатност дека карактеристиките f_A and f_B би ги имале нивните *i*-та и *j*-та вредност, соодветно. Веројатноста дека карактеристиката f_A ја има нејзината *i*-та вредност се означува како $p(f_{A,i})$, додека веројатноста карактеристиката f_B да ја има нејзината *j*-та вредност се означува како $p(f_{B,j})$. При сумирањето, се земаат сите парови на вредности од карактеристиките f_A и f_B. Потоа, релевантноста *D* и редундантноста *R* во рамки на дадено множество на карактеристики *SF* со N_F карактеристики се пресметува како

$$\begin{aligned} D(FS) &= \frac{1}{N_F} \sum_{f_i \in FS} I(f_i, f_{class}) \\ R(FS) &= \frac{1}{{N_F}^2} \sum_{f_i \in FS} \sum_{f_j \in FS} I(f_i, f_j), \end{aligned} \tag{4.23}$$

каде f_{class} е класната карактеристика и таа не припаѓа во *SF*. Методот ја оптимизира објективната функција $\Phi(D,R)$ со цел за да ја максимизира релевантноста и да ја минимизира редундантноста во исто време. Во (195) дадени се две шеми за комбинирање на релевантноста и редундантноста, тоа се: разлика на взаемната информација (Mutual Information Difference - MID) и сооднос на взаемната информација (Mutual Information Quotient - MIQ). Во MID објективната функција е $\Phi(D,R)=D-R$, додека во MIQ објективната функција е $\Phi(D,R)=D/R$. Во оваа докторска дисертација се користат двете шеми. При анализата, пред да се примени mRMR техниката, прво се дискретизираат карактеристиките во десет интервали со еднаква ширина бидејќи mRMR техниката работи само со дискретни атрибути.

Како последна техника за филтрирање во ова истражување се користи техниката предложена од (196), каде се прави рекурзивна елиминација наназад. Релевантноста на карактеристиките се одредува според тежините добиени со класификаторот за градење на машини со носечки вектори (Support Vector Machines - SVM), при што за секое испитувано подмножество од карактеристики се гради посебен SVM модел.

Како што беше спомнато претходно, техниките на обвиткување (185) го користат методот за индукција на модел за да го најдат оптималното подмножество на карактеристики. Подмножествата се евалуираат користејќи ја истата мерка за евалуација која ќе биде користена за да се евалуира предиктивниот модел. На овој начин, може директно да се оптимизира финалната објективна функција (на пример класификациската точност при класификација). При тоа се стремиме да се најде оптималното подмножество на карактеристики во насока нанапред. Во однос на методот за индукција на модел, во ова истражување се применети следните добро познати класификатори: C4.5 дрва на одлука (153), Наизменично одлучувачко стебло (Alternating Decision Tree - ADTree) (159), Наивен Баесов класификатор (Naïve Bayes) (154), Наивно Баесово дрво (Naïve Bayes Tree - NBTree) (163), Баесова мрежа (Bayesian Network) (164) и *k*-најблиски соседи (*k*-nearest neighbors – *k*-nn) (155). Во оваа докторска дисертација се користи вкрстена валидација со два превои за да се одреди класификациската точност која е објективната функција која се оптимизира со техниката на обвиткување.

За mRMR се користи имплементацијата достапна на (197), додека за другите техники за избор на карактеристики се користи Weka софтверот (156). За секоја техника се користат предефинираните поставки, освен ако не е поинаку наведено.

4.3.2. Експериментални резултати

Во оваа секција ќе бидат направени повеќе анализи за предиктивната моќ на моделите добиени со користење на различни класификатори во комбинација со различни техники за избор и трансформација на карактеристики. За градење на моделите се користат неколку класични класификатори, и тоа: C4.5 дрва на одлука (153), Наизменично одлучувачко стебло (Alternating Decision Tree - ADTree) (159), Функционално дрво (Functional Tree - FTree) (161), Наивен Баесов класификатор (Naïve Bayes) (154), Наивно Баесово дрво (Naïve Bayes Tree - NBTree) (163) и Баесова мрежа (BayesNet) (164). При генерирање на моделите исто така предвид се земени методите од-дното-нагоре (174) и од-врвот-надолу (175) за градење на НДП.

Опис на податочните множества

Во првите анализи кои ќе бидат презентирани во оваа секција се користи истото податочно множество кое се користеше во секција 4.1. Имено од BIND (48) базата на податоци која содржи знаење за сврзните делови од протеинските структури откриено по експериментален пат, со користење на ASTRAL методот (145) се одбира репрезентативно множество преку филтрирање на протеинските ланци кои имаат помалку од 20% секвентна сличност. Множеството за тестирање се формира од амино-киселинските остатоци на 3530-те ланците кои имаат помалку од 10% секвентна сличност, а од остатоците на преостанатите 633 ланци се формира множеството за обука. Потоа се филтрираат површинските остатоци, по што остануваат 115579 остатоци во множеството за обука и 625939 остатоци во множеството за тестирање. Во множеството за обука само 15696 (13.58%) амино-киселински остатоци се дел од сврзен регион според BIND базата. Затоа за да се избегне градење на модели кои се наклонети кон доминантната класа, множеството за обука се балансира преку одбирање на неговите примероци се до 27% од неговата големина. При тоа даден примерок само еднаш се зема предвид, и се формира множество во кое има балансираност на двете класи. Балансирањето се прави само над множеството за обука, а множеството за тестирање, кое е означено како B3530, останува небалансирано. Потоа влезните карактеристики се нормализираат така што се доведуваат во интервалот [0,1]. Множеството B3530 се користи за евалуација на моделите добиени со користење на различни класификатори во комбинација со различни техники за избор и трансформација на карактеристики.

Во оваа секција исто така ќе биде направена споредба со неколку познати методи за детекција на сврзните делови од протеинската структура. При тоа предвид се земени методи кои се базираат на растојанието помеѓу остатоците (198), (152), (199), методи кои го испитуваат зачувувањето на секвенцата и/или структурата (28), (129), (200), како и методи кои се базираат на идентификација на џебови (англ. pocket) (201), (202), (203). Исто така предвид се зема и ConCavity методот (204) кој ги комбинира конзервацијата на секвенцата и идентификацијата на џебови. За споредба, исто така предвид се зема методот за одредување на протеинските интеракции преку структурно поклопување (Protein Interactions by Structural Matching - PRISM) (129), (130), (131), кој ги зема предвид зачувувањето на секвенцата и структурата за да ги одреди сврзните делови на структурите кои се користат како шаблони. Потоа, со PRISM методот сврзните делови на даден примерок се одредуваат преку структурно порамнување со сите шаблони структури. За евалуација на методите предложени во (198), (152), (28), (129) се користи PSAIA софтверот (152), додека за останатите методи се користат веќе пресметаните предвидувања кои се достапни на PRISM, PRINT и ConCavity веб страните. Во оваа споредба се

користат три податочни множества. B1549 множеството за тестирање е формирано така што од B3530 се земени предвид амино-киселинските остатоци на протеинските ланци за кои беа добиени предвидувања со користење на методите презентирани во (131), (198), (152), (199), (28), (129), (200) и (204).

Останатите податочни множества кои се користат во анализата го содржат знаењето кое е сместено во LigASite верзија 7.0 базата на податоци (205), која содржи податоци за биолошки релевантните сврзни делови со познатите апо-структури. Оваа база на податоци содржи две множества, редундантно и нередундантно множество (во кое се опфатени протеински ланци со помалку од 25% секвентна сличност). Множеството за обука се формира така што предвид се земаат 703-те ланци кои ги има во редундантното, а ги нема во нередундантното множество. За да се избегне градење на модели кои се наклонети кон доминантната класа, множеството за обука се семплира до 20% од неговата големина без замена на примероците и следејќи рамномерна дистрибуција на класниот атрибут. Потоа L542 множеството за тестирање се формира од 542-та ланци кои се содржат во нередундантното множество. Ова податочно множество се користи за споредба со методите за кои се добија резултати за овие 542 ланци. Бидејќи за дел од методите нема достапни резултати за сите 542 ланци, затоа конечно се формира и L213 множеството за тестирање така што од L542 множеството предвид се земаат остатоците на ланците за кои се добија предвидувања со сите методи кои се споредуваат во оваа секција.

Во Табела 4.8 е даден преглед на сумарните статистики на податочните множества кои се користат во анализите кои ќе бидат презентирани во оваа секција.

Множество за обука	**BIND**		**LigASite**	
#протеински ланци	663		703	
#остатоци (примероци)	115579		132773	
процент на примероци во позитивната класа	13.58%		9.86%	
Множество за тестирање	**B3530**	**B1549**	**L542**	**L213**
#протеински ланци	3530	1549	542	213
#остатоци (примероци)	625939	277735	105408	37886
процент на примероци во позитивната класа	14.74%	16.42%	10.29%	11.3%

Табела 4.8 Сумарни статистики на податочните множества. За множествата за обука прикажани се карактеристиките пред балансирање на множеството.

Евалуација на моделите

Прво се користи BIND множеството за обука на моделите, и B3530 множеството за евалуација на моделите. Најнапред, беа споредени моделите добиени користејќи ги сите карактеристики и моделите добиени со користење само на карактеристики кои се користеа во анализите направени во претходните секции (тоа се totalASA, avgDPX, avgCX и хидрофобичноста). Резултатите од оваа анализа дадени во Табела 4.9 покажуваат дека генерално кога се зема целото множество на карактеристики, предиктивната моќ на моделите се намалува (освен за ADTree и BayesNet). Од тука може да се заклучи дека со користење на повеќе карактеристики не значи дека ќе се добијат подобри модели. Исто така времињата за тренирање и тестирање и комплексноста на моделот се зголемуваат со порастот на бројот на карактеристики. На пример, првото C4.5 дрво (користејќи ги сите карактеристики) е генерирано за 54 секунди и има 3015 јазли, додека второто C4.5 дрво (користејќи 4 карактеристики) е генерирано за 11 секунди и има 243 јазли. Тестирањето за сите примероци користејќи ги овие C4.5 дрва трае 13 и 5 секунди, соодветно.

Следно, беа применети претходно опишаните техники за избор и трансформација на карактеристики. Во Табела Д1 во Додатокот се дадени детали за селектираните карактеристики со секоја техника. TotalASA, non-polarASA, totalRASA, трите варијанти на CX и хидрофобичноста се селектираните карактеристики од повеќето од техниките. Во Табела 4.10 прикажани се детали за потребното време (во секунди) за секоја од техниките. Селекцијата на карактеристики трае подолго кога се применува Relief техниката, што е очекувано бидејќи секој примерок се споредува со сите други примероци. Селекцијата со SVM трае подолго бидејќи за секое испитувано подмножество од карактеристики се генерира SVM модел. Извршувањето на техниките за обвиткување трае значително повеќе од другите, бидејќи за секое испитувано подмножество од карактеристики се генерира посебен класификациски модел.

Беа направени експерименти со филтрирање на прворангираните 4, 8 и 10 карактеристики, и резултатите покажуваат дека дополнителните карактеристики не секогаш го подобруваат предвидувањето. Поради тоа, во понатамошната анализа од техниките кои прават рангирање на карактеристики се земаат само четирите најдобро рангирани карактеристики. Користејќи го C4.5 класификаторот, најпрецизниот модел се добива со користење на MID шемата од mRMR. Исто така карактеристиките идентификувани со некои техники за обвиткување, и множеството од 4 карактеристики кои се користеше претходно (во секција 4.1 и секција 4.2) се соодветни за комбинирање со C4.5. Користејќи InfoGain и GainRatio за проценување на значајноста на карактеристиките, не се добија најдобрите C4.5 дрва и покрај тоа што C4.5 ја користи стапката на информациона добивка за изборот на најдобрите карактеристики во секој јазол. Причината за овие

резултати е изборот на карактеристики каде се рангираат карактеристиките и се наоѓа најоптимална карактеристика што треба да биде испитувана во коренот на дрвото. Најдобро рангираните карактеристики имаат висока добивка во коренот, но во јазлите подолу во дрвото може да се добие дека се ирелевантни бидејќи може да имаат висока корелација со карактеристиките кои се испитуваат во погорните јазли кои се поблиску до коренот од дрвото.

Класификатор	C4.5	NB	NBTree	ADTree	FTree	BayesNet	Од-дното-нагоре НДП триаголна	Од-дното-нагоре НДП трапезоидна	Од-дното-нагоре НДП Гаусова	Од-врвот-надолу НДП
сите карактеристики	0.564	0.565	0.576	**0.562**	0.582	0.579	NA	NA	NA	NA
totalASA, avgDPX, avgCX, хидрофоб.	0.587	0.567	0.586	0.546	0.589	0.576	0.541	0.563	0.541	**0.586**
PCA	0.577	**0.574**	0.570	0.551	0.581	0.572	0.555	0.531	0.545	0.572
ChiSquared	0.568	0.565	0.572	0.560	0.576	0.575	**0.569**	**0.564**	0.567	0.578
InfoGain	0.568	0.565	0.572	0.560	0.576	0.575	**0.569**	**0.564**	0.567	0.578
GainRatio	0.564	0.560	0.566	0.552	0.578	0.572	0.561	0.563	0.543	0.569
Relief unweighted	0.582	0.554	**0.589**	0.546	0.584	**0.588**	0.541	0.535	0.541	0.568
Relief weighted	0.561	0.552	0.568	0.552	0.566	0.568	0.547	0.530	0.548	0.568
Exhaustive/Best-first/ Genetic	0.583	0.567	0.583	**0.562**	0.589	0.587	NA	NA	NA	NA
Forward selection	0.577	0.567	0.571	0.552	0.578	0.579	0.567	**0.564**	0.568	0.575
Backward elimination	0.577	0.567	0.571	0.552	0.578	0.579	0.567	**0.564**	0.568	0.575
mRMR (MID)	**0.590**	0.555	0.586	0.546	**0.591**	0.586	0.554	0.549	0.556	0.560
mRMR (MIQ)	0.566	0.558	0.568	0.541	0.569	0.563	0.555	0.559	0.555	0.547
SVM	0.583	0.567	0.570	0.553	0.585	0.580	0.565	**0.564**	0.566	0.579
Wrapper_C4.5	0.583	0.565	0.578	0.546	0.585	**0.588**	0.541	0.562	0.542	0.569
Wrapper_ADTree	0.569	0.542	0.567	0.560	0.571	0.562	0.536	0.561	0.554	0.558
Wrapper_NB	0.568	0.565	0.570	0.561	0.573	0.573	0.557	0.562	0.558	0.564
Wrapper_NBTree	0.585	0.547	**0.589**	0.546	0.590	0.587	0.541	0.531	0.541	0.576
Wrapper_BayesNet	0.587	0.565	0.577	**0.562**	0.589	0**.588**	0.545	**0.564**	**0.570**	0.585
Wrapper_KNN	0.586	0.548	**0.589**	0.546	0.589	0.587	0.541	0.533	0.541	0.570

Со задебелени бројки се означени најголемите вредности на AUC-ROC за секој од класификаторите.
NA означува дека не се добил резултат во разумно време.

Табела 4.9 AUC-ROC добиен со различни класификатори и множества од карактеристики. Кај рангирачките техники земени се првите четири најдобро рангирани карактеристики.

Користејќи го NB класификаторот се добива помал AUC-ROC. Највисокиот AUC-ROC користејќи NB се добива на множеството од карактеристики добиено со PCA, бидејќи NB претпоставува дека карактеристиките се независни, а PCA ги трансформира оригиналните корелирани карактеристики во нови некорелирани карактеристики. NBTree класификаторот постигнува највисока AUC-ROC користејќи ги множествата од карактеристики идентификувани од Wrapper_KNN, нетежински Relief (Relief unweighted) и Wrapper_NBTree. ADTree класификаторот

генерираше модели со мала предиктивна моќ. Треба да се спомне дека за другите класификатори користејќи техники за обвиткување, најлошиот модел се добива со користење на Wrapper_AD-Tree. Сепак, користејќи ADTree класификатор, обвиткувачот Wrapper_ADTree прави подобар избор од повеќето други обвиткувачи. FTree моделите генерално покажуваат најдобри перформанси. Со користење на FTree класификаторот со четирите карактеристики избрани од MID шемата од mRMR се добива највисок AUC-ROC. Исто така, FTree дава подобри резултати на множествата од карактеристики избрани со исцрпно пребарување, пребарување прво-по-најдобриот, генетски алгоритми и некои обвиткувачи. Што се однесува до BayesNet класификаторот, најдобри резултати се добиваат со користење на карактеристиките избрани со соодветниот обвиткувач (Wrapper_BayesNet).

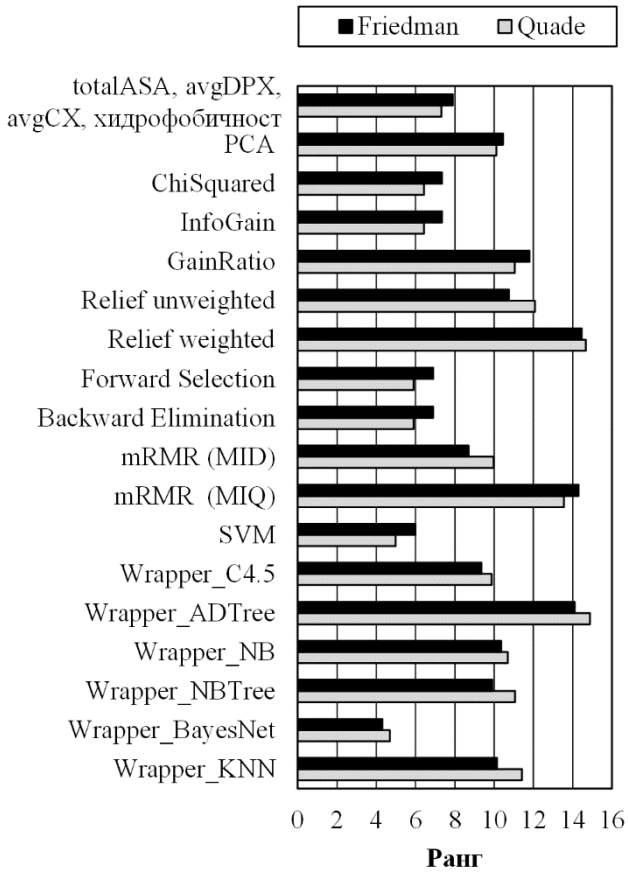
Воопшто, методот од-врвот-надолу за градење на НДП генерира попрецизни модели отколку методот од-дното-нагоре. Сепак, со пристапот од-врвот-надолу најдобри резултати се добиваат со користење на множеството од карактеристики кое се користеше во претходните секции. Имено, ниедна од техниките за избор на карактеристики не најде подобро множество од четири карактеристики соодветни за овој метод. Што се однесува до методот од-дното-нагоре, со користење на техниките за избор на карактеристики се идентификуваат пооптимални карактеристики отколку четирите карактеристики кои се користеа претходно. Кај mRMR, MID се покажа како подобра шема од MIQ. Нетежинскиот Relief (Relief unweighted) метод дава подобри резултати отколку тежинскиот Relief (Relief weighted). Кај техниките за обвиткување, ако се користи методот за индукција на модел при изборот на карактеристики се добиваат попрецизни модели. Највисок AUC-ROC од 0.591 се добива користејќи FTree класификатор врз множеството од четирите најдобро рангирани карактеристики избрани со MID шемата од mRMR техниката. Вториот најдобар модел се добива со C4.5 класификаторот користејќи го истото множество од карактеристики. Речиси во сите експерименти поголема прецизност се постигнува користејќи подмножество од карактеристики наместо со користење на сите карактеристики. Генерално, со избор на најрелевантните карактеристики се добиваат попрецизни модели.

Техника за избор и трансформација на карактеристики	Време (секунди)
PCA	19
ChiSquared	40
InfoGain	34
GainRatio	35
Relief unweighted	530
Relief weighted	534
Exhaustive search	39
Forward selection	34
Backward elimination	34
Best-first search	36
Genetic search	29
mRMR (MID и MIQ)	30
SVM	210
Wrapper_C4.5	333
Wrapper_ADTree	20685
Wrapper_NB	820
Wrapper_NBTree	11840
Wrapper_BayesNet	869
Wrapper_KNN	22631

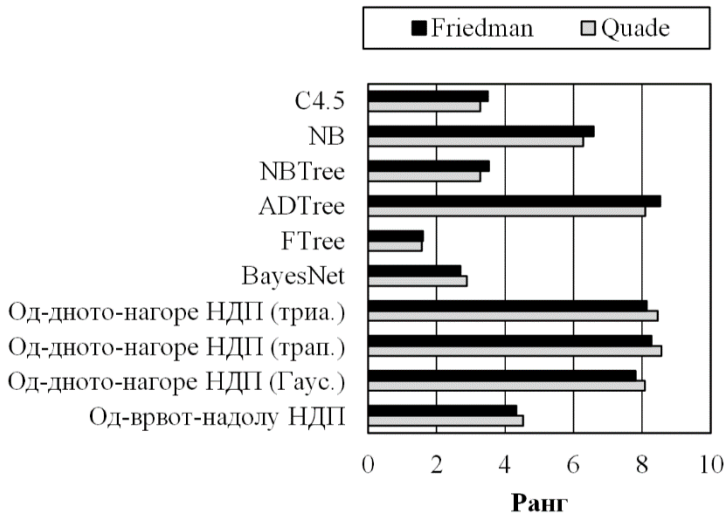
Табела 4.10 Време (секунди) потребно за селекција на најрелевантните карактеристики користејќи различни техники за избор и трансформација на карактеристики.

Во Табела Д2 дадена во Додатокот се прикажани времињата потребни за тренирање и тестирање при користење на различни класификатори и множества од карактеристики. Може да се забележи дека генерирањето на ADTree, од-врвот-надолу НДП и NBTree моделите трае подолго отколку кај другите модели. Од друга страна времето на тестирање е подолго кога се користат FTree и од-дното-нагоре НДП методите.

Со цел да се добие појасна слика за севкупниот ранг на класификаторите и техниките за избор на карактеристики, беше направено севкупно рангирање користејќи ги Friedman и Quade тестовите (206). Бидејќи за НДП методите генерирани се модели само со користење на четири карактеристики, затоа само множествата со четири карактеристики се земени предвид при рангирањето. Резултатите од рангирањето се покажани на Слика 4.8 и Слика 4.9. Помала вредност на рангот означува подобро рангирање. Може да се забележи дека множеството од карактеристики идентификувано со Wrapper_BayesNet има најдобар ранг, додека множеството добиено со тежинскиот Relief (Relief weighted) има најлош ранг. Од класификаторите, FTree има најдобра предиктивна моќ, додека најлош класификатор е од-дното-нагоре НДП класификаторот користејќи трапезоидна функција на припадност.



Слика 4.8 Рангирање на множествата од четири карактеристики со користење на Friedman и Quade тестовите.



Слика 4.9 Рангирање на класификаторите со користење на Friedman и Quade тестовите.

Споредба со неколку постоечки методи за детекција на сврзните делови од протеинската структура

Следно, предложениот пристап беше спореден со неколку постоечки методи за предикција на сврзните делови од протеинските структури. Во споредбата предвид се земени методи кои се базираат на растојанието помеѓу остатоците (198), (152), (199), методи кои ја анализираат конзервацијата на секвенцата и/или структурата (28), (129), (200), како и неколку методи кои вршат детекција на џебови (англ. pocket) (201), (202), (203). Исто така направена е споредба и со ConCavity методот (204), кај кој со Jensen-Shannon divergence (JSD) методот (200) се врши анализа на зачувувањето на секвенцата, и дополнително се врши и детекција на џебови. Во споредбата исто така предвид е земен и PRISM методот (129), (130), (131), кај кој врз основа на конзервацијата на секвенцата и структурата се врши детекција на сврзните делови на структурите кои се користат како шаблони. Во процесот на детекција со PRISM методот, испитуваниот протеин структурно се споредува со сите шаблони структури, и предвидувањето на сврзните делови се базира на сврзните делови од најсличниот шаблон. Може да се забележи дека Concavity и PRISM комбинираат по два различни пристапи за донесување на одлуки во предвидувањето. За предложениот пристап беа направени експерименти користејќи ги класификаторите и техниките за избор на карактеристики кои беа подобро рангирани во претходните анализи. Во Табела 4.11 се презентирани резултатите од споредбата на предложениот пристап со постоечките методи за детекција на сврзните делови од протеинската структура. Во оваа анализа се користат B1549, L542 и L213 множествата за тестирање.

Метод	**Тип**	**Референца**	**B1549**	**L542**	**L213**
Предложениот пристап		[A35]	0.595	0.626	0.607
PRISM	ЗДП	(131)	**0.787**	NA	0.538
Atom nucleus distance	Р	(198)	**0.833**	0.522	0.530
PIADA	Р	(152)	**0.832**	0.523	0.531
PRINT	Р	(199)	**0.775**	NA	0.548
ASA change	К	(28)	**0.820**	0.530	0.543
Van der Waals distance	К	(129)	**0.840**	0.517	0.525
JSD	К	(200)	0.537	0.609	0.608
LigSite	Џ	(201)	NA	**0.792**	**0.744**
PocketFinder	Џ	(202)	NA	**0.804**	**0.773**
Surfnet	Џ	(203)	NA	**0.785**	**0.741**
ConCavity LigSite	К+Џ	(204)	0.587	**0.835**	**0.800**
ConCavity PocketFinder	К+Џ	(204)	NA	**0.836**	**0.809**
ConCavity Surfnet	К+Џ	(204)	NA	**0.819**	**0.791**

NA означува дека не се добиени предвидувања за даденото множество.
Тип ЗДП означува ЗД порамнување, Р означува базиран на растојание,
К означува базиран на конзервација и Џ означува базиран на детекција на џебови.

Табела 4.11 AUC-ROC добиен со различни методи.

Предложениот пристап покажува подобри перформанси од JSD методот (200) која е базиран на анализа на конзервација на секвенцата. Резултатите покажуваат дека постоечките методи кои се земани предвид во оваа споредба покажуваат различни перформанси на различни множества. Ова е затоа што овие методи се дизајнирани за одредување на сврзните региони кај специфична група на интеракции, па затоа се соодветни за одредена група на протеините, но не нудат генерален модел за идентификација на сврзните делови. Од друга страна, предложениот пристап постигнува приближно слични резултати на сите податочни множества, па со тоа е многу постабилен. Предложениот пристап е генерален, бидејќи моделот се гради користејќи множество за обука од различни протеини, па со тоа не се фокусираме на градење на модел само за одредени специфични интеракции. Се очекува дека со предложениот пристап ќе се добијат модели со подобра предиктивна моќ доколку истиот се примени за специфична група на протеини и интеракции. Со предложениот пристап се овозможува само-адаптибилност бидејќи при градење на модели за одредена група на протеини/интеракции, со техниките за избор и трансформација на карактеристики ќе се одберат најзначајните карактеристики за таа специфична група на протеини кои се релевантни за донесување на одлуките. Од друга страна, постоечките методи со кои е направена споредбата, не нудат можност за само-адаптација за градење на посебни модели кои се прилагодени за одредена група на интеракции. Дополнително, предложениот пристап може да биде применет за градење на генерален модел кој ќе се формира како каскада од посебни подмодели, при што секој подмодел ќе се самоадаптира за соодветниот тип на интеракции за кои е наменет тој подмодел.

5

ОДРЕДУВАЊЕ НА ФУНКЦИИТЕ НА ПРОТЕИНСКИ СТРУКТУРИ

Главната цел на истражувањето кое е направено во оваа докторска дисертација е да се одредат функциите кои ги вршат протеинските структури во процесите во живите организми во кои учествуваат. Како што беше кажано уште во воведното поглавје, знаењето за протеинските функции може да се искористи за дизајн на лекарства со што би се стимулирале или спречиле разни процеси во организмите. Во ова поглавје ќе бидат презентирани два пристапа за одредување на протеинските функции. Кај првиот пристап, функциите на даден непознат протеин се одредуваат врз основа на функциите на неговите најблиски соседи, при што најблиските соседи се одредуваат со примена на методите за пребарување на слични протеински структури кои беа презентирани во поглавје 3. Вториот пристап се базира на анализа на карактеристиките на амино-киселинските остатоци кои формираат сврзен регион (интерфејс), а предиктивниот модел се гради со примена на методи за повеќезначна класификација.

Проблемот за функционално анотирање на протеински структури кој се анализира во ова поглавје претставува проблем на повеќезначна класификација, при што треба да се предвиди кои од функциите треба да му се доделат како анотации на испитуваниот протеин. Затоа прво ќе биде дадена формална дефиниција за повеќезначна класификација, а потоа ќе бидат презентирани евалуациските метрики кои може да се користат за одредување на предиктивната моќ на изградените модели.

5.1. Повеќезначна класификација (дефиниција и евалуациски мерки)

При класификација на примероци, целта е според дадени влезни карактеристики, кои можат да бидат номинални (дискретни) или континуални, да се изгради модел со кој ќе се предвидува вредноста на класниот атрибут кој претставува номинална променлива. Кај повеќекласна класификација се гради модел со кој се предвидува еден класен атрибут кој може да прима произволен број на вредности. Доколку моделот треба наеднаш да предвидува вредности за повеќекласни атрибути, тогаш станува збор за решавање на проблем за повеќезначна класификација, како што е проблемот на функционална анотација на протеински структури.

Повеќекласна класификација и повеќезначна класификација

Проблемот на ***повеќекласна*** (англ. multi-class) ***класификација*** формално може да се дефинира на следниот начин. Нека X е просторот на влезни карактеристики, каде секоја карактеристика може да биде номинална или континуална. Ако бројот на влезни карактеристики се означи со D, тогаш за i-тиот примерок од податочното множество влезните карактеристики се дефинирани со векторот $\mathbf{x_i}=(x_{i,1}, x_{i,2},\ldots, x_{i,D})$. Нека класниот атрибут кој треба да се предвиди со повеќекласна класификација прима L различни номинални вредности, тогаш множеството $C=\{c_1, c_2,\ldots, c_L\}$ е множеството од сите класи. Податочното множество S може да се дефинира како $S = \{(\mathbf{x_i},c_i)|\mathbf{x_i} \in X, c_i \in C, i = 1, 2,\ldots N\}$, каде N е бројот на примероци. Целта на повеќекласната класификација е да се најде функција со која се прави мапирање $X\rightarrow\{c_1, c_2,\ldots, c_L\}$ при што се оптимизира одреден критериум според кој се мерат перформансите на моделите.

Кај ***повеќезначната*** (англ. multi-label) ***класификација*** се предвидуваат одреден конечен број на класни карактеристики. Класните карактеристики уште ќе се нарекуваат ознаки или лабели. Целта е да се генерира модел со кој за секој непознат примерок ќе се предвидува множеството на релевантни функционални лабели. Нека има L ознаки кои припаѓаат во множеството $F=\{f_1, f_2,\ldots f_L\}$. Бидејќи целта е да се одреди кои ознаки треба да му се доделат на даден непознат примерок, затоа проблемот се сведува на одредување на L бинарни променливи кои одговараат на функционалните лабели $f_1, f_2,\ldots f_L$. Податочното множество S може да се дефинира како $S=\{(\mathbf{x_i},A_i)|\mathbf{x_i} \in X, A_i \subseteq F, i=1, 2,\ldots N\}$, каде N е бројот на примероци, а A_i е множеството од ознаки кои ги има i-тиот примерок. Целта на повеќезначната класификација е да се најде функција со која се прави мапирање $X\rightarrow 2^F$ при што треба да оптимизира некој критериум според кој се одредуваат перформансите на моделите. Кај повеќезначната класификација не се зема предвид евентуалната зависност помеѓу излезните карактеристики (ознаките). Доколку сакаме да ја земеме предвид зависноста на ознаките, тогаш треба да се користи хиерархиска повеќезначна класификација, но во ова истражување се задржуваме на повеќезначната класификација.

Евалуациски мерки

При градење на модели за решавање на проблеми за повеќезначна класификација, треба да се користат евалуациски мерки кои се соодветни за евалуација на ваков тип на модели. Во продолжение накратко ќе бидат презентирани евалуациските мерки кои се предложени во (207) и кои се користат во оваа дисертација. Евалуациските мерки за повеќезначна класификација може да се поделат во две групи: мерки за бинарна поделба и мерки базирани на рангирање. Кај првата група се врши споредба на вистинските ознаки на примероците со предвидените ознаки. При тоа во оваа група се разликуваат мерки базирани на примерок и мерки базирани на ознака. Кај мерките базирани на примерок се врши споредба на вистинските и предвидените мерки врз множеството од примероци и се прави усреднување врз сите примероци, додека кај мерките базирани на ознака се одредуваат мерките посебно за секоја ознака, и потоа се прави усреднување врз сите ознаки. Во оваа докторска дисертација од мерките базирани на примерок предвид се земаат следниве мерки: Хамингова загуба (Hamming loss), прецизност (Precision), одзив (Recall), F_1, точност (Accuracy) и класификациска точност (Classification accuracy). Од мерките базирани на ознака во ова истражување се користат следниве мерки: макро прецизност ($\text{Precision}_{\text{macro}}$), макро одзив ($\text{Recall}_{\text{macro}}$), макро F_1 ($F_{1\ \text{macro}}$), макро AUC-ROC ($\text{AUC-ROC}_{\text{macro}}$), микро прецизност ($\text{Precision}_{\text{micro}}$), микро одзив ($\text{Recall}_{\text{micro}}$), микро F_1 ($F_{1\ \text{micro}}$), микро AUC-ROC ($\text{AUC-ROC}_{\text{micro}}$). Овие мерки можат да се користат ако излезните карактеристики се бинарни, т.е. ако при предвидувањето се одредува дали примерокот ќе ја има или ќе ја нема дадена ознака. Доколку при градење на моделот се користи метод која на излез дава континуални вредности кои покажуваат колкава е можноста даден примерок да ја има дадена ознака, тогаш треба да се усвои некој праг за да се добијат бинарни вредности со цел потоа да може да се пресметаат овие мерки. За разлика од групата на мерки за бинарна поделба, кај мерките за рангирање излезот од моделите треба да биде вектор во кој е дефиниран предвидениот ранг на секоја ознака за испитуваниот примерок, па одредувањето на мерката се прави преку споредба на рангираните ознаки со вистинските ознаки. Во оваа група припаѓаат повеќе мерки, како на пример: една грешка (One-error), опфатност (Coverage), загуба при рангирање (Ranking loss) и просечна прецизност (Average precision). Во ова истражување ќе бидат користени мерките базирани на бинарна поделба.

Нека со A_i и P_i се означени множествата кои ги содржат вистинските и предвидените ознаки за i-тиот примерок за тестирање, соодветно. Мерките базирани на примерок се пресметуваат со равенствата (5.1), каде со $|S|$ е означена големината на дадено множество S, а со Q е означен бројот на примероци за тестирање. Во равенството за Хамингова загуба, со $A_i \Delta P_i$ се бележи симетричната разлика помеѓу двете множества, што всушност одговара на XOR операција во

Булова логика. Од равенството за класификацискатa точност може да се забележи дека за даден примерок се смета дека се точно предвидени ознаките само доколку целосно се совпаѓаат множествата со вистинските и предвидените ознаки, A_i и P_i соодветно. Со тоа оваа мерка е премногу строга. Доколку се погреши само една од ознаките тогаш ќе се добие иста класификациска точност како и во случај на комплетно грешење на сите ознаки за примерокот. Па затоа оваа мерка нема да се користи во анализите.

$$\text{Hamming\_loss} = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{L}\left|A_i \Delta P_i\right|$$

$$\text{Precision} = \frac{1}{Q}\sum_{i=1}^{Q}\frac{\left|A_i \cap P_i\right|}{\left|P_i\right|} \qquad \text{Recall} = \frac{1}{Q}\sum_{i=1}^{Q}\frac{\left|A_i \cap P_i\right|}{\left|A_i\right|} \tag{5.1}$$

$$\text{F}_1 = \frac{1}{Q}\sum_{i=1}^{Q}\frac{2\left|A_i \cap P_i\right|}{\left|A_i\right|+\left|P_i\right|} \qquad \text{Accuracy} = \frac{1}{Q}\sum_{i=1}^{Q}\frac{\left|A_i \cap P_i\right|}{\left|A_i \cup P_i\right|}$$

$$\text{Classification accuracy} = \frac{1}{Q}\sum_{i=1}^{Q} I(A_i = P_i), \text{каде} \quad I(x) = \begin{cases} 1, x = \text{вистина} \\ 0, x = \text{невистина} \end{cases}$$

Во однос на мерките базирани на ознака, предвид може да се земе било која евалуациска мерка за бинарна класификација, пример: прецизност, одзив, точност, F_1, AUC-ROC и други. Нека со B(*TP*, *TN*, *FP*, *FN*) се означи дадена евалуациската мерка B која е дефинирана преку *TP*, *TN*, *FP* и *FN*, каде *TP* е бројот на точни позитивни примероци, *TN* е бројот на точни негативни примероци, *FP* е бројот на грешни позитивни примероци и *FN* е бројот на грешни негативни примероци. Евалуациските мерки добиени со макро и микро усреднување по ознака за дадената мерка за бинарна евалуација може да се пресметаат како

$$\begin{aligned} \text{B}_{\text{macro}} &= \frac{1}{L}\sum_{i=1}^{L}\text{B}(TP_i, TN_i, FP_i, FN_i) \\ \text{B}_{\text{micro}} &= \text{B}\left(\sum_{i=1}^{L}TP_i, \sum_{i=1}^{L}TN_i, \sum_{i=1}^{L}FP_i, \sum_{i=1}^{L}FN_i\right), \end{aligned} \tag{5.2}$$

каде со L е означен бројот на ознаки (протеински функции во ова истражување), а TP_i, TN_i, FP_i и FN_i се бројот на точни позитивни, точни негативни, грешни позитивни и грешни негативни примероци за i-тата ознака, соодветно. Од равенството (5.2) може да се забележи дека со макро верзиите на мерките, поединечните вредностите за мерките добиени од секоја ознака на крај се усреднуваат со што на некој начин мерката добиена за секоја ознака добива иста тежина. Од

друга страна, кај микро мерките се пресметува просечниот број на точни позитивни, точни негативни, грешни позитивни и грешни негативни примероци од сите ознаки, и потоа се пресметува соодветната микро мерка. Но, на овој начин во микро мерките позастапените ознаки многу повеќе доминираат, што не е случај кај макро верзиите од мерките.

5.2. Метод за одредување на функции на протеински структури врз основа на структурно порамнување

Во поглавје 3 беа презентирани неколку методи за пребарување на протеински структури. Анализите кои беа направени покажаа дека протеинскиот дескриптор базиран на рамномерна интерполација на протеинскиот скелет иако е многу едноставен за имплементација сепак овозможува компактна и ефикасна репрезентација на протеинската структура. Имено, при евалуацијата на методите се покажа дека овој дескриптор е помеѓу подобрите за пребарување на протеински структури. Во [A36] беше воведен нов метод за функционално анотирање на протеинските структури кој се базира на структурно порамнување на протеински структури. Во продолжение ќе биде опишан овој метод, и ќе биде направена негова евалуација.

5.2.1. Одредување на протеинските функции со безтежинско гласање

Опис на методот

Анотирањето на протеинските структури со овој метод се одвива во две фази. Во првата фаза се врши екстракција на протеинскиот дескриптор базиран на рамномерна интерполација на скелетот, а потоа испитуваниот протеин се споредува со сите протеини за тренирање. Во втората фаза, се идентификуваат k-те најблиски соседи на испитуваниот протеин за тестирање, и потоа се врши одредување на функциите на испитуваниот протеин врз основа на анотациите на неговите најблиски соседи. Методот се базира на k-nn (k nearest neighbours) методот за класификација (155), но истиот се адаптира за решавање на повеќезначен проблем. Нека за дадена протеинска структура за тестирање q се идентифицирале нејзините k најблиски соседи помеѓу протеинските структури за обука врз основа на растојанијата помеѓу нивните дескриптори. Потоа се идентификуваат k множества од функции S_1, S_2,…, S_k, каде S_i е множеството на функции кои ги има i-тиот најблизок сосед. Со предложениот методот се предвидува множеството на функции на испитуваната протеинска структура S_q, така што од множеството со функции $F=\{f_1, f_2, \ldots f_L\}$ предвид се земаат функциите за кои е исполнет условот

$$v_{min} \le v(f_j) = \sum_{i=1}^{k} \text{припаѓа}(f_j, S_i) \,, \qquad \text{припаѓа}(f_j, S_i) = \begin{cases} 1, & f_j \in S_i \\ 0, & f_j \notin S_i \end{cases} \,, \tag{5.3}$$

каде $f_j \in F$ е испитуваната функција за која се проверува условот, а v_{min} е предефиниран праг за гласот на функцијата $v(f_j)$ кој треба да биде задоволен за да функцијата f_j биде земена во множеството со предвидени функции S_q. Прагот v_{min} одговара на минималниот број на протеински структури помеѓу најблиските соседи кои треба да ја имаат функцијата f_j за да таа биде доделена како ознака на испитуваниот протеин. Во оваа анализа се користи безтежинско гласање, односно гласот на секој сосед е со иста тежина независно од неговото растојание до испитуваниот примерок.

Опис на податочните множества

Во оваа анализа се користат анотациите од Gene Ontology (GO) (3) кои беа достапни на 24 мај 2013 година. При тоа се формираат три множества. Множеството за обука *множество1* е формирано од анотираните протеински ланци кои се членови во таргет множеството кое се користи кај PRISM методот (131). Авторите на PRISM ова множество го формираат со филтрирање на ланците кои имаат помалку од 50% секвентна сличност користејќи го методот достапен на (2). Потоа множеството за тестирање се формира со филтрирање на анотираните ланци кои не се во множеството за обука, а кои според ASTRAL (145) имаат помалку од 10% секвентна сличност.

Во рамки на оваа анализа анотациите кои се предвидени со предложениот метод ќе бидат споредени со анотациите со код за евиденција = "IEA" (IEA), како и со множеството на анотации со код за евиденција ≠ "IEA" (nonIEA). За таа цел се формира множеството *множество2* на следниот начин. Од *множество1* се филтрираат протеинските ланци кои имаат и IEA и nonIEA анотации, и кои се анотирани со функциите кои се присутни во двете подмножества од анотации (IEA и nonIEA). Овие ланци го формираат множеството за тестирање. Множеството за обука се формира од ланците кои се во *множество1* а не се во *множество2*, и кои се анотирани со функциите кои ги има во двете подмножества од анотации.

За секоја анотација во GO има информација за изворот од кој доаѓа анотацијата. Беше направена анализа на дистрибуцијата на изворите на анотациите со што се утврди дека околу 59.89% од анотациите доаѓаат од InterPro (208), околу 35.68% од анотациите доаѓаат од UniProtKB (209), а останатите анотации доаѓаат од други извори. InterPro (208) и UniProtKB (209) претставуваат бази кои содржат податоци за анотации кои се откриени по експериментален пат или на автоматизиран начин. Во оваа секција анотациите кои се добиени со предложениот метод ќе бидат споредени со анотациите кои се снимени во овие бази со анотации, со цел да се процени предиктивната моќ на предложениот метод. Идејата која лежи зад оваа споредба, е предложениот метод да се спореди со методите и алатките кои се користат за прибирање на знаењето кое се складира во овие бази со анотации. За таа цел се формира множеството

множество3, и тоа е добиено на сличен начин како *множество2* така што предвид се земаат анотациите кои доаѓаат од изворите InterPro и UniProtKB.

Во Табела 5.1 прикажани се карактеристиките на податочните множества кои се користат во оваа анализа. При тоа прикажани се и кардиналноста на лабелите (Label cardinality) C и густината на лабелите (Label density) ρ, кои се едни од позначајните карактеристики за опишување на податочни множества при решавање на повеќезначни проблеми бидејќи тие даваат информација за комплексноста на проблемот кој се решава. Овие мерки се дефинирани во (207) и се пресметуваат со равенствата (5.4). Кардиналноста на лабелите е еднаква на просечниот број на ознаки по примерок. Сепак, кај проблеми со поголем број на различни ознаки комплексноста на проблемот е значително поголема во однос на проблеми со мал број на ознаки, па затоа порелевантна карактеристика за податочните множества е густината на лабелите. Кај оваа карактеристика се одредува колкава фракција од функциите ги има даден примерок, и потоа се усреднува по сите примероци за тестирање.

$$C = \frac{1}{Q}\sum_{i=1}^{Q}\left|A_i\right| \qquad \rho = \frac{1}{Q}\sum_{i=1}^{Q}\frac{\left|A_i\right|}{L} \tag{5.4}$$

множество	#примероци за обука	#примероци за тестирање Q	Број на лабели L	Кардиналност C	Густина ρ (%)
множество1	7806	827	2268	5.42	0.24
множество2	6643	442	218	2.70	1.24
множество3	3803	4526	451	3.93	0.87

Табела 5.1 Карактеристики на податочните множества.

Резултати од евалуацијата на методот со користење на безтежинско гласање

Прво беше направена анализа за влијанието на метриката за растојание и должината на дескрипторот (бројот на карактеристики) D врз предиктивната моќ на моделите При тоа се користи *множество1*, а v_{min} е поставено на 2. Беа направени експерименти со користење на различен број на соседи (k=3 и k=5) во комбинација со L_1 и L_2 метриките за растојание. Експерименталните резултати од оваа анализа се прикажани во Табела 5.2 и Табела 5.3. Резултатите покажуваат дека не секогаш е подобро да се користи подолг дескриптор. Генерално, со намалување на бројот на карактеристики од 512 до 64, предиктивната моќ малку се зголемува. Ова е резултат на тоа што со користење на повеќе карактеристики предвид се земаат повеќе информации кои може да не бидат релевантни. Потоа, со намалувањето на дескрипторот од 64 до 8 карактеристики, предиктивната моќ значително се намалува. Генерално може да се заклучи

дека најдобрите поставки се D=64 со L_2 норма и D=128 со L_1 норма. Па затоа во следните анализи се користат овие поставки за должината на дескрипторот и мерката за растојание.

Метрика за растојание	**L_2 норма**							**L_1 норма**						
Должина на дескриптор	**512**	**256**	**128**	**64**	**32**	**16**	**8**	**512**	**256**	**128**	**64**	**32**	**16**	**8**
Precision	0.72	0.72	0.72	0.73	0.72	0.68	0.57	0.73	0.73	0.74	0.73	0.72	0.68	0.57
Recall	0.72	0.72	0.72	0.73	0.72	0.68	0.55	0.72	0.72	0.72	0.72	0.71	0.67	0.55
F_1	0.71	0.71	0.71	0.72	0.71	0.66	0.55	0.71	0.71	0.72	0.71	0.70	0.66	0.55
Accuracy	0.69	0.69	0.69	0.69	0.69	0.65	0.53	0.69	0.69	0.69	0.69	0.68	0.64	0.53
$Precision_{macro}$	0.24	0.24	0.24	0.23	0.23	0.22	0.19	0.23	0.23	0.24	0.23	0.23	0.22	0.19
$Recall_{macro}$	0.22	0.22	0.22	0.22	0.22	0.21	0.17	0.22	0.22	0.22	0.22	0.22	0.21	0.17
$F_{1\ macro}$	0.22	0.22	0.22	0.22	0.22	0.21	0.17	0.22	0.22	0.22	0.22	0.22	0.21	0.17
$Precision_{macro}$ *	0.75	0.75	0.75	0.75	0.73	0.71	0.60	0.75	0.75	0.75	0.75	0.73	0.71	0.60
$Recall_{macro}$ *	0.71	0.71	0.71	0.71	0.70	0.66	0.55	0.72	0.72	0.72	0.71	0.70	0.66	0.54
$F_{1\ macro}$ *	0.72	0.72	0.71	0.71	0.70	0.67	0.56	0.72	0.72	0.72	0.71	0.70	0.66	0.55
$Precision_{micro}$	0.82	0.81	0.81	0.81	0.80	0.75	0.67	0.81	0.81	0.82	0.81	0.79	0.76	0.67
$Recall_{micro}$	0.74	0.74	0.73	0.74	0.73	0.69	0.57	0.74	0.74	0.74	0.74	0.73	0.69	0.57
$F_{1\ micro}$	0.77	0.77	0.77	0.77	0.76	0.72	0.62	0.77	0.78	0.78	0.77	0.76	0.72	0.62

Табела 5.2 Резултати добиени со предложениот метод со користење на L_1 и L2 норма, k=3 најблиски соседи и v_{min} = 2.

Метрика за растојание	**L_2 норма**							**L_1 норма**						
Должина на дескриптор	**512**	**256**	**128**	**64**	**32**	**16**	**8**	**512**	**256**	**128**	**64**	**32**	**16**	**8**
Precision	0.58	0.58	0.58	0.59	0.57	0.54	0.46	0.59	0.58	0.59	0.59	0.57	0.54	0.48
Recall	0.74	0.74	0.74	0.75	0.74	0.70	0.60	0.74	0.74	0.75	0.75	0.74	0.70	0.60
F_1	0.62	0.62	0.62	0.63	0.61	0.58	0.50	0.62	0.62	0.63	0.63	0.61	0.58	0.51
Accuracy	0.56	0.55	0.56	0.57	0.55	0.52	0.44	0.56	0.56	0.56	0.57	0.55	0.52	0.45
$Precision_{macro}$	0.21	0.21	0.21	0.21	0.20	0.19	0.17	0.21	0.21	0.21	0.21	0.20	0.20	0.17
$Recall_{macro}$	0.23	0.23	0.23	0.23	0.22	0.21	0.19	0.23	0.23	0.23	0.23	0.22	0.21	0.18
$F_{1\ macro}$	0.21	0.21	0.21	0.21	0.20	0.20	0.17	0.21	0.21	0.21	0.21	0.20	0.20	0.17
$Precision_{macro}$ *	0.66	0.66	0.66	0.66	0.64	0.62	0.54	0.66	0.66	0.67	0.67	0.65	0.63	0.55
$Recall_{macro}$ *	0.73	0.73	0.73	0.73	0.72	0.68	0.59	0.73	0.73	0.73	0.73	0.72	0.68	0.59
$F_{1\ macro}$ *	0.67	0.67	0.67	0.67	0.65	0.63	0.54	0.67	0.67	0.67	0.67	0.65	0.63	0.55
$Precision_{micro}$	0.57	0.56	0.56	0.57	0.55	0.52	0.45	0.57	0.57	0.57	0.57	0.55	0.52	0.48
$Recall_{micro}$	0.76	0.76	0.76	0.77	0.76	0.72	0.62	0.76	0.76	0.77	0.77	0.76	0.71	0.62
$F_{1\ micro}$	0.65	0.65	0.65	0.66	0.64	0.60	0.52	0.65	0.65	0.65	0.65	0.64	0.60	0.54

Табела 5.3 Резултати добиени со предложениот метод со користење на L_1 и L2 норма, k=5 најблиски соседи и v_{min} = 2.

Макро мерките постигнаа ниски вредности бидејќи само 709 од вкупно 2268 функции (ознаки) се присутни во множеството за тестирање. Преостанатите функции доаѓаат од анотациите на примероците за обука, па мерките добиени за овие ознаки се нули бидејќи бројот на точни позитивни примероци за нив е нула. При пресметувањето на макро мерките,

соодветната мерка се усреднува врз сите 2268 ознаки, па од тука максималната можна вредност (во случај на перфектна предикција) на ова податочно множество е 1*709/2268 ≈ 0.312. Па затоа исто така се презентирани и резултатите добиени со усреднување врз сите 709 ознаки кои се присутни во множеството за тестирање. Овие вредности за макро мерките се прикажани со ѕвездичка (*).

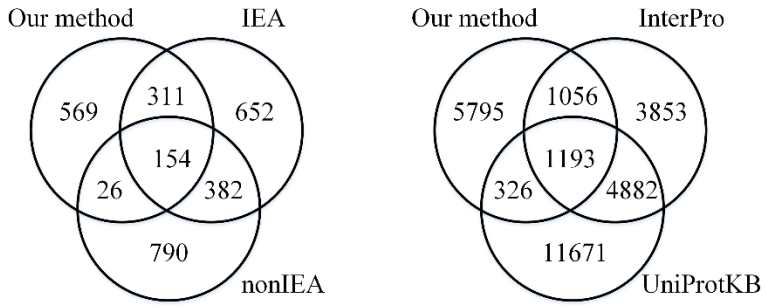
Во Табела 5.4 прикажани се резултатите од анализата за оптималните вредности за k и v_{min}. Со користење на $k = 5$ се добива повисок одѕив отколку за $k = 3$ бидејќи поголем број од релевантните функции ќе бидат опфатени со најблиските соседи. Сепак, со користење на поголемо k расте и бројот на грешни позитивни примероци бидејќи и понерелевантни соседи ќе гласаат при одлучувањето. За $k = 3$ се добива значително повисока прецизност бидејќи само трите најслични структури гласаат за анотациите на испитуваниот примерок. Како и да е, доколку испитуваниот примерок има поголем број на слични структури, со користење на три соседи може да се испуштат некои релевантни функции кои се опфатени од пооддалечените соседи. Со користење на $v_{min} = 1$ се добива повисок одѕив, но микро прецизноста е значително помала. Со користење на поголемо v_{min} се добива помал број на грешни позитивни примероци што резултира со повисока микро прецизност и помал одѕив. Генерално, може да се заклучи дека оптималните поставки се $k = 3$ и $v_{min} = 2$. Ако предложениот метод се користи како филтер за одредување на кандидатските функции, кои потоа се верифицираат со некој друг попрецизен метод, тогаш може да се користи поголемо k и помало v_{min} со цел да се опфатат сите релевантни функции.

	***D* = 64, L_2 норма**						***D* = 128, L_1 норма**					
k	**5**				**3**		**5**				**3**	
v_{min}	**1**	**2**	**3**	**4**	**1**	**2**	**1**	**2**	**3**	**4**	**1**	**2**
Precision	0.44	0.59	0.54	0.43	0.60	0.73	0.44	0.59	0.53	0.44	0.60	0.74
Recall	0.86	0.75	0.49	0.36	0.84	0.73	0.85	0.75	0.49	0.36	0.84	0.72
F_1	0.54	0.63	0.49	0.37	0.66	0.72	0.54	0.63	0.49	0.38	0.67	0.72
Accuracy	0.43	0.57	0.46	0.35	0.59	0.69	0.44	0.56	0.46	0.35	0.60	0.69
$\text{Precision}_{\text{macro}}$	0.19	0.21	0.16	0.14	0.22	0.23	0.18	0.21	0.16	0.14	0.22	0.24
$\text{Recall}_{\text{macro}}$	0.27	0.23	0.14	0.10	0.27	0.22	0.27	0.23	0.14	0.10	0.27	0.22
$F_{1\ \text{macro}}$	0.21	0.21	0.14	0.11	0.23	0.22	0.21	0.21	0.14	0.11	0.23	0.22
$\text{Precision}_{\text{macro}}$ *	0.59	0.66	0.52	0.44	0.69	0.75	0.59	0.67	0.52	0.45	0.69	0.75
$\text{Recall}_{\text{macro}}$ *	0.87	0.73	0.43	0.31	0.86	0.71	0.86	0.73	0.44	0.32	0.86	0.72
$F_{1\ \text{macro}}$ *	0.66	0.67	0.45	0.34	0.74	0.71	0.66	0.67	0.46	0.35	0.74	0.72
$\text{Precision}_{\text{micro}}$	0.37	0.57	0.74	0.93	0.53	0.81	0.38	0.57	0.73	0.93	0.53	0.82
$\text{Recall}_{\text{micro}}$	0.88	0.77	0.49	0.35	0.86	0.74	0.87	0.77	0.49	0.36	0.86	0.74
$F_{1\ \text{micro}}$	0.52	0.66	0.59	0.51	0.65	0.77	0.53	0.65	0.59	0.52	0.66	0.78

Табела 5.4 Резултати добиени со предложениот метод со користење на различен број на најблиски соседи k и различна вредност за прагот v_{min}.

Споредба на предвидените анотации со IEA, nonIEA, InterPro и UniProtKB анотациите

Следно, анотациите кои се предвидени со предложениот метод ги беа споредени со IEA и nonIEA анотациите. Слично, предвидените анотации ги беа споредени со анотациите сместени во InterPro (208) и UniProtKB (209). Во оваа анализа се користат следните поставки $D = 128$, L_1 норма, $k = 3$ и $v_{min} = 2$. Резултатите од оваа анализа, кои се прикажани на Слика 5.1, покажуваат дека предложениот метод точно предвидува 154 анотации кои се присутни помеѓу IEA и nonIEA анотациите, 311 анотации кои ги има само помеѓу IEA анотациите и 26 анотации кои ги има само помеѓу nonIEA анотациите. Преостанатите 569 анотации кои се предвидени од методот не се помеѓу познатите анотации во GO. Во предвидувањето на IEA анотациите предложениот метод постигна 0.44 микро прецизност и 0.31 микро одзив, додека во предвидувањето на nonIEA анотациите методот постигна 0.17 микро прецизност и 0.13 микро одзив. Предложениот метод ја предвиде унијата од IEA и nonIEA анотациите со микро прецизност од 0.46 и микро одзив од 0.21. Микро F_1 мерката добиена од споредбата на IEA и nonIEA анотациите е 0.38, додека микро F_1 мерката добиена од споредбата на предвидените и IEA анотациите е 0.36. Треба да се спомене дека во оваа анализа во множеството за тестирање се земаат примероци за кои има анотации во двете множества (IEA и nonIEA), но со тоа може да се добијат примероци за тестирање кои се значително различни од примероците за обука. Од споредбата со InterPro и UniProtKB анотациите, може да се утврди дека предложениот метод е посличен со методите кои се користат за добивање на знаењето кое е сместено во InterPro базата со анотации (микро F_1 мерката е 0.23), наспрема методите кои се користат кај UniProtKB (микро F_1 мерката е 0.11).



а) б)

Слика 5.1 Резултати од споредбата на предвидените анотации со а) IEA и nonIEA анотациите; б) InterPro и UniProtKB анотациите.

Претходната анализа покажува дека анотациите кои се предвидени од предложениот метод се послични со IEA и InterPro анотациите, што значи дека овој метод е посличен со алатките кои се користат за откривање на овие анотации. Сепак треба да се спомене дека евалуациските мерки кои се добиени во анализата не покажуваат дали некој метод е подобар во однос на другиот, туку

покажуваат дали двете множества од анотации добиени со методите кои се споредуваат опфаќаат слично знаење или не. Доколку две множества опфаќаат различно знаење, тоа не значи дека методите не се прецизни, туку едноставно истите се фокусирани на откривање на различни делови од сите можни интеракции кои постојат помеѓу протеинските структури.

5.2.2. Одредување на протеинските функции со тежинско гласање

Опис на методот

Како што беше споменато погоре, во анализите кои беа презентирани погоре се користи безтежинско гласање при што секој од најблиските соседи гласа со иста тежина при одлучувањето. Дополнително беше направено истражување каде на различен сосед му се доделува различна тежина врз основа на неговото растојание до испитуваниот протеински ланец. За таа цел равенството (5.3) се проширува во

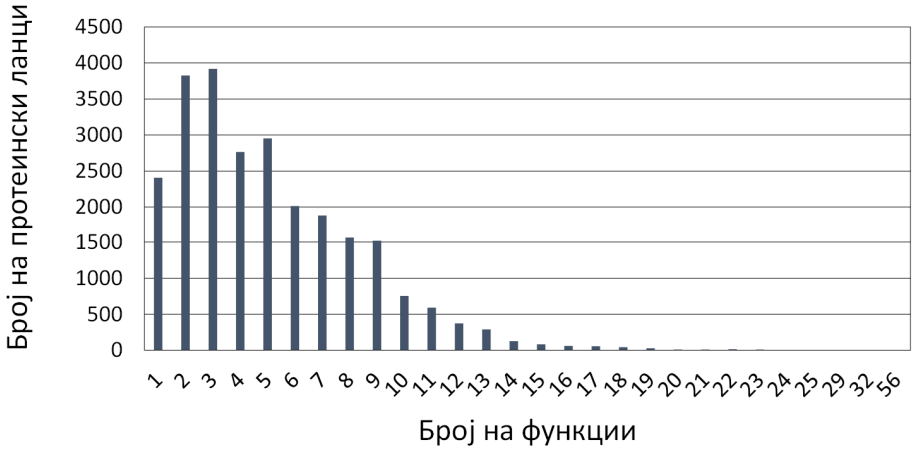
$$v_{min} \le v(f_j) = \sum_{i=1}^{k} \text{припаѓа}(f_j, S_i) \frac{1}{d(i,q)^{n_1} * i^{n_2}}, \qquad \text{припаѓа}(f_j, S_i) = \begin{cases} 1, & f_j \in S_i \\ 0, & f_j \notin S_i \end{cases}, \qquad (5.5)$$

каде $d(i,q)$ е растојанието помеѓу i-тиот најблизок сосед и испитуваниот примерок q, n_1 е степенот на тоа растојание, додека преку параметарот n_2 се дефинира дали при гласањето предвид ќе се земе и редната позиција на најблискиот сосед (за $n_2>0$). Преку параметарот n_1 може да се дефинира дали ќе се користи тежинско ($n_1 > 0$) или безтежинско гласање ($n_1 = 0$).

Опис на податочното множество

Во продолжение ќе биде направена детална анализа на перформансите на овој метод. За споредба на две протеински структури ќе се користат дескрипторот базиран на рамномерна интерполација на протеинскиот скелет (секција 3.2), дескрипторите базирани на бранчиња (секција 3.3) и MASASW методот (секција 3.4) за порамнување на матриците на растојанија. Оваа анализа е направена со користење на податочно множество кое е добиено така што од GO анотациите достапни на 12 јули 2013 година предвид се земени анотациите на репрезентативните протеински ланци. При тоа за репрезентативни ланци се сметаат ланците за кои важи дека имаат помалку од 100% секвентна сличност според BLASTClust методот со кој се прави кластерирање на секвенци врз основа на нивното растојание добиено со BLUST методот (9). Од ова репрезентативно множество се формира множеството за тестирање кое ги содржи протеинските ланци кои имаат помалку од 30% секвентна сличност според BLASTClust, а останатите репрезентативни ланци го формираат множеството за обука. Од анотациите на ланците за тестирање потоа се филтрираат само оние функции кои ги има барем еден од протеините за обука, бидејќи во спро-

тивно истата функција секогаш ќе биде грешно негативно предвидена. Со ова се добива множество за обука од 16363 протеински ланци и множество за тестирање од 9029 протеински ланци. Бројот на различни функции кои ги имаат овие репрезентативни протеини е 3379, а вкупниот број на анотации на ланците кои се земаат предвид е 129226. Ова податочно множество нека го именуваме како *GO100_30*. За да се добие подобар поглед на дистрибуцијата на функциите по ланци, на Слика 5.2 е прикажана дистрибуција на функциите по протеински ланци, а на Слика 5.3 е прикажана дистрибуцијата на протеински ланци по функција. Од Слика 5.2 може да се забележи дека повеќето ланци имаат од 1 до 10 функции, а од Слика 5.3 може да се забележи дека повеќето функции се специфични функции кои ги имаат мал дел од протеинските ланци, наспрема ретките генерални функции кои се застапени кај голем број на протеинските ланци. Од ова се забележува комплексноста на проблемот кој го решаваме.



Слика 5.2 Дистрибуција на функциите по протеински ланци.

Слика 5.3 Дистрибуција на протеинските ланци по функции.

Резултати од евалуацијата на методот со користење на дескрипторот базиран на рамномерна интерполација на протеинскиот скелет

Прво беше направена анализа на предиктивната моќ на методот доколку споредбата на структурите се базира на споредба на дескрипторите базирани на рамномерна интерполација на скелетот. Беа направени анализи со користење на дескриптори со различна должина D во комбинација со L_1 и L_2 нормите како метрики за растојание. Во овие експерименти предвид се земаат k=3 најблиски соседи, прагот v_{min} е поставен на 2 и се користи безтежинско гласање (n_1=n_2=0). Од резултатите од оваа анализа дадени во Табела 5.5 може да се забележи дека со намалување на должината на дескрипторот D од 512 до 64 има незначителна промена во вредноста на микро F_1 мерката, а потоа со намалување на должината од 64 до 8 има значителен пад на предиктивната моќ. За поголема вредност на D посоодветна е L_1 нормата, а за помало D подобро е да се користи L_2 нормата. Ова е и очекувано однесување бидејќи при споредба со поголема резолуција (за поголемо D) големите разлики во растојанијата не треба дополнително да се земаат на степен 2, додека при анализа со помала резолуција (за помало D) поголемите растојание треба многу повеќе да доминираат при пресметување на вкупното растојание меѓу дескрипторите. Највисока вредност за микро F_1 мерката од 0.39 се доби за D=512 и D=128 во комбинација со L_1 нормата. Следно, беше направена анализа користејќи го истиот дескриптор, при тоа беа направени анализи за различни вредности на k и v_{min} користејќи безтежинско гласање (n_1=n_2=0). Во оваа анализа се користат најдобрите поставки за D и метрика за растојание (D=128 и D=512, L_1 норма). Резултатите од оваа анализа се прикажани во Табела 5.6.

Метрика за растојание	**L_2 норма**							**L_1 норма**						
Должина на дескриптор	**512**	**256**	**128**	**64**	**32**	**16**	**8**	**512**	**256**	**128**	**64**	**32**	**16**	**8**
Precision	0.31	0.31	0.31	0.31	0.30	0.24	0.15	0.32	0.32	0.32	0.31	0.30	0.24	0.15
Recall	0.29	0.29	0.29	0.29	0.28	0.21	0.13	0.30	0.30	0.30	0.29	0.28	0.21	0.13
F_1	0.28	0.28	0.28	0.28	0.27	0.21	0.13	0.29	0.29	0.29	0.28	0.27	0.21	0.13
Accuracy	0.25	0.25	0.25	0.25	0.24	0.19	0.11	0.26	0.26	0.26	0.25	0.24	0.19	0.11
$\text{Precision}_{\text{macro}}$	0.28	0.27	0.28	0.27	0.25	0.21	0.12	0.28	0.28	0.28	0.28	0.26	0.21	0.13
$\text{Recall}_{\text{macro}}$	0.24	0.24	0.24	0.24	0.22	0.18	0.12	0.24	0.24	0.24	0.24	0.22	0.18	0.12
$F_{1\ \text{macro}}$	0.22	0.22	0.23	0.22	0.21	0.17	0.10	0.23	0.23	0.23	0.23	0.21	0.17	0.10
$\text{Precision}_{\text{macro}}$ *	0.37	0.37	0.38	0.37	0.34	0.28	0.17	0.38	0.38	0.38	0.37	0.35	0.28	0.17
$\text{Recall}_{\text{macro}}$ *	0.32	0.32	0.32	0.32	0.30	0.25	0.16	0.33	0.33	0.33	0.32	0.30	0.25	0.16
$F_{1\ \text{macro}}$ *	0.30	0.30	0.31	0.30	0.28	0.22	0.14	0.31	0.31	0.31	0.30	0.28	0.22	0.14
$\text{Precision}_{\text{micro}}$	0.46	0.46	0.46	0.46	0.43	0.35	0.25	0.46	0.46	0.46	0.46	0.43	0.35	0.25
$\text{Recall}_{\text{micro}}$	0.32	0.32	0.32	0.32	0.30	0.24	0.16	0.33	0.33	0.33	0.32	0.31	0.24	0.15
$F_{1\ \text{micro}}$	0.38	0.38	0.38	0.38	0.36	0.29	0.19	**0.39**	0.38	**0.39**	0.38	0.36	0.29	0.19

Табела 5.5 Резултати добиени со предложениот метод со користење на протеинскиот дескриптор базиран на рамномерна интерполација на скелетот користејќи дескриптори со различна должина D, k=3 најблиски соседи и v_{min} = 2.

	***D*=128, L_1 норма**			***D*=512, L_1 норма**		
k	**1**	**3**	**5**	**1**	**3**	**5**
v_{min}	**1**	**2**	**2**	**1**	**2**	**2**
Precision	0.398	0.317	0.290	0.399	0.316	0.291
Recall	0.417	0.297	0.336	0.418	0.297	0.336
F_1	0.391	0.289	0.287	0.391	0.287	0.288
Accuracy	0.358	0.259	0.245	0.358	0.258	0.245
$Precision_{macro}$	0.342	0.282	0.222	0.347	0.282	0.224
$Recall_{macro}$	0.370	0.243	0.260	0.373	0.245	0.259
$F_{1\ macro}$	0.322	0.231	0.206	0.325	0.231	0.207
$Precision_{macro}$ *	0.458	0.378	0.298	0.465	0.378	0.300
$Recall_{macro}$ *	0.497	0.326	0.348	0.500	0.328	0.348
$F_{1\ macro}$ *	0.431	0.310	0.276	0.436	0.310	0.278
$Precision_{micro}$	0.410	0.463	0.322	0.412	0.462	0.323
$Recall_{micro}$	0.458	0.330	0.370	0.460	0.331	0.370
$F_{1\ micro}$	0.433	0.385	0.345	0.435	0.385	0.345

Табела 5.6 Резултати добиени со предложениот метод со користење на дескрипторот базиран на рамномерна интерполација на скелетот користејќи различни вредности за *k* и v_{min}.

Од овие резултати може да се утврди дека генерално за *D*=512 во комбинација со L_1 норма се добиваат подобри резултати. Интересно е да се напомене дека за *k*=1, v_{min}=1 се доби повисок одзив отколку за *k*=3, v_{min}=2, што не е случај и со прецизноста. Ова значи дека доколку гласа само првиот сосед се опфаќаат поголем дел од релевантните функции, но за одредени протеини кои немаат доволно висока сличност со ниту еден протеин за обука, најблискиот сосед ќе биде носител на нерелевантни функции што придонесува до намалување на прецизноста. Ова одне-сување е затоа што множеството содржи репрезентативни протеински ланци, па затоа најчесто протеините немаат два соседи со кои делат исти функции, но исто така има и ланци за кои во репрезентативното множество нема ланец кој е доволно сличен со нив и при тоа да делат исти функции. Доколку се споредат резултатите за *k*=3, v_{min}=2 и *k*=5, v_{min}=2 може да се забележи дека во првиот случај се добива повисока прецизност бидејќи гласаат само поблиски соседи, но затоа пак се добива помал одзив, т.е. дел од функциите кои испитуваниот протеин ги дели со неговиот четврти и петти сосед може да не бидат предвидени.

Како што беше опишано погоре, наместо безтежинско гласање може да се користи и тежин-ско гласање. Во Табела 5.7 се прикажани резултати добиени користејќи различни вредности за параметрите n_1 и n_2 преку кои се дефинира тежината на гласот на даден сосед како функција од растојанието до испитуваниот протеин, како и редната позиција на тој сосед. Оваа анализа е направена користејќи *D*=512 и L_1 норма. Кај тежинското гласање беа направени повеќе експери-менти користејќи различен праг v_{min}, а во табелата се дадени резултатите за најдобрата вредност на прагот. Од резултатите може да се забележи дека со користење на тежинско гласање се доби-ваат подобри резултати отколку со безтежинско гласање. При тоа како најдобро се покажа

гласањето за n_1=n_2=1 во кое предвид се зема и растојанието на соседот до испитуваниот протеин, како и неговата редна позиција помеѓу соседите.

k	1	3	3	3	3	5	5	5	5
v_{min}	1	2	0.035	0.02	0.0005	2	0.035	0.02	0.0009
n_1	0	0	1	1	2	0	1	1	2
n_2	0	0	0	1	0	0	0	1	0
Precision	0.40	0.32	0.32	0.37	0.33	0.29	0.32	0.37	0.30
Recall	0.42	0.30	0.32	0.38	0.37	0.34	0.35	0.38	0.31
F_1	0.39	0.29	0.30	0.35	0.33	0.29	0.31	0.36	0.29
Accuracy	0.36	0.26	0.28	0.33	0.30	0.25	0.27	0.33	0.26
$\text{Precision}_{\text{macro}}$	0.35	0.28	0.41	0.41	0.40	0.22	0.36	0.40	0.42
$\text{Recall}_{\text{macro}}$	0.37	0.24	0.30	0.34	0.34	0.26	0.31	0.35	0.30
$F_{1\ \text{macro}}$	0.33	0.23	0.31	0.34	0.34	0.21	0.29	0.34	0.32
$\text{Precision}_{\text{macro}}$ *	0.46	0.38	0.55	0.56	0.54	0.30	0.48	0.54	0.56
$\text{Recall}_{\text{macro}}$ *	0.50	0.33	0.40	0.46	0.46	0.35	0.41	0.47	0.40
$F_{1\ \text{macro}}$ *	0.44	0.31	0.42	0.46	0.45	0.28	0.39	0.46	0.42
$\text{Precision}_{\text{micro}}$	0.41	0.46	0.61	0.54	0.51	0.32	0.45	0.51	0.57
$\text{Recall}_{\text{micro}}$	0.46	0.33	0.36	0.42	0.41	0.37	0.38	0.43	0.35
$F_{1\ \text{micro}}$	0.43	0.39	0.45	**0.47**	0.45	0.34	0.41	**0.46**	0.44

Табела 5.7 Резултати добиени со предложениот метод со користење на дескрипторот базиран на рамномерна интерполација на скелетот (за D=512, L_1 норма) користејќи различни вредности за k, n_1 и n_2.

Дополнително беа направени анализи каде се користи хибрид од тежинско и безтежинско гласање. За таа цел равенството (5.5) е проширено во

$$v_{min1} \le v(f_j) = \sum_{i=1}^{k} \text{припаѓа}(f_j, S_i)\ , \qquad \text{припаѓа}(f_j, S_i) = \begin{cases} 1, & f_j \in S_i \\ 0, & f_j \notin S_i \end{cases},$$
$$v_{min2} \le v(f_j) = \sum_{i=1}^{k} \text{припаѓа}(f_j, S_i) \frac{1}{d(i,q)^{n_1} * i^{n_2}}, \tag{5.6}$$

каде v_{min1} и v_{min2} се праговите кои се користат при безтежинско и тежинско гласање, па така функцијата f_j се зема како член на предвидените функции за испитуваниот протеин ако е исполнет било кој од двата услови. Резултатите од оваа анализа се дадени во Табела 5.8. Од резултатите може да се забележи дека со користење на хибридното гласање (Табела 5.8) кое ги комбинира безтежинското и тежинското гласање се добиваат полоши резултати отколку со користење на тежинското гласање (Табела 5.7). Истото однесување се покажа доколку се користат и дескрипторите со бранчиња, како и MASASW методот, па затоа понатаму нема да бидат презентирани резултатите со хибридното гласање.

k	3	3	3	5	5	5
v_{min1}	2	2	2	2	2	2
v_{min2}	0.03	0.02	0.0005	0.03	0.02	0.0005
n_1	1	1	2	1	1	2
n_2	0	1	0	0	1	0
Precision	0.34	0.36	0.34	0.30	0.32	0.30
Recall	0.35	0.39	0.38	0.39	0.42	0.41
F_1	0.33	0.35	0.34	0.32	0.33	0.32
Accuracy	0.30	0.32	0.30	0.27	0.28	0.27
$Precision_{macro}$	0.36	0.38	0.37	0.30	0.32	0.31
$Recall_{macro}$	0.32	0.35	0.35	0.33	0.36	0.35
$F_{1\ macro}$	0.30	0.33	0.32	0.27	0.30	0.29
$Precision_{macro}$ *	0.48	0.50	0.49	0.40	0.43	0.42
$Recall_{macro}$ *	0.43	0.47	0.46	0.44	0.48	0.48
$F_{1\ macro}$ *	0.41	0.44	0.43	0.37	0.40	0.39
$Precision_{micro}$	0.47	0.47	0.45	0.34	0.34	0.33
$Recall_{micro}$	0.39	0.43	0.42	0.43	0.46	0.45
$F_{1\ micro}$	0.43	0.45	0.43	0.38	0.39	0.38

Табела 5.8 Резултати добиени со предложениот метод со користење на дескрипторот базиран на рамномерна интерполација на скелетот (D=512, L_1 норма) користејќи хибрид од безтежинско и тежинско гласање. Направени се анализи за влијанието на k, v_{min2}, n_1 и n_2.

Резултати од евалуацијата на методот со користење на дескрипторите базирани на бранчиња

Следно, беа направени анализи со користење на дескрипторите со бранчиња за одредување на најблиските соседи. Бидејќи споредбата на моделите во оваа анализа се базира на микро F_1 мерката, затоа при презентирање на резултатите во оваа секција ќе биде прикажана само вредноста на микро F_1 мерката. Прво беа направени експерименти во кои се користат Haar дескриптори со различна должина D, користејќи безтежинско гласање со различни вредности за k и v_{min}. Од резултатите дадени во Табела 5.9 може да се увиди дека со зголемување на бројот на апроксимативни коефициенти од 20 до 100 значително расте $F_{1\ micro}$, а потоа со зголемување на должината на дескрипторот има незначителни промени во предиктивната моќ. Слично беа направени анализи каде беа применети и останатите бранчиња кои се користеа во поглавје 3. Во оваа анализа се користат D=100, D=150 и D=200. Од резултатите дадени во Табела 5.10 може да се утврди дека со Haar бранчето се добиваат значително подобри резултати отколку со другите бранчиња кои се користат во ова истражување. Тука треба да се спомене следново. Во секција 3.5 се покажа дека со Haar бранчето се добиваат најлоши резултати при класификацијата на протеинските ланци во SCOP домени, а со Daubechies2 бранчето се добија најдобри резултати. За разлика од тоа, во оваа секција се прави одредување на функциите на протеините врз основа на функциите на најблиските соседи, и при тоа Haar бранчето се покажа како најдобро, а Daubechies2 како најлошо бранче за оваа намена. Ова е затоа што со Haar бранчето многу

поуспешно се наоѓаат првите неколку најблиски протеини со кои испитуваниот протеин дели исти функции. Од друга страна при класификација на протеини во SCOP домени може структурно многу слични ланци да припаѓаат во различен домен, иако истите делат голем број функции. Па затоа при класификација ова резултира со опаѓање на прецизноста. Од овие анализи може да се заклучи дека со Haar бранчето се обезбедува дескриптор со кој најдобро се препознаваат сличните протеините кои делат исти функции.

D	**20**	**64**	**100**	**128**	**150**	**200**	**255**
***k*=1, v_{min}=1**	0.39	0.45	0.47	**0.48**	**0.48**	**0.48**	**0.48**
***k*=3, v_{min}=2**	0.35	0.42	0.43	0.44	0.44	0.44	0.44
***k*=5, v_{min}=2**	0.31	0.37	0.38	0.39	0.39	0.39	0.39

Табела 5.9 $F_{1\ micro}$ со предложениот метод со користење на дескрипторот базиран на Haar бранчето со различна должина *D* и безтежинско гласање.

D	**100**			**150**			**200**		
k	**1**	**3**	**5**	**1**	**3**	**5**	**1**	**3**	**5**
v_{min}	**1**	**2**	**2**	**1**	**2**	**2**	**1**	**2**	**2**
Haar	**0.47**	**0.43**	**0.38**	**0.48**	**0.44**	**0.39**	**0.48**	**0.44**	**0.39**
Daubechies2	0.33	0.28	0.26	0.35	0.30	0.27	0.36	0.31	0.28
Daubechies3	0.42	0.37	0.33	0.42	0.38	0.34	0.43	0.39	0.34
Daubechies4	0.41	0.36	0.33	0.41	0.37	0.33	0.42	0.38	0.34
Symlet4	0.40	0.36	0.32	0.41	0.36	0.33	0.41	0.36	0.32
Coiflet1	0.42	0.37	0.33	0.42	0.38	0.34	0.43	0.39	0.35

Табела 5.10 $F_{1\ micro}$ со предложениот метод со користење на дескриптори базирани на бранчиња со различна должина *D* и безтежинско гласање.

Дополнително беше направена анализа каде се користи Haar дескриптор со должина D=200 при што со тежинско гласање се одредуваат функциите на протеините за тестирање. Резултатите од оваа анализа се дадени во Табела 5.11. Може да се забележи дека исто како и со дескрипторот базиран на рамномерна интерполација на протеинскиот скелет, така и со Haar дескрипторот најдобри резултати се добиваат за k=3 и n_1=n_2=1.

k	**1**	**3**	**3**	**3**	**3**	**5**	**5**	**5**	**5**
v_{min}	**1**	**2**	**50**	**40**	**2000**	**2**	**60**	**40**	**2500**
n_1	**0**	**0**	**1**	**1**	**2**	**0**	**1**	**1**	**2**
n_2	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**1**	**0**
$F_{1\ micro}$	0.48	0.44	0.48	**0.52**	0.50	0.39	0.44	**0.52**	0.48

Табела 5.11 $F_{1\ micro}$ со предложениот метод со користење на дескрипторот базиран на Haar бранчето со должина D=200 користејќи различни вредности за k, v_{min}, n_1 и n_2.

Резултати од евалуацијата на методот со користење на MASASW

Следно, беа направени експерименти за одредување на предиктивната моќ на методот доколку MASASW методот се користи за одредување на сличноста меѓу две протеински структури. При тоа беа направени експерименти каде се работи со матрици со димензии 64x64 и 128x128. Во оваа анализа големините на лизгачките прозорци *w* и *W* се поставени на 5, а прагот е поставен на 10. Од резултатите може да се забележи дека подобри резултати се добиваат доколку се користат матрици со димензија 64x64, наспрема матрици со димензија 128x128. Исто така од резултатите може да се забележи дека и со MASASW методот најдобрите поставки се k=3 и n_1=n_2=1.

Од сите анализи презентирани претходно може да се заклучи дека со предложениот метод најдобри резултати се добиваат со MASASW методот ($F_{1\ micro}$=0.59), а потоа следат дескрипторот базиран на Haar бранчето ($F_{1\ micro}$=0.52), па дескрипторот базиран на рамномерна интерполација на скелетот на протеинот ($F_{1\ micro}$=0.47).

k	**1**	**3**	**3**	**3**	**3**	**5**	**5**	**5**	**5**
n_1	**0**	**0**	**1**	**1**	**2**	**0**	**1**	**1**	**2**
n_2	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**1**	**0**
v_{min}	**1**	**2**	**3000**	**3000**	**15E+6**	**2**	**3500**	**3000**	**20E+6**
64x64	0.56	0.51	0.53	**0.59**	0.55	0.45	0.46	0.57	0.51
v_{min}	**1**	**2**	**25000**	**12000**	**20E+6**	**2**	**25000**	**12000**	**20E+6**
128x128	0.54	0.49	0.52	0.58	0.56	0.43	0.48	0.56	0.50

Табела 5.12 $F_{1\ micro}$ со предложениот метод со користење на MASASW методот користејќи матрици со различни димензии, како и различни вредности за k, v_{min}, n_1 и n_2.

Споредба на предвидените анотации со IEA, nonIEA, InterPro и UniProtKB анотациите

Следно, беше направена споредба на анотациите предвидени од предложениот метод со IEA (код за евиденција="IEA") и nonIEA (код за евиденција≠"IEA") анотациите, како и со анотациите од InterPro (208) и UniProtKB (209) базите. Од множеството *GO_100*, кое се користеше во претходните анализи, се формира множество за споредба со IEA анотациите, при што за ланците за обука предвид се земаат сите анотации, а за ланците за тестирање предвид се земаат само IEA анотациите. Слично беше формирано множество каде и за ланците за обука и за ланците за тестирање предвид се земаат само IEA анотациите кои се однесуваат на функциите кои се присутни и кај двете множества од ланци. Со овие две множества се прави евалуација на nonIEA анотациите за ланците за тестирање, како и на анотациите за ланците за тестирање добиени со предложениот метод користејќи различен метод за одредување на сличноста помеѓу протеините. При тоа за трите методи кои беа користени во претходните анализи во оваа секција, се избираат

најдобрите поставки за кои се доби највисока вредност за $F_{1\ micro}$. Според истата процедура се генерирани и останатите множества потребни за споредување со nonIEA, InterPro и UniProtKB анотациите. Резултатите од оваа анализа се прикажани во Табела 5.13.

обука	**Сите**				**IEA**			
тест	**IEA**				**IEA**			
анотации	**nonIEA**	**рамн. инт.**	**бранчиња**	**MASASW**	**nonIEA**	**рамн. инт.**	**бранчиња**	**MASASW**
$Precision_{micro}$	0.31	0.53	0.60	0.61	0.44	0.55	0.62	0.63
$Recall_{micro}$	0.02	0.43	0.58	0.46	0.02	0.43	0.58	0.46
$F_{1\ micro}$	0.03	0.47	0.59	0.52	0.03	0.48	0.60	0.53
обука	**Сите**				**nonIEA**			
тест	**nonIEA**				**nonIEA**			
анотации	**IEA**	**рамн. инт.**	**бранчиња**	**MASASW**	**IEA**	**рамн. инт.**	**бранчиња**	**MASASW**
$Precision_{micro}$	0.17	0.02	0.02	0.02	0.22	0.35	0.30	0.35
$Recall_{micro}$	0.27	0.28	0.32	0.29	0.24	0.23	0.25	0.24
$F_{1\ micro}$	0.21	0.04	0.04	0.04	0.23	0.28	0.27	0.29
обука	**Сите**				**InterPro**			
тест	**InterPro**				**InterPro**			
анотации	**Uni-ProtKB**	**рамн. инт.**	**бранчиња**	**MASASW**	**Uni-ProtKB**	**рамн. инт.**	**бранчиња**	**MASASW**
$Precision_{micro}$	0.32	0.35	0.42	0.41	0.34	0.58	0.67	0.66
$Recall_{micro}$	0.20	0.44	0.63	0.49	0.21	0.44	0.62	0.48
$F_{1\ micro}$	0.25	0.39	0.50	0.45	0.26	0.50	0.64	0.56
обука	**Сите**				**UniProtKB**			
тест	**UniProtKB**				**UniProtKB**			
анотации	**InterPro**	**рамн. инт.**	**бранчиња**	**MASASW**	**InterPro**	**рамн. инт.**	**бранчиња**	**MASASW**
$Precision_{micro}$	0.41	0.26	0.27	0.29	0.50	0.48	0.51	0.54
$Recall_{micro}$	0.27	0.44	0.55	0.46	0.27	0.41	0.50	0.43
$F_{1\ micro}$	0.33	0.33	0.36	0.36	0.35	0.45	0.50	0.48

Табела 5.13 Резултати од споредбата на анотациите предвидени од предложениот метод со IEA, nonIEA, InterPro и UniProtKB анотациите.

Резултатите покажуваат дека прецизноста добиена од споредбата на nonIEA со IEA анотациите е 0.31, додека прецизноста од споредбата на IEA со nonIEA анотациите е 0.17. Слично однесување се доби и во анализата која беше направана на друго податочно множество погоре во оваа секција [A36], а тоа се должи на фактот што IEA анотациите се значително поброjни и опфаќаат поширок спектар на знаење во однос на nonIEA анотациите. Па затоа во споредбата сите IEA анотации кои не се помеѓу познатите nonIEA анотации се сметаат за грешно позитивни, но тоа е резултат на ограниченото знаење претставено преку nonIEA анотациите, а не како резултат на грешните одлуки кои се донесени за извлекување на IEA анотациите. Во однос на одзивот, nonIEA анотациите опфаќаат мал дел од познатите IEA анотации, па затоа при нивната споредба со IEA анотациите се доби микро одзив од 0.02, додека IEA анотациите

опфаќаат значително поголем дел од nonIEA анотациите што резултира со микро одзив од 0.27. Во предвидувањето на IEA анотациите предложениот метод најдобри резултати постигна со Haar бранчето, при што малку подобри резултати се добија доколку се користат само IEA анотациите наместо сите анотации. Од друга страна, во предвидувањето на nonIEA анотациите, доколку предвид се земат сите анотации се добива многу помала прецизност отколку ако се користат само nonIEA анотациите, што се должи на фактот што со земањето предвид на сите анотации со методот ќе се предвидат релевантни анотации кои сепак не се помеѓу nonIEA анотациите. Значи оваа ниска прецизност не се должи на тоа дека методите не се прецизни, туку на тоа што методите предвидуваат знаење кое не е опфатено во множеството кое се користи како вистина, а е опфатено од останатите GO анотации. Интересно е да се забележи дека иако со Haar бранчето генерално се добиваат подобри резултати, сепак доколку предвид се земат само nonIEA анотациите на ланците за обука и тестирање, тогаш со Haar бранчето се добиваат послаби резултати отколку со другите методи за споредба на протеински структури. Ова покажува дека со Haar бранчето сепак потешко се открива знаењето кое е опфатено од nonIEA анотациите. Од резултатите од експериментите во кои беше направена споредба со InterPro и UniProtKB анотациите може да се утврди дека анотациите предвидени од предложениот метод имаат поголемо поклопување со множеството од InterPro анотациите наспрема множеството од UniProtKB анотациите. Ова укажува дека со предложениот метод се носат послични одлуки како со алатките кои се користат за генерирање на знаењето сместено во InterPro базата, наспрема алатките кои се користат за одредување на UniProtKB анотациите.

5.3. Метод за одредување на функции на протеински структури базиран на локалните и глобалните карактеристики

Покрај методот за одредување на функциите на протеините врз основа на функциите на неговите најблиски соседи [A36], во [A37], [A38] беше воведен нов метод со кој анотирањето на протеините се прави врз основа на локалните својства на сврзните региони, како и на глобалните својства на протеинската структура.

5.3.1. Локални, глобални и излезни карактеристики

Во оваа анализа предвид се земаат податоците за сврзните региони од протеинската структура сместени во BIND базата (48). Целта е да се изгради предиктивен модел врз основа на **локалните** карактеристики (тоа се карактеристиките на амино-киселинските остатоци кои го формираат сврзниот регион) и **глобалните** карактеристики (тоа се елементите кои го формираат дескрипторот базиран на рамномерна интерполација на протеинскиот скелет). Во поглавје 4 се правеше детекција на сврзните делови од протеинската структура, при што за секој амино-киселински

остаток врз основа на неговите карактеристики се одредува дали тој би припаѓал во сврзен регион или не. За разлика од тоа, во оваа секција целта е за даден сврзен регион, кој се состои од повеќе амино-киселински остатоци, да се одредат функциите кои би ги имал соодветниот протеин во интеракциите кои ќе настанат кај тој сврзен регион. Ова значи дека во оваа анализа примероците не се поединечните амино-киселински остатоци, туку цел сврзен регион. Па затоа за сврзниот регион се одредуваат неговите својства така што од карактеристиките totalASA, totalRASA, avgDPX, avgCX и хидрофобичноста од сите амино-киселини кои се дел од сврзниот регион се одредува вкупната, средната, минималната и максималната вредност, како и варијансата на соодветната карактеристика во рамки на испитуваниот сврзен регион. Исто така предвид се зема и бројот на амино-киселински остатоци кои се дел од сврзниот регион. Со ова се добиваат 26 локални карактеристики. Како што беше кажано претходно, покрај локалните карактеристики на сврзниот регион предвид се земени и глобалните карактеристики на протеинската структура кои соодветствуваат на карактеристиките кои се содржат во протеинскиот дескриптор базиран на рамномерна интерполација на протеинскиот скелет. Во ова анализа скелетот се интерполира со 64 интерполациски точки, па глобалните карактеристики се Евклидовите растојанија од интерполациските точки до центарот на маса.

Излезните карактеристики кои треба да ги предвидуваат моделите се однесуваат на тоа кои функции би ги имал соодветниот сврзен регион. Бидејќи во GO анотациите се однесуваат за протеински ланци, а не за сврзни региони (интерфејси), затоа на сврзниот регион кој се зема како примерок во податочното множество му се придружуваат сите функции кои ги има тој ланец. Со ова варијанта се внесува и мал шум, бидејќи сепак дел од функциите се релевантни за некој од останатите сврзни региони на истиот ланец. Друга варијанта е како точни анотации за даден интерфејс од даден ланец предвид да се земат само функциите кои тој ланец ги дели со ланецот со кој се стапува во интеракција. Сепак, и оваа варијанта не нуди комплетно решение бидејќи во дадена интеракција ланците кои стапуваат во интеракција може да вршат различни функции. Ако се работи со првата верзија тогаш ќе има повеќе грешни позитивни предвидувања, а со втората варијанта ќе има повеќе грешни негативни предвидувања. При анотација на протеини многу поважно е да се опфати што поголем дел од релевантните функции, кои потоа може да се верифицираат со некој друг пософистициран метод. Па затоа во оваа докторска дисертација се користи првата варијанта каде на даден интерфејс му се придружуваат сите функции кои ги има соодветниот ланец.

5.3.2. Методи за повеќезначна класификација

Врз основа на влезните и излезните карактеристики се гради предиктивен модел кој треба да ги одредува функциите кои ги има испитуваниот сврзен регион. За оваа намена може да се користи некој од методите за решавање на повеќезначни проблеми. Во (207) е даден преглед на методите кои може да се користат за решавање на проблеми за повеќезначна класификација, и истите се поделени во две групи: *методи за повеќезначна класификација* и *методи за рангирање на ознаки*. Кај првата група се решава класификациски проблем, односно за секоја ознака на излез се дава бинарна вредност која кажува дали соодветната ознака (во овој случај функција) ќе се придружи на испитуваниот примерок (во овој случај сврзен регион). Од друга страна, кај втората група се решава проблем на регресија, односно за секој излезен атрибут се предвидува неговата релевантност изразена преку континуален атрибут. Друга можна категоризација на методите е поделбата на *методи кои го трансформираат проблемот* и *методи кои го адаптираат алгоритамот*. Првата група методи го трансформираат проблемот во повеќе класификациски проблеми со еден класен атрибут. При тоа може да се примени било кој метод за градење на класификациски модели за предвидување на еден класен атрибут. Втората група на методи врши адаптација на постојните методи за решавање на проблеми со еден класен атрибут, со цел да истите може да се применат за решавање на повеќезначни проблеми.

Во продолжение ќе биде даден краток опис на методите кои се користат во ова истражување. Нека податочното множество содржи 4 примероци $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ и $\mathbf{x}_4$ кои имаат четири различни ознаки f_1, f_2, f_3 и f_4. Во Табела 5.14 се прикажани ознаките (функциите) кои ги имаат примероците (сврзните региони) во множеството за кое ќе бидат опишани методите.

примерок	множество од ознаки
$\mathbf{x}_1$	$\{f_1, f_3\}$
$\mathbf{x}_2$	$\{f_3, f_4\}$
$\mathbf{x}_3$	$\{f_2\}$
$\mathbf{x}_4$	$\{f_1, f_2, f_3\}$

Табела 5.14 Пример за повеќезначно множество.

Методи кои го трансформираат проблемот

Од методите кои вршат трансформација на проблемот, во ова истражување се користат методот на бинарна релевантност (Binary relevance - BR) (207), (210), методот множество на ознаки (Label Powerset - LP) (207), (211) и HOMER (Hierarchy Of Multilabel classifiERs) методот (212). Кај методот на бинарна релевантност (207), (210), кој уште е познат и како еден против сите, се градат L бинарни модели, односно се гради посебен модел за секоја ознака од

множеството $F=\{f_1, f_2,\ldots f_L\}$, каде L е бројот на ознаки. При тренирање на i-тиот модел примероците кои ја имаат i-тата ознака f_i се означуваат како позитивни, а сите преостанати примероци се означуваат како негативни. Во Табела 5.15 е прикажан изгледот на податочното множество после трансформацијата. Во процесот на тестирање, даден примерок за тестирање се презентира на сите L бинарни класификатори, и множеството од ознаки за испитуваниот протеин се наоѓа како унија од сите ознаки кои се позитивно предвидени од класификаторите.

примерок	ознака
$\mathbf{x}_1$	f_1
$\mathbf{x}_2$	$\neg f_1$
$\mathbf{x}_3$	$\neg f_1$
$\mathbf{x}_4$	f_1

примерок	ознака
$\mathbf{x}_1$	$\neg f_2$
$\mathbf{x}_2$	$\neg f_2$
$\mathbf{x}_3$	f_2
$\mathbf{x}_4$	f_2

примерок	ознака
$\mathbf{x}_1$	f_3
$\mathbf{x}_2$	f_3
$\mathbf{x}_3$	$\neg f_3$
$\mathbf{x}_4$	f_3

примерок	ознака
$\mathbf{x}_1$	$\neg f_4$
$\mathbf{x}_2$	f_4
$\mathbf{x}_3$	$\neg f_4$
$\mathbf{x}_4$	$\neg f_4$

Табела 5.15 Трансформација на множеството со методот бинарна релевантност.

Кај методот множества на ознаки (207), (211) се прави трансформација така што секое единствено множество од ознаки кое постои во множеството за обука се зема предвид како една од вредностите за единствениот класен атрибут добиен по трансформацијата, види Табела 5.16. Непознатиот примерок се презентира на единствениот модел кој на излез ја предвидува најверојатната класа, односно тоа е множество од ознаки. Доколку класификаторот кој се користи за градење на моделот на излез дава веројатност за припадност во овие класи, тогаш со LP методот може да се направи рангирање на ознаките. За таа цел се користи пристапот презентиран во (211). Во Табела 5.17 е даден пример за рангирање на ознаките f_1, f_2, f_3 и f_4 со користење на методот множества на ознаки.

Примерок	ознака
$\mathbf{x}_1$	$m_{\{1,3\}}$
$\mathbf{x}_2$	$m_{\{3,4\}}$
$\mathbf{x}_3$	$m_{\{2\}}$
$\mathbf{x}_4$	$m_{\{1,2,3\}}$

Табела 5.16 Трансформација на множеството со методот множества на ознаки.

ознака m_i	i	$p(m_i\|\mathbf{x})$	припаѓа(1,i)	припаѓа(2,i)	припаѓа(3,i)	припаѓа(4,i)
$m_{\{1,3\}}$	{1,3}	0.3	1	0	1	0
$m_{\{3,4\}}$	{3,4}	0.2	0	0	1	1
$m_{\{2\}}$	{2}	0.4	0	1	0	0
$m_{\{1,2,3\}}$	{1,2,3}	0.1	1	1	1	0
$p(f_j\|\mathbf{x})$			**0.4**	**0.5**	**0.6**	**0.2**

Табела 5.17 Рангирање на ознаките со користење на методот множества на ознаки.

Нека со трансформација на множеството со методот множества на ознаки се добиени M различни ознаки m_i, i =1, 2,…., M (Табела 5.16). На моделот добиен со одреден метод за класификација му се дава на влез примерокот за тестирање **x**, и нека за ознаките m_i се добиени веројатностите $p(m_i|\mathbf{x})$, кои ја покажуваат веројатноста да примерокот **x** ја има дадена ознаката m_i. Од тука со користење на методот множества од ознаки може да се најдат веројатностите $p(f_j|\mathbf{x})$ испитуваниот примерок да ги има оригиналните ознаки f_j, j =1, 2,…., L како (207), (211)

$$p(f_j \mid \mathbf{x}) = \sum_{i=1}^{M} p(m_i \mid \mathbf{x})\text{припаѓа}(j,i)\ , \qquad \text{припаѓа}(j,i) = \begin{cases} 1, & j \in i \\ 0, & j \notin i \end{cases}. \tag{5.7}$$

На овој начин може да се изврши рангирање на ознаките за испитуваниот примерок. Со овој метод може да се решаваат проблеми на повеќезначно рангирање.

Сепак, со оваа трансформацијата на проблемот може да се добијат голем број на ознаки кои ги имаат мал број на примероци, па затоа при градењето на моделот потешко е да се извлече релевантно знаење врз основа на тие примероци. Од друга страна, постои зависност помеѓу ознаките, т.е. можат да се најдат групации од ознаки кои често се појавуваат заедно. Па затоа воведен е Hierarchy Of Multilabel classifiERs (HOMER) методот (212) кој повеќезначниот проблем го трансформира во помали повеќезначни проблеми кои се организирани во форма на хиерархија. При тоа се прави хиерархија од повеќезначни подмодели кои се обучени да донесуваат одлука за подмножество од ознаките, со што има и побалансирана дистрибуција на ознаките во рамки на подмножествата кои се користат за тренирање на подмоделите. Со користење на HOMER, ознаките се организираат во хиерархија во форма на стебло, што се изведува со хиерархиско кластерирање на ознаките. Потоа за секој јазел во стеблото се гради посебен модел. Со користењето на овој метод се обезбедува повисока предиктивна моќ на моделот бидејќи погорните класификатори во хиерархијата ќе пронајдат погенерални разлики помеѓу функциите, а класификаторите во подолните нивоа се фокусираат на подеталните разлики помеѓу сличните функции.

Метод кој го адаптира алгритамот

Од методите кои вршат адаптација на алгоритамот, во оваа докторска дисертација предвид е земен ML_*k*NN (213) методот. ML_*k*NN (213) е метод за повеќезначна класификација кој се базира на методот *k* најблиски соседи (*k* nearest neighbours - *k*-nn) (155). Во литературата можат да се најдат повеќе методи за повеќезначна класификација кои се базираат на *k*-nn. Кај овие методи заедничко е тоа што прво се одредуваат најблиските соседи на испитуваниот примерок, но потоа одлучувањето за анотациите на примерокот врз основа на анотациите на најблиските соседи се прави на различен начин кај различен метод. Кај ML_*k*NN методот (213) со користење на Баесова теорема се одредуваат постериорните веројатности испитуваниот примерок да ја има

или нема дадена ознака, при што предвид се зема фреквенцијата на секоја ознака во рамки на соседите. За таа цел се дефинираат два настани, а тоа се: примерокот да ја има ознаката и примерокот да ја нема ознаката. Врз основа на постериорните веројатности од двата настани се одбира настанот за кој е добиена повисока веројатност, со што се одлучува дали ознаката ќе му се додели на испитуваниот примерок.

5.3.3. Евалуација на методот

Експериментални поставки

При градење на моделите со методот на бинарна релевантност (Binary relevance - BR) (207) и методот на множество на ознаки (Label Powerset - LP) (207), (211) ќе се користат класификаторите: C4.5 дрвата на одлука (153), случајни шуми (Random Forest - RF) (162) и Наивниот Баесов класификатор (Naïve Bayes - NB) (154), а при градење на моделите со HOMER (212) методот покрај овие три класификатори дополнително предвид се зема и методот *k* најблиски соседи (*k* nearest neighbours - *k*NN) (155). За HOMER методот направени се анализи со користење на различен број на кластери (*k*=3, *k*=5 и *k*=10). Во анализите кои ќе бидат презентирани, кај ML_*k*NN методот (213) се користи безтежинско гласање, и направени се експерименти со користење на различен број на најблиски соседи (*k*=1, *k*=2, *k*=3, *k*=5 и *k*=10). Во ова истражување се користат имплементациите на методите за повеќезначна класификација достапни во MULAN (207), кој претставува библиотека со методи за учење на повеќезначни проблеми.

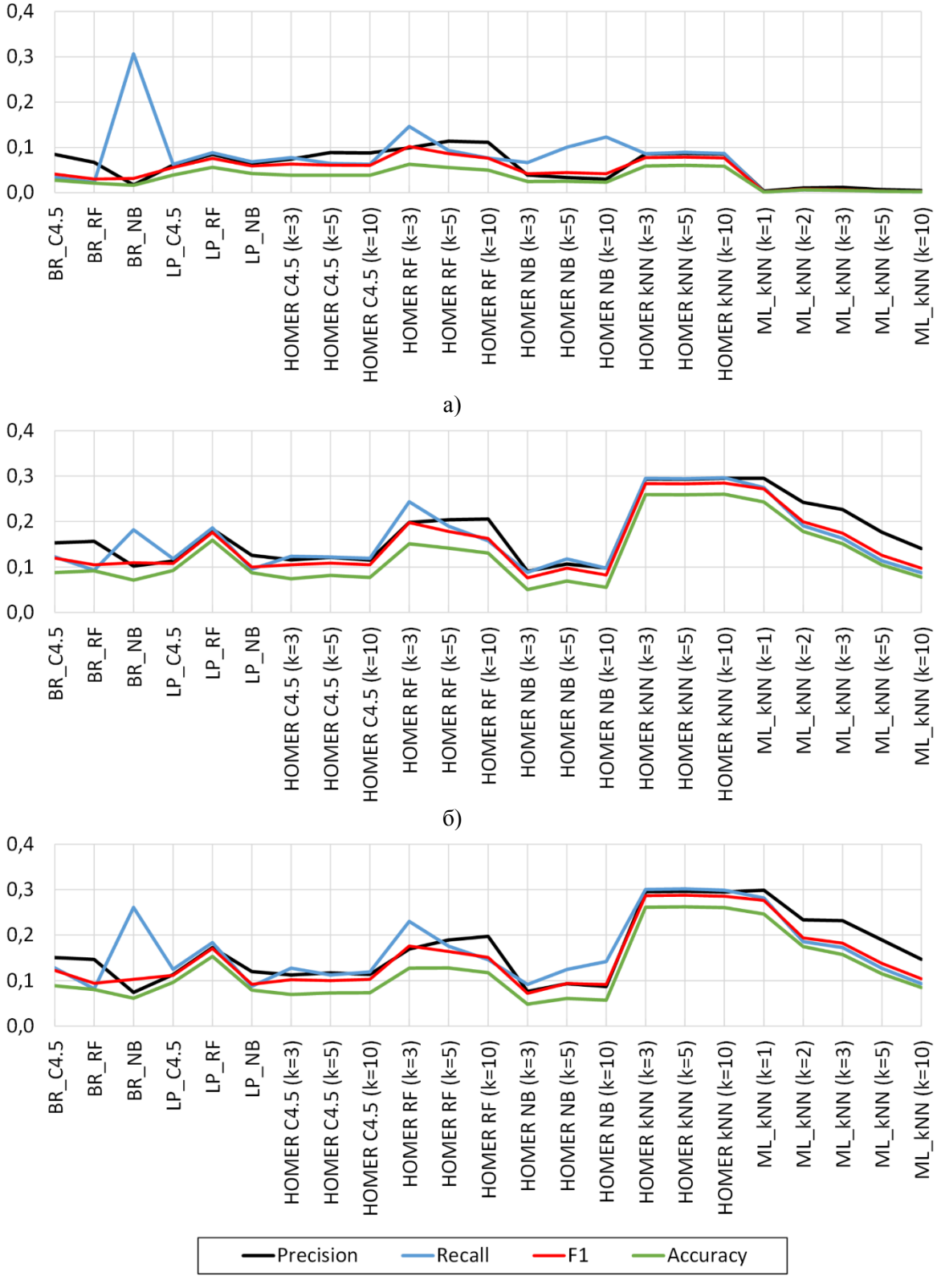
Опис на податочното множество

Податочното множество кое се користи во оваа анализа се добива така што од множеството *GO100_30*, кое се користи во анализите презентирани во секција 5.2, се филтрираат ланците за кои во BIND базата (48) има податоци за нивните сврзни региони. Потоа предвид се земаат ланците кои имаат барем една анотација за функција која ја има и помеѓу ланците за обука и помеѓу ланците за тестирање. Со ова се добива множество за обука кое содржи податоци за 2136 ланци, и множество за тестирање кое содржи податоци за 960 ланци. Бидејќи во податочното множество примероците соодветствуваат на сврзните региони, а не на соодветните протеински ланци, се добива множество за обука со 3167 примероци и множество за тестирање со 1449 примероци. Како што беше кажано, предвид се земаат функциите на ланците за кои има податоци и во множеството за обука и во множеството за тестирање, и со тоа се добиваат 757 различни протеински функции (ознаки).

Експериментални резултати

На Слика 5.4 даден приказ на резултатите добиени за мерките базирани на примерок со користење на локалните, глобалните и унијата од локални и глобални карактеристики. Слично, на Слика 5.5. се прикажани вредностите добиени за мерките базирани на ознака доколку се користат само локалните карактеристики, само глобалните карактеристики, или пак ако се користат сите карактеристики заедно. Во Табела Д3, Табела Д4, Табела Д5 и Табела Д6 дадени во Додатокот може да се погледнат подетални информации за вредностите на евалуациските мерки добиени за моделите.

Од резултатите може да се забележи дека предиктивната моќ на моделите добиени само со користење на локалните карактеристики е значително помала отколку на моделите кои ги користат глобалните карактеристики. Во однос на мерките базирани на примерок, со користење само на локалните карактеристики најдобри резултати се добиваат со користење на HOMER методот во комбинација со *k*NN класификаторот, додека со користење на ML_*k*NN се добиваат најлоши резултати. Од друга страна, ако се користат глобалните карактеристики тогаш најдобрите резултати се добиваат со користење на HOMER метод во комбинација со *k*NN класификаторот и ML_*k*NN за k=1. Ова покажува дека при одлучување врз основа на најблиските соседи, ако се користат само локалните карактеристики подобро е да се работи со помали податочни множества во кои се поделени примероците според поспецифични функции (како што прави HOMER со користење на *k*NN класификаторот), наместо да се работи со сите функции наеднаш (како што прави ML_*k*NN). При користење на глобалните карактеристики, кај ML_*k*NN методот перформансите опаѓаат со зголемувањето на бројот на најблиски соседи кои се земаат предвид. При користење на методот множество ознаки најдобри резултати се добиваат со користење на Random Forest, а C4.5 и Naïve Bayes се покажаа како помалку соодветни класификатори. Со користење на HOMER методот највисоки вредности за мерките базирани на примерок се добиваат со користење на *k*NN класификаторот, а потоа следат моделите добиени со Random Forest, C4.5 и најслаби резултати се добиваат со Naïve Bayes класификаторот. Naïve Bayes класификаторот е најмалку соодветен за да се користи кај HOMER методот бидејќи при градењето на подмоделите предвид се земаат подмножества од примероци преку кои неможе да се извлечат релевантни одлуки врз основа на дистрибуцијата на атрибутите во тоа помало податочно множество.

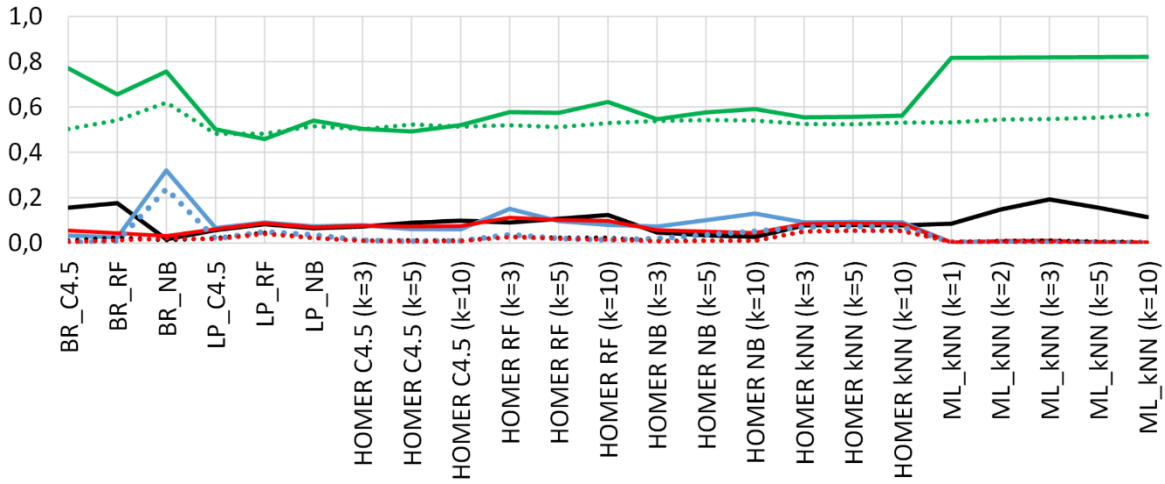


а)

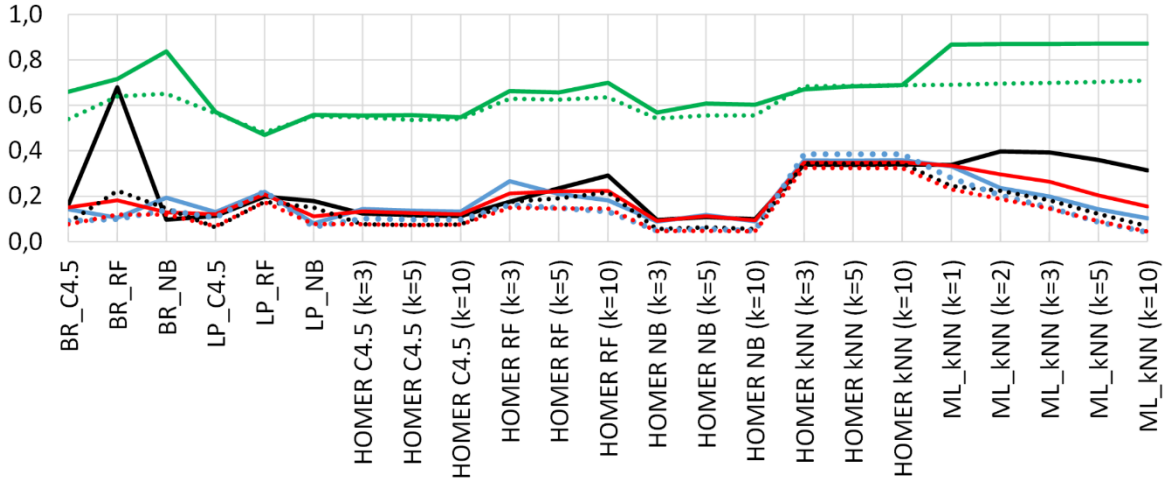
б)

в)

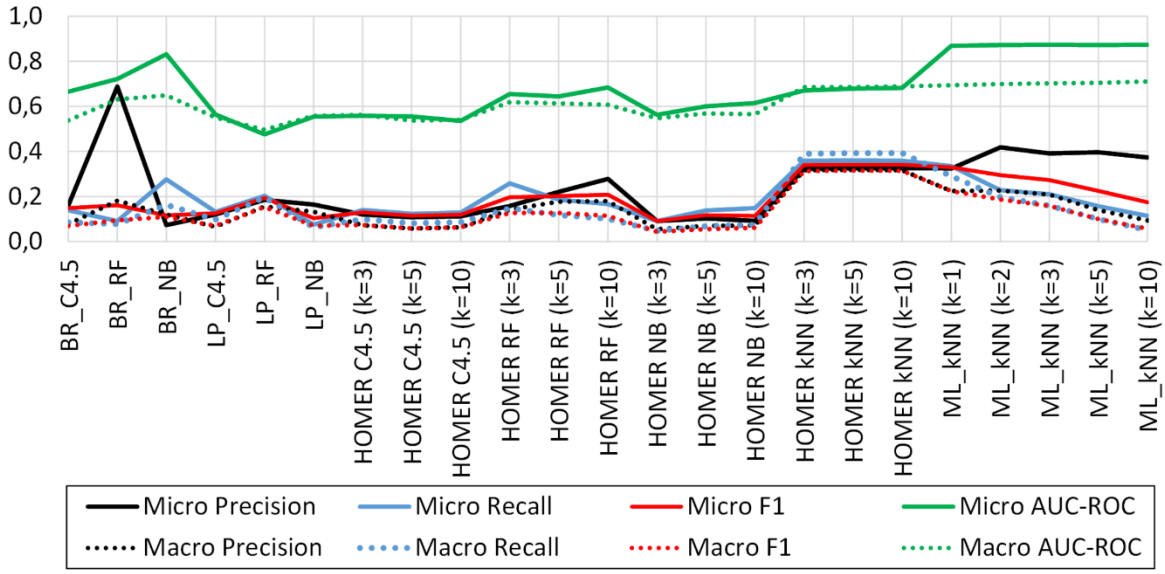
Слика 5.4 Резултати за мерките базирани на примерок за повеќезначните модели со користење на а) локалните карактеристики, б) глобалните карактеристики и в) сите карактеристики.



а)



б)



в)

Слика 5.5 Резултати за мерките базирани на ознака за повеќезначните модели со користење на а) локалните карактеристики, б) глобалните карактеристики и в) сите карактеристики.

Ако се разгледаат мерките базирани на ознака може повторно да се утврди дека со користење на глобалните карактеристики се добиваат подобри резултати отколку ако се користат само локалните карактеристики. При тоа најдобро е да се користат и локалните и глобалните карактеристики. Со користење на глобалните карактеристики, слично како што беше за мерките базирани на примерок, така и за мерките базирани на ознака најдобри резултати се добиваат со ML_*k*NN и HOMER во комбинација со *k*NN класификаторот. Вреди да се напомене дека кај моделите добиени со ML_*k*NN има најголема разлика помеѓу микро и макро мерките за AUC-ROC и Precision, за разлика од моделите кое се добиени со другите методи. Имено микро мерките се значително повисоки во однос на макро мерките што се должи на фактот дека со ML_*k*NN поточно се предвидуваат погенералните (позастапените) функции наспрема поспецифичните (ретките) функции. Ова се должи на тоа што за поспецифичните функции има мал број на анотации, па многу е потешко во репрезентативното множество од ланци кои немаат висока сличност помеѓу себе да се пронајде како близок сосед протеински ланец кој ја има дадената функција. Во однос на HOMER методот може да се забележи дека вредноста на параметарот *k* многу малку влијае врз перформансите на моделите (освен кај Random Forest класификаторот), а исто така кај овој метод има најмали разлики помеѓу микро и макро мерките. Ова покажува дека овој метод не ги фаворизира ниту генералните ниту ретките ознаки. AUC-ROC мерките се зголемуваат со користење на поголем број на најблиски соседи кај ML_*k*NN методот, што не е случај за останатите мерки. Во однос на методот бинарна релевантност интересно е да се напомене дека при користење на Naïve Bayes класификаторот се добива значително повисок одзив отколку со другите методи ако предвид се земат локалните карактеристики. Тоа значи дека доколку за секоја ознака се гради посебен модел за да се одлучи дали да се додели одредена ознака или не, без при тоа да се земе предвид зависноста меѓу ознаките, тогаш врз основа на карактеристиките на сврзниот регион со Наивниот Баесов класификатор се генерира модел кој успешно предвидува поголем дел од релевантните функции. Со користење на Random Forest класификаторот кај сите три методи (BR, LP и HOMER) се добива повисока прецизност отколку со останатите класификатори.

Од оваа анализа може да се заклучи дека најдобри резултати се добиваат со HOMER и ML_*k*NN методите. Како што беше кажано погоре, со ML_*k*NN подобро се препознаваат поопштите функции што доведува до позначителна разлика помеѓу микро и макро мерките базирани на ознака. Од друга страна HOMER методот се стреми да изгради модел претставен преку хиерархија од подмодели со цел да не се фаворизираат подоминантните ознаки, како и да се изградат подмодели кои се фокусирани да попрецизно ги препознаваат поспецифичните функции.

5.4. Споредба на предложените методи за одредување на функции на протеински структури

Во претходните две секции беа презентирани два методи за одредување на функции на протеински структури. Кај првиот метод анотирањето се прави врз основа на анотациите на најблиските соседи на испитуваниот примерок. Во вториот метод со користење на методи за повеќезначна класификација се гради предиктивен модел, при што предвид се земаат локалните карактеристики на сврзниот регион за која се прави предвидувањето, како и глобалните карактеристики од протеинскиот дескриптор базиран на рамномерна интерполација на скелетот. Двата методи беа евалуирани, но во претходните две секции се користат различни податочни множества. Со цел да се добие подобра слика за моќта на методите, во оваа секција ќе бидат претставени резултатите за двата пристапи добиени врз множеството кое се користи во секција 5.3. Во Табела 5.18 се прикажани резултатите за најдобрите модели.

Метод базиран на структурно поранмнување			
Метод	**рамн. интер.**	**бранчиња**	**MASASW**
	***D*=128 L_1 норма**	**Haar *D*=200**	**64x64**
k	**3**	**3**	**3**
v_{min}	**0.09**	**30**	**3000**
n_1	**1**	**1**	**1**
n_2	**1**	**1**	**1**
Precision	0.23	0.31	0.36
Recall	0.22	0.31	0.35
F_1	0.21	0.29	0.34
Accuracy	0.20	0.26	0.30
$Precision_{micro}$	0.61	0.44	0.52
$Recall_{micro}$	0.27	0.36	0.41
$F_{1\ micro}$	0.37	0.40	0.46
$Precision_{macro}$	0.46	0.45	0.49
$Recall_{macro}$	0.33	0.40	0.44
$F_{1\ macro}$	0.35	0.39	0.43

Метод базиран на локалните и глобалните карактеристики				
Метод	**Бинарна релевантност**	**Множество на ознаки**	**HOMER**	**ML_*k*NN**
	Random Forest	**Random Forest**	***k*NN, *k*=5**	***k*=1**
Precision	0.15	0.17	0.30	0.30
Recall	0.08	0.18	0.30	0.28
F_1	0.09	0.17	0.29	0.28
Accuracy	0.08	0.15	0.26	0.25
$Precision_{micro}$	0.69	0.18	0.32	0.32
$Recall_{micro}$	0.09	0.20	0.36	0.33
$F_{1\ micro}$	0.16	0.19	0.34	0.33
$Precision_{macro}$	0.18	0.16	0.32	0.22
$Recall_{macro}$	0.08	0.20	0.39	0.29
$F_{1\ macro}$	0.09	0.15	0.32	0.22

Табела 5.18 Сумарни резултати од споредбата на двата методи за одредување на функциите на протеинските структури.

Од резултатите може да се забележи дека со користење на првиот метод кој се базира на структурно порамнување на протеинските структури се добиваат подобри перформанси отколку со вториот метод. Најдобри резултати се добија со првиот метод со користење на MASASW методот за одредување на сличноста помеѓу две протеински структури при одредување на најблиските соседи на испитуваниот примерок. Потоа, следниот најдобар модел е добиен со првиот метод со користење на протеинскиот дескриптор базиран на бранчиња, по кого следи моделот генериран со првиот метод користејќи го дескрипторот базиран на рамномерна интерполација на протеинскиот скелет. Интересно е да се забележи дека со првиот метод генерално се добиваат многу помали разлики помеѓу микро и макро мерките, што укажува на тоа дека овој пристап не се фокусира само на погенералните (позастапените) функции, туку успешно извлекува знаење и за препознавање на поспецифичните (поретките) функции. Во однос на вториот метод, со користење на методот бинарна релевантност се добиваат најслаби резултати, по кого следи методот множества на ознаки. Методот множества на ознаки е подобар од методот бинарна релевантност бидејќи предвид ја зема зависноста меѓу ознаките, односно се прави анализа на тоа кои ознаки се појавуваат заедно како анотации на еден ист примерок. Токму на тоа се должат и подобрите предиктивни перформанси на методот множество на ознаки. Од друга страна со користење на методите HOMER и ML_*k*NN се добиваат подобри резултати, што покажува дека методите кои градат хиерархија од повеќезначни подмодели, како и методите кои го адаптираат алгоритамот за решавање на повеќезначни проблеми генерално се посоодветни од методите кои проблемот на повеќезначна класификација го трансформираат во еден повеќекласен (како кај методот множества на ознаки) или во неколку бинарни (како кај методот бинарна релевантност) проблеми. Од методите за повеќезначна класификација, со користење на HOMER методот се добиваат многу помали разлики помеѓу микро и макро верзиите на евалуациските мерки, за разлика од другите методи, што се постигнува преку градењето на хиерархија од подмодели кои не ги фаворизираат позастапените ознаки.

Ако се направи споредба на мерките базирани на примерок, може да се забележи дека со користење на првиот метод во комбинација со дескрипторот базиран на рамномерна интерполација на скелетот се добиваат послаби вредности за мерките базирани на примерок во однос на вредностите кои се добиени за HOMER и ML_*k*NN. Ова покажува дека при користење на првиот метод базирајќи се на дескрипторот добиен со интерполација на скелетот успешно се извлекува знаење за препознавање на ознаките (мерките базирани на ознака се подобри), но за примероците кои немаат доволно слични соседи неможе да се донесат точни одлуки за нивните анотации. Како резултат на ова, кај овој модел се добиваат помали вредности за мерките базирани на примерок. Исто така кај овој модел може да се забележи дека кај микро мерките многу е повисока прецизноста наспрема одзивот, додека кај макро мерките нема толку голема

разлика помеѓу овие две мерки. Ова покажува дека овој модел донесува голем број на точни одлуки за погенералните функции (висок $Precision_{micro}$), но не кај сите испитувани ланци успева да одлучи дека дадената генерална функција е релевантна (низок $Recall_{micro}$). Ова се должи на тоа што испитуваниот протеин со своите најблиски соседи во принцип би делел голем број на исти специфични функции кои ќе се препознаваат со висока прецизност и одзив. Но од друга страна, кај голем број на ланци кои се со мала структурна сличност со ланците за обука, за дадена релевантна функција соседите нема да го обезбедат потребниот минимален глас за да се додели таа функција на испитуваниот протеински ланец. Како резултат на ова, со користење на дескрипторот базиран на рамномерна интерполација на скелетот се добива голема микро прецизност, но мал микро одзив. Ако се погледнат резултатите за микро мерките за останатите методи кои вршат структурно порамнување, ќе се забележи дека со овие методи за ланците кои имаат мала структурна сличност со ланците за обука многу поточно како соседи се пронаоѓаат ланците со кои дадениот ланец дели исти генерални функции.

6

ЗАКЛУЧОК

Во оваа докторска дисертација беа претставени неколку методи за пребарување на протеински структури, кои можат да се користат при класификација на протеински структури, како и при одредување на функциите на тие структури. Беше направена детална евалуација на овие методи, и истите беа споредени со неколку познати методи за порамнување на протеински структури. Со користење на овие методи може да се одреди структурната сличност на протеинските структури, која потоа може да се користи за класификација на протеински ланци во SCOP домени, или пак за функционално анотирање на тие ланци. Беше предложен метод за одредување на протеинските функции со кој одлучувањето за анотациите на испитуваниот ланец се прави врз основа на анотациите на неговите најблиски соседи. За таа намена беа применети методи за пребарување на протеински структури, со цел да се одредат најблиските соседи на примерокот кој се испитува. Наместо преку структурно порамнување на протеински структури, одредувањето на протеинските функции може да се прави и врз основа на карактеристиките на сврзните делови каде што настанува интеракција со друга структура. За таа намена, во оваа дисертација беа предложени неколку методи за препознавање на амино-киселинските остатоци кои формираат сврзен регион. Потоа, беше предложен метод за градење на модели за анотирање на протеински структури врз основа локалните и глобалните карактеристики на структурата користејќи методи за повеќезначна класификација. Локалните карактеристики се однесуваат на

својствата кои ги има даденият сврзен регион кој се испитува, додека глобалните карактеристики носат информација за поставеноста на тродимензионалната протеинска структура во просторот. Дополнително беше направена споредба на двата предложени методи за одредување на протеински функции.

Прво беше претставен протеинскиот воксел-базиран дескриптор во кој покрај карактеристиките на терциерната структура се вклучени и поважните карактеристики на примарната и секундарната структура. Во процесот на споредба, на различна карактеристика и се доделува различна тежина, па во иднина може да се испита како изборот на тежините на карактеристиките да се прави на автоматизиран начин преку кој ќе се определат некои оптимални тежини за даденото податочно множество. Во овој дескриптор дополнително може да се воведат и некои други својства на примарната и секундарната структура со цел да се обезбеди попрецизно пребарување. Протеинскиот воксел-базиран дескриптор предвид ја зема распореденоста на целата структура во тродимензионалниот простор. Сепак, анализите покажуваат дека подобро е предвид да се земат само Cα атомите кои го сочинуваат протеинскиот скелет. Па затоа останатите методи кои се развиени во ова истражување предвид ја земаат само просторната поставеност на Cα атомите. Потоа беше презентиран протеинскиот дескриптор базиран на интерполација на протеинскиот скелет. Бидејќи целиме кон извлекување на дескриптори чија должина нема да зависи од протеинските структури, а од друга страна протеинските ланци имаат различен број на Cα атоми, затоа прво се врши интерполација на скелетот на протеинот со предефиниран број на интерполациски точки. Беа презентирани два начини на интерполација на протеинскиот скелет, а тоа се рамномерната и нерамномерната интерполација. Со рамномерна интерполација секој дел од скелетот се анализира со иста резолуција, додека при нерамномерна интерполација пооддалечените Cα атоми се анализираат со поголема резолуција. Резултатите покажаа дека рамномерната интерполација е посоодветна бидејќи обезбедува пофина апроксимација на протеинскиот скелет. Со користење на рамномерна интерполација не само што се обезбедува повисока прецизност, туку и времето потребно за екстракција е значително помало отколку кај нерамномерна интерполација. Кај овој дескриптор може да се користат различен број на карактеристики, со што се дефинира и бројот на интерполациски точки, па преку одбирање на овој број може да се прави анализа на посакуваната резолуција. Експерименталните резултати покажаа дека намалувањето на бројот на карактеристики од 512 на 64 резултира со мало опаѓање во прецизноста, па со тоа користејќи релативно мал број на карактеристики може да обезбедиме висока прецизност при пребарувањето. Овој метод е наједноставен за имплементација од сите презентирани методи за пребарување на протеински структури, а сепак има споредлива прецизност со другите методи кои се значително покомплексни. Потоа беа презентирани дескрипторите базирани на бранчиња. Екстракцијата на овие дескриптори

започнува со извлекување на матрицата на растојанија, а потоа врз оваа матрица се применуваат различни трансформации на бранчиња за да се извлечат карактеристиките на дескрипторот. Во дескрипторите кои се презентирани во оваа докторска дисертација се врши декомпозиција на оваа матрица до последното ниво. Беше направена детална анализа за прецизноста која се добива со дескрипторите базирани на бранчиња користејќи различен број на карактеристики. Од анализата се утврди дека доволно е да се земат помеѓу 150 и 200 карактеристики, бидејќи понатамошните апроксимативни коефициенти содржат информации за некои поситни локални детали, а не се однесуваат на глобалниот облик на протеинската структура. Овие дескриптори беа споредени помеѓу себе, а исто така беа споредени и со протеинскиот дескриптор базиран на Haar бранчето предложен во (73). При екстракција на дескрипторот предложен во (73) декомпозицијата се прави до четврто ниво, и на крај предвид се земаат 36 карактеристики. Сепак, анализите покажаа дека овие 36 карактеристики не се доволни за да обезбедат дескриптор со висока дискриминаторна моќ. Резултатите покажаа дека Daubechies2 бранчето обезбедува најдобро препознавање на протеинските ланци кои припаѓаат во ист SCOP домеин. Покрај овие три методи за одредување на сличноста помеѓу протеински структури, дополнително беше воведен и MASASW методот со кој се прави порамнување на матриците на растојанија на двете структури кои се споредуваат. Овој метод се базира на постоечки метод за порамнување на секвенци, каде истиот е проширен во 2Д простор за порамнување на матрици. Со цел да се избегне исцрпното пребарување кое се прави кај DALI (19) и MatAlign (70) методите, кај MASASW се користи лизгачки прозорец преку кој се одредува контекстниот прозорец во рамки на кој дадена редица од првата матрица ќе се порамнува со редиците од втората матрица. Со воведувањето на лизгачки прозорец се постигна значително забрзување, без при тоа да има значително опаѓање во предиктивната моќ на методот. Од сите методи кои беа презентирани во оваа дисертација, MASASW методот се покажа како најпрецизен, но за сметка на тоа има и поголема временска комплексност од останатите методи кои се базираат на споредба на карактеристични вектори. Резултатите покажаа дека методите кои предвид ги земаат само Cα атомите постигнуваат значително повисока прецизност за разлика од методите кои предвид ја земаат целата протеинска структура. Предложените методи дополнително беа споредени со неколку познати методи кои се среќаваат во литературата, и тоа: DALI (19), CE (20), MSVNS (71), протеинскиот фрактален дескриптор (72) и протеинскиот дескриптор базиран на Haar бранчето предложен во (73). Резултатите покажаа дека со DALI методот се обезбедува најпрецизно пребарување, но истото трае значително подолго. Потоа следи предложениот MASASW метод. Потоа со приближно иста прецизност следат CE методот, дескрипторот базиран на интерполација на скелетот и предложените дескриптори базирани на бранчиња. Потоа со значително пониска прецизност доаѓаат MSVNS и воксел-базираниот дескриптор. При

тоа MSVNS метод иако е временски многу комплексен, сепак обезбедува значително помала прецизност од претходните методи. Потоа следи дескрипторот базиран на Haar бранчето предложен во (73), и на крај со убедливо најслаби предиктивни перформанси се покажа фракталниот дескриптор кај кој двете карактеристики кои ги содржи дескрипторот се покажаа како недоволни за да се претстави комплетната информација за комплексната протеинска структура.

Со цел да се подобри прецизноста на протеинскиот дескриптор базиран на рамномерна интерполација на протеинскиот скелет, во него дополнително беа вклучени неколку карактеристики на амино-киселинските остатоци, и тоа дофатливата површина (ASA), релативната дофатлива површина (RASA), индексот на длабочина (DPX), индексот на испакнатост (CX) и хидрофобичноста. За таа цел, прво се врши рамномерна интерполација на протеинскиот скелет, а потоа за секоја интерполациска точка покрај Евклидовото растојание од дадената точка до центарот на маса, дополнително предвид се земаат и карактеристиките на најблискиот амино-киселински остаток. Беа направени анализи за да се одреди дали овој проширен дескриптор обезбедува подобро пребарување во однос на основниот дескриптор. Резултатите покажаа дека воведувањето на дополнителни карактеристики доведува до зголемување на прецизноста при пребарување на протеински структури кои припаѓаат во ист SCOP домен. Исто така беа изградени и класификациски модели во кои на влез покрај карактеристиките кои ги содржи протеинскиот дескриптор базиран на рамномерна интерполација на скелетот, дополнително му се проследуваат и останатите карактеристики кои се додаваат во проширениот дескриптор. Резултатите покажаа дека класификациските модели со кои се врши класификација на протеинските ланци во SCOP домени постигнуваат повисока класификациска точност со воведувањето на дополнителните карактеристики. Понатаму може да се разгледаат и други дополнителни карактеристики кои може да се земат предвид во дескрипторите, со цел да се зголеми прецизноста при пребарување и класификација.

При анотација на протеински структури, покрај глобалните карактеристики на структура предвид може да се земат и некои локални својства кои го опишуваат регионот од структурата каде настанува интеракција со некој друг протеински ланец. Затоа, во оваа докторска дисертација беа предложени неколку методи за детекција на сврзните делови од протеинската структура. За таа цел прво се извлекуваат карактеристиките на амино-киселинските остатоци, а потоа користејќи одреден класификациски метод се гради предиктивен модел. Во оваа дисертација предвид се земаат следниве карактеристики на атомите: дофатливата површина (ASA), релативната дофатлива површина (RASA), индексот на длабочина (DPX) и индексот на испакнатост (CX). Бидејќи еден амино-киселински остаток се состои од повеќе атоми, затоа се прави агрегирање на карактеристиките на атомите кои влегуваат во составот на тој остаток, и тоа се

одредува вкупната, просечната, минималната и максималната вредност. Дополнително, за секој остаток предвид се зема и хидрофобичноста на амино-киселината на која се однесува тој остаток. Со користење на четири карактеристики на амино-киселинските остатоци, и тоа totalASA, avgDPX, avgCX и хидрофобичноста, беа изградени класификациски модели за одредување на сврзните делови од протеинската структура. Прво, беа изградени индивидуални модели со користење на неколку класични класификациски методи. Со цел да се подобрат перформансите на моделите, беа генерирани ансамбли кои агрегираат повеќе модели. Целта на ансамблите е со комбинирање на моделите да се постигне надминување на недостатоците на индивидуалните модели. Во ова истражување предвид се земени bagging и boosting техниките. Експерименталните резултати ги потврдија очекувањата дека ансамблите ќе обезбедат попрецизно предвидување во однос на претходно генерираните индивидуални модели. Со анализата се покажа дека со boosting техниката се добиваат попрецизни модели отколку со bagging, што се должи на фактот дека кај boosting се форсира учењето на потешките примероци кои се погрешно класифицирани од моделите добиени во претходните чекори. Сепак кај boosting техниката треба да се внимава да не се преобучи моделот, што и се случи во анализата при користењето на C4.5 класификатор.

Сепак, класичните методи за класификација се многу осетливи на мали промени на карактеристиките во податочното множество. Од друга страна, во текот на еволуцијата кај протеинските структури настануваат одредени промени што резултира и со промени во карактеристиките на нивните амино-киселински остатоци. Па од тука дојде идејата да се воведе непрецизираната логика за одредување на сврзните региони од структурата, со цел да се надминат недостатоците на класичните методи. Во оваа дисертација беа презентирани методите од-дното-нагоре (174) и од-врвот-надолу (175), и истите се применети за препознавање на амино-киселинските остатоци каде настануваат интеракции. Со овие методи се градат Непрецизирани дрва на припадност (НДП) кои се базираат на непрецизираната логика. Главната разлика помеѓу овие два методи е тоа што кај методот од-врвот-надолу насоката на градење на стеблото е сменета така што стеблото се гради од коренот кон листовите. Друга важна разлика помеѓу овие методи е тоа што кај методот од-врвот-надолу се користи поинаков критериум за прекинување на градењето на стеблото, со што се овозможува комплексноста на моделот да се адаптира според комплексноста на проблемот кој се решава. За методот од-дното-нагоре беа направени анализи со користење на различен број на непрецизирани термини по карактеристика, како и со различни типови на функции на припадност. Покрај познатата мерка за сличност RMSE, исто така беа воведени и две нови мерки, а тоа се $RMSE^2$ и G_MAX. Во истражувањето за одредување на индикаторските карактеристики на дијатомеите кои живеат до водните екосистеми, со овие мерки се генерираа попрецизни модели. Во истражувањето направено за одредување на сврзните делови од

протеинот, $RMSE^2$ мерката се покажа како помалку соодветна, додека со G_MAX се добива приближна прецизност како и со RMSE мерката. При користењето на двата методи за градење НДП беа направени анализи со користење на различни оператори за агрегација, и резултатите покажаа дека генерално воведувањето на дополнителни оператори доведува до зголемување на предиктивната моќ на моделите, но од друга страна времето потребно за обука на моделите расте со бројот на оператори за агрегација кои се земаат предвид. Методот од-врвот-надолу се покажа како попрецизен, со што се потврдија очекувањата. Имено, при градењето на моделот со користење на метод од-дното-нагоре, тековното најдобро стебло се агрегира со друго стебло, така што тековното најдобро стебло се поставува како дете на новодобиеното стебло. Со ова при агрегирањето се добива модел кој има значителни разлики со моделот од претходниот чекор. Од друга страна кај методот од-врвот-надолу во секој чекор се прават мали промени во моделот, со цел да истиот пофино се прилагоди кон примероците.

При градењето на претходните модели базирани на класификациски методи од класичната и непрецизираната логика предвид беа земени четири карактеристики на амино-киселинските остатоци. Со цел да се подобри предиктивната моќ на моделите, дополнително предвид се земаат и останатите карактеристики на амино-киселинските остатоци кои беа извлечени при формирање на проширениот дескриптор базиран на рамномерна интерполација на скелетот. Со користење на техники за избор и трансформација на карактеристики, од овие карактеристики беа селектирани множествата од најрелевантни карактеристики. Во оваа докторска дисертација е даден детален преглед на најпознатите техники за избор и трансформација на карактеристики, и при тоа се наведени и нивните предности и недостатоци. Од техниките за избор на карактеристики предвид беа земени неколку техники за филтрирање и неколку техники за обвиткување. Од техниките за филтрирање применети се техники кои индивидуално ги рангираат карактеристиките, како и техники кои евалуираат подмножества од карактеристики со цел да ја максимизираат корелацијата со класниот атрибут и истовремено да ја минимизираат редундантноста помеѓу избраните карактеристики. Беа направени детални анализи со користење на повеќе техники за избор на карактеристики во комбинација со различни класификатори за градење на моделите. Резултатите ги потврдија очекувањата дека наместо да се користат сите карактеристики, подобро е да се земат предвид само најрелевантните својства. Со ваквото селектирање не само што се подобрува прецизноста, туку дополнително се намалува комплексноста на моделот, како и времињата на обука и тестирање. Од техниките за филтрирање, се потврди очекувањето дека со одбирање на најдоброто подмножество од карактеристики се добива попрецизен модел отколку со одбирање на најдобрите карактеристики кои се рангирани индивидуално. За да се добие подобра слика за предиктивната моќ на моделите, дополнително беше направена споредба со неколку познати методи за одредување на сврзните делови од

протеинската структура. Методите кои се земени во споредбата се базираат на различни типови на информации, па така едни методи се базираат на растојанието помеѓу остатоците, друга група методи го анализираат зачувувањето (конзервацијата) на секвенцата и/или структурата, трета група вршат детекција на џебови, а исто така предвид е земен и PRISM методот кај кој се врши структурно порамнување. Дел од методите кои се земени предвид во анализата комбинираат различни типови на информации со цел да обезбедат поточно предвидување. Анализите покажаа дека постоечките методи кои се користат во оваа споредба имаат различна предиктивна моќ на различни множества. Имено, овие методи се прилагодени за поточно донесување на одлуки за специфична група на интеракции, но сепак не обезбедуваат генерален модел кој ќе биде применлив за различни групи на протеини. Со предложениот пристап се добиваат споредливи перформанси на сите множества кои беа земени предвид, што покажува дека овој метод е постабилен и неговите перформанси не зависат премногу од податочното множество кое се користи за обука. Предложениот пристап обезбедува генералност бидејќи истиот е применлив за различни типови на интеракции. Иако овој пристап бележи скромни перформанси, сепак се очекува дека истиот ќе обезбеди поточно предвидување доколку моделот се гради за специфични интеракции и протеини. При тоа, бидејќи во предложениот пристап се користат техники за избор и трансформација на карактеристики, се очекува дека овој пристап ќе се обезбеди само-адаптибилност на моделот. Имено, со самиот пристап во првата фаза автоматски ќе се одберат најрелевантните карактеристики за испитуваната група на протеини. Како резултат на ова, се очекува дека ако овој пристап се употреби за генерирање на модел кој преставува каскада од подмодели, тогаш преку само-адаптибилноста обезбедена со овој пристап ќе се овозможи секој подмодел во каскадата автоматски да се адаптира така што ќе го одбере најдоброто множество од карактеристики за примероците кои ќе се користат при обука на тој подмодел. Во иднина, покрај градењето на генерален модел кој ќе претставува каскада од само-адаптибилни подмодели кои се наменети за предвидувања за потесна група на интеракции и протеини, дополнително предиктивната моќ може да се зголеми така што предвид ќе се земат и други карактеристики на амино-киселинските остатоци кои можеби ќе понудат подобра дискриминација помеѓу остатоците кои се дел од сврзен регион, наспроти останатите остатоци.

Во оваа докторска дисертација методите за пребарување на протеински структури и методите за препознавање на сврзните делови од протеините беа воведени со цел предвидувањето од моделите потоа да се искористи за функционално анотирање на протеински структури. Во таа насока, во ова истражување дополнително беа воведени два методи за одредување на протеинските функции. Првиот метод се базира на структурно порамнување користејќи ги предложените методи за пребарување на слични протеински структури. Со овој метод анотирањето на испитуваниот протеин се прави врз основа на анотациите на неговите најблиски соседи. Кај

вториот метод, со користење на методи за повеќезначна класификација се гради модел за предвидување на протеинските функции врз основа на локалните карактеристики на сврзниот регион за кој се прави предвидувањето, како и врз основа на глобалните карактеристики кои даваат информација за поставеноста на структурата во просторот.

Првиот метод е варијанта на методот на *k* најблиски соседи, при што е направена адаптација за решавање на повеќезначни проблеми. Кај овој метод беше направена детална анализа со цел оптимално да се одберат можните поставки, како на пример должината на дескрипторот, користење на различна метрика за растојание, избор на најсоодветното бранче итн. Исто така беа евалуирани моделите добиени со користење на безтежинско и тежинско гласање. Во однос на тежинското гласање, тежината на гласот на даден сосед може да се дефинира преку растојанието од тој сосед до испитуваниот примерок, како и врз основа на редната позиција на тој сосед. Беа направени експерименти со користење на безтежинско и тежинско гласање, и резултатите покажаа дека најдобри резултати се добиваат со моделите каде тежините се дефинираат преку двата параметри (растојанието и редната позиција). Анализите покажаа дека кај првиот предложен метод за анотација на протеински структури најпрецизен модел се добива со користење на MASASW методот ($F_{1\ micro}$=0.59) при одредување на сличноста помеѓу испитуваниот примерок и даден примерок за обука. Потоа следи дескрипторот базиран на Haar бранчето ($F_{1\ micro}$=0.52), па дескрипторот базиран на рамномерна интерполација на скелетот ($F_{1\ micro}$=0.47).

Кај вториот метод, врз основа на локалните карактеристики на сврзниот регион и врз основа на глобалните карактеристики на целата протеинска структура, се гради предиктивен модел со кој се решава повеќезначен класификациски проблем. За оваа намена предвид се земени методи кои повеќезначниот проблем го трансформираат во еден или повеќе проблеми со еден класен атрибут, а исто така предвид е земен и метод кој наместо трансформација на проблемот прави адаптирање на одреден методот за повеќекласна класификација за решавање на повеќезначни проблеми. Од првата категорија предвид е земен методот на бинарна релевантност каде одлучувањето се прави така што се гледаат разликите помеѓу примероците кои ја имаат и кои ја немаат испитуваната функција. Сепак, кај овој метод не се земаат предвид зависностите помеѓу функциите, односно не се зема предвид информацијата за тоа кои функции се појавуваат заедно како анотации на исти протеини. Од друга страна, кај вториот метод од оваа категорија, тоа е методот множества на ознаки, проблемот се трансформира така што повеќезначниот проблем се трансформира во еден повеќекласен проблем, при што кај новодобиениот проблем се користи посебна ознака за секое различно множество од ознаки кои се сретнуваат како анотации на примероците за обука. Во првата категорија припаѓа и HOMER методот кај кој се прави хиерархиски модел составен од повеќе подмодели кои се обучени за препознавање на потесна група на протеински функции. Со ова кај HOMER методот се обезбедува да во погорните нивоа

се анализираат погенералните разлики помеѓу разнородни функции, а потоа во пониските нивоа се разгледуваат разликите помеѓу послични функции. Од втората категорија, а тоа е категоријата на методи кај кои даден метод за повеќекласна класификација се адаптира за повеќезначна класификација, предвид е земен ML_*k*NN методот. Методот ML_*k*NN, како што покажува и самото име, се базира на методот на *k* најблиски соседи, каде врз основа на анотациите на најблиските соседи со користење на Баесовата теорема се одлучува дали дадена ознака да му се додели на испитуваниот примерок или не. Беа направени детални анализи при што предвид беа земени неколку класификатори при градење на повеќекласните модели кои го формираат главниот модел. Исто така беа направени анализи за испитување на влијанието на различните поставки кои може да се користат при градење на моделите. Резултатите покажаа дека со методот на бинарна релевантност се добиваат најслаби резултати што се должи на фактот што при градење на моделите не се зема предвид зависноста помеѓу различните ознаки. Кај овој метод се појавува најголема разлика помеѓу микро и макро мерките што покажува дека тој е наклонет кон погенералните (позастапените) ознаки. Потоа следи методот множества на ознаки кој предвид ја зема информацијата за тоа кои ознаки се појавуваат заедно, што доведува и до поточни предвидувања. Сепак со HOMER и со ML_*k*NN методите се добива попрецизен модел отколку со претходните два методи. При тоа ML_*k*NN подобро ги препознава позастапените ознаки, па затоа и се добиваат поголеми разлики помеѓу микро и макро мерките за евалуација. Од друга страна кај HOMER методот има најмала разлика помеѓу микро и макро мерките, што покажува дека овој метод не прави фаворизирање на погенералните ознаки, што се обезбедува преку градењето на хиерархија од подмодели. Резултатите покажаа дека значително подобри резултати се добиваат при користење на глобалните наместо локалните карактеристики, а притоа најдобро е да се користат сите карактеристики заедно.

На крај беше направена споредба на двата предложени методи за функционално анотирање на протеински структури. Резултатите покажуваат дека со првиот метод се обезбедуваат попрецизни модели. При тоа најдобри резултати се добиваат доколку MASASW методот се користи при одредување на најблиските соседи на испитуваниот примерок. Потоа следи моделот добиен со користење на дескрипторот базиран на бранчиња, а најслаби перформанси се добиваат со користењето на едноставниот дескриптор базиран на рамномерна интерполација на скелетот на протеинот. Генерално, првиот метод не се фокусира на градење на модели кои се наклонети кон погенералните функции, што резултира со мали разлики помеѓу микро и макро мерките. Во однос на вториот метод, најслаби перформанси се добиваат со методот на бинарна релевантност и методот множества на ознаки, додека со HOMER и ML_*k*NN методите се добиваат споредливи резултати со моделите добиени со методот базиран на структурно порамнување. При тоа кај вториот метод најдобри резултати се добиваат со користење на HOMER методот.

Во иднина може да се направат анализи каде ќе се вклучат и други карактеристики на амино-киселинските остатоци кои можеби ќе се покажат како порелевантни за одредување на протеинските функции. Во однос на вториот предложен метод за одредување на протеинските функции, покрај четирите методи за повеќезначна класификација кои се користени во ова истражување, дополнително може да се применат и други методи кои можеби ќе понудат попрецизно предвидување на функциите. Во анализите направени во ова истражување предвид не е земена хиерархијата на функциите која е дефинирана со онтологијата Gene Ontology (GO). Доколку предвид се земе оваа хиерархија, можат да се подобрат перформансите на моделите со земањето во предвид на сличноста помеѓу функциите. Дополнително може да се направи модификација на евалуациските мерки така што во процесот на евалуација доколку за даден примерок се предвиди некоја функција која ја нема дадениот примерок, и ако таа функција има голема сличност со некоја од релевантните функции за тој примерок, тогаш тоа да не се смета како грешка, туку да се смета како точно предвидување со некоја помала тежина. На овој начин грешните позитивни предвидувања всушност ќе добијат некакви тежини при што предвид ќе се земе сличноста помеѓу дадената грешно предвидена функција и релевантните функции за тој протеин. Со користење на вакви модифицирани верзии на евалуациските мерки ќе се добие подобра слика за перформансите на методите.

7

РЕФЕРЕНЦИ

7.1. Листа на објавени трудови во областа во кои кандидатот е (ко)автор

[A1] **Мирчева Г.**, "Систем за пребарување на 3Д протеински структури", Дипломска работа, Факултет за електротехника и информациски технологии (УКИМ), Скопје, Р. Македонија, 2007.

[A2] **Мирчева Г.**, "Компаративна анализа на постоечки и нови дескриптори и нивна примена за класификација на протеини", Магистерска работа, Факултет за електротехника и информациски технологии (УКИМ), Скопје, Р. Македонија, 2009.

[A3] **Mirceva G.**, Kalajdziski S., Trivodaliev K., Davcev D., "Protein Retrieval by Matching 3D Structures", 8th National Conference ETAI 2007, Ohrid, Macedonia, 2007, pn. I5-2.

[A4] Kalajdziski S., **Mirceva G.**, Trivodaliev K., Davcev D., "Protein Classification by Matching 3D Structures", IEEE Conference Frontiers in the Convergence of Bioscience and Information Technologies 2007 (FBIT '07), Jeju Island, Korea, 2007, pp. 147-152.

[A5] Trivodaliev K., Kalajdziski S., Kulakov A., Davcev D., **Mirceva G.**, "Efficient protein classification by using 3D structure content representation", IASTED International Conference on Artificial Intelligence and Soft Computing 2008 (ASC '08), Palma de Mallorca, Spain, 2008, pp. 151-156.

[A6] **Mirceva G.**, Kalajdziski S., Trivodaliev K., Davcev D., "Comparative Analysis of three efficient approaches for retrieving protein 3D structures", 4-th IEEE Cairo International Biomedical Engineering Conference 2008 (CIBEC '08), Cairo, Egypt, 2008, pp. 1-4.

[A7] **Mirceva G.**, Kulakov A., Davcev D., "SGNG Protein Classifier by matching 3D structures", In E. Corchado et al. (Eds.) 4-th International Conference on Hybrid Artificial Intelligence Systems (HAIS '09), Salamanca, Spain, 2009, LNAI 5572, pp. 425-432, Springer-Verlag Berlin Heidelberg 2009.

[A8] Dimov Z., **Mirceva G.**, Davcev D., "An Efficient Approach in Comparing Protein Structures using Matrix Alignment Techniques", International Conference on Bioinformatics, Computational Biology, Genomics and Chemoinformatics (BCBGC '09), Orlando, FL, USA, 2009, pp. 73-77.

[A9] **Mirceva G.**, Davcev D., "Novel protein classifier based on SVM", In Book of Abstracts of the international Conference on Computational and Systems Biology (iCCSB 2009), Shanghai, China, 2009, pn. OR-020.

[A10] Cingovska I., **Mirceva G.**, Kalajdziski S., Davcev D., "Comparative analysis of wavelet based protein descriptors", 9th National Conference ETAI 2009, Ohrid, Macedonia, 2009, pn. IE3-6.

[A11] **Mirceva G.**, Davcev D., "HMM Approach for Classifying Protein Structures", Future Generation Information Technology 2009, Jeju Island, Korea, Y.-h. Lee et al. (Eds.): FGIT 2009, LNCS 5899, pp. 34-41, 2009, Springer-Verlag Berlin Heidelberg 2009.

[A12] Dimov Z., **Mirceva G.**, Davcev D., "Protein distance matrices comparison using sequence alignment techniques", In CD Proceedings of ICT Innovations 2009, Ohrid, Macedonia, 2009.

[A13] Cingovska I., **Mirceva G.**, Dimov Z., Kalajdziski S., Davcev D., "Novel wavelet based protein descriptors", In CD Proceedings of ICT Innovations 2009, Ohrid, Macedonia, 2009.

[A14] **Mirceva G.**, Davcev D., "HMM based approach for classifying protein structures", International Journal of Bio-Science and Bio-Technology, vol. 1, no. 1, 2009, pp. 37-46.

[A15] **Mirceva G.**, Dimov Z., Kalajdziski S., Davcev D., "Protein Classification Based on 3D Structures and Fractal Features", ICT Innovations 2009, Springer-Verlag Berlin, Heidelberg, (Eds. D.Davcev and Jorge Marx Gomez), 2010, pp. 115-124.

[A16] Kalajdziski S., Pepik B., Ivanovska I., **Mirceva G.**, Trivodaliev K., Davcev D., "Automated Structural Classification of Proteins by Using Decision Trees and Structural Protein Features", ICT Innovations 2009, Springer-Verlag Berlin Heidelberg, (Eds. D. Davcev and Jorge Marx Gomez), 2010, pp. 135-144.

[A17] **Mirceva G.**, Davcev D., "Incorporating several features in the protein ray descriptor for more accurate protein 3D structure retrieval", ACM Multimedia 2010, 3D object retrieval, Florence, Italy, 2010, pp. 51-56.

[A18] Ivanoska I., **Mirceva G.**, Trivodaliev K., Kalajdziski S., "Hierarchical Protein Classification based on Gene Ontology and Decision Trees", ICT Innovations 2010, Ohrid, Macedonia, (Ed.) M. Gusev, Web proceedings, ISSN 1857-7288, pp.31-40.

[A19] **Mirceva G.**, Naumoski A., Davcev D., "A Protein Classifier Based on SVM by Using the Voxel Based Descriptor", Rough Sets and Current Trends in Computing - 7th International Conference (RSCTC 2010), M. Szczuka et al. (Eds.): RSCTC 2010, LNAI 6086, pp. 640-648, 2010, Springer-Verlag Berlin Heidelberg 2010.

[A20] **Mirceva G.** et al., "Hidden Markov Models for classifying protein secondary and tertiary structures", Journal of Convergence, vol. 1, no. 1, 2010, pp. 57-64.

[A21] Piskachev G., **Mirceva G.**, Davcev D., "Novel Consensus Approach for Protein Active Sites Detection", Sixth International Scientific Conference - Computer Science'2011, Ohrid, Macedonia, 2011, pp. 122-127.

[A22] **Mirceva G.**, Davcev D., "Three Approaches for Classifying Protein Tertiary Structures", Book chapter in Lauren M. Haggerty (Ed.): Protein Structure, Nova Publisher, pp. 51-69, 2011.

[A23] **Mirceva G.**, Davcev D., "SVM based approaches for classifying protein tertiary structures", International Conference on Data and Knowledge Engineering 2011 (ICDKE 2011), Milan, Italy, 2011, pp. 1-7.

[A24] Naumoski A., **Mirceva G.**, Mitreski K., "Fuzzy models with GIS for water quality diatom-indicator classification", 8th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2011), Shanghai, China, 2011, pp. 829-833.

[A25] **Mirceva G.**, Davcev D., "Incorporating the depth and protrusion index in the protein ray based descriptor", 8th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2011), Shanghai, China, 2011, pp. 1592-1596.

[A26] **Mirceva G.**, Davcev D., "Protein classification by including several features in the protein ray based descriptor", The 7-th Structural Bioinformatics and Computational Biophysics Meeting (3DSIG 2011), Vienna, Austria, 2011, pp. 94.

[A27] **Mirceva G.**, Naumoski A., Stojkovik V., Temelkovski D., Davcev D., "Method for Protein Active Sites Detection Based on Fuzzy Decision Trees", Database Theory and Application / Bio-Science and Bio-Technology, DTA/BSBT 2011, CCIS 258, pp. 143-150, 2011, Springer-Verlag Berlin Heidelberg 2011.

[A28] **Mirceva G.**, Kulakov A., "Fuzzy pattern trees for predicting the protein binding sites", The 9th Conference for Informatics and Information Technology (CIIT 2012), Bitola, Macedonia, 2012.

[A29] **Mirceva G.**, Naumoski A., Davcev D., "A Novel Fuzzy Decision Tree Based Method for Detecting Protein Active Sites", L. Kocarev (Ed.): ICT Innovations 2011, AISC 150, pp. 51-60, Springer-Verlag Berlin Heidelberg 2012.

[A30] Naumoski A., **Mirceva G.**, Mitreski K., "A novel fuzzy based approach for inducing diatom habitat models and discovering diatom indicating properties", Ecological informatics, vol. 7, no. 1, pp. 62-70, 2012, (**impact factor 1.961**).

[A31] **Mirceva G.**, Cingovska I., Dimov Z., Davcev D., "Efficient Approaches for Retrieving Protein Tertiary Structures", IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 9, no. 4, pp. 1166-1179, 2012, (**impact factor 1.616**).

[A32] **Mirceva G.**, Kulakov A., "Top-down approach for protein binding sites prediction based on fuzzy pattern trees", S. Markoski, M. Gusev (Eds.): ICT Innovations 2012, AISC 207, pp. 325-334, Springer-Verlag Berlin Heidelberg 2013.

[A33] **Mirceva G.**, Kulakov A., "Fuzzy pattern trees for protein binding sites prediction using weighted averaging fuzzy aggregation operators", The 10th Conference for Informatics and Information Technology (CIIT 2013), Bitola, Macedonia, 2013.

[A34] **Mirceva G.**, Kulakov A., "Protein Binding Sites Prediction Using Ensembles", ICT Innovations 2013, Ohrid, Macedonia, 2013.

[A35] **Mirceva G.**, Kulakov A., "Improvement of Protein Binding Sites Prediction by Selecting Amino Acid Residues' Features", Journal of Structural Biology, (under review).

[A36] **Mirceva G.**, Kalajdziski S., Kulakov A., "A method for predicting protein functions based on the protein ray-based descriptor", Journal of Bioinformatics, vol. 1, no. 1, 2014, pp. 26-33.

[A37] **Mirceva G.**, Kulakov A., "Annotating protein structures by using multi-label classifier", The 11th Conference for Informatics and Information Technology (CIIT 2014), Bitola, Macedonia, 2014.

[A38] **Mirceva G.**, Naumoski A., Kulakov A., "Protein function prediction by using binary relevance multi-labeling method", Barcelona GPCR Spring Conference 2014, Barcelona, Spain, 2014.

7.2. Листа на користени трудови во истражувањето

1. Berman H.M., Westbrook J., Gilliland Z. Feng G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E., "Protein Data Bank", Nucleic Acids Research, 2000, 28, 235-242.
2. Protein Data Bank, 1971, http://www.rcsb.org.
3. The Gene Ontology Consortium. The Gene Ontology project in 2008. Nucleic Acids Res., 2008, 36(Database issue), D440–D444.
4. Du Plessis L., Škunca N., Dessimoz C., "The what, where, how and why of gene ontology-a primer for bioinformaticians", Brief. Bioinform., 2011, 12(6), 723–735.
5. Casari G., Sander C., Valencia A., "A method to predict functional residues in proteins", Nat. Struct. Biol., 1995, 2, 171-178.
6. Andrade M.A., Casari G., Sander C., Valencia A., "Classification of protein families and detection of the determinant residues with an improved self-organizing map", Biol. Cybern., 1997, 76, 441-450.
7. Hannenhalli S.S., Russell R.B., "Analysis and prediction of functional sub-types from protein sequence alignments", J. Mol. Biol., 2000, 303, 61-76.
8. Li L., Shakhnovich E.I., Mirny L.A., "Amino acids determining enzyme-substrate specificity in prokaryotic and eukaryotic protein kinases", Proc. Natl. Acad. Sci., 2003, 100, 4463-4468.
9. Altschul, S., Gish, W., Miller, W, Myers, E.W., Lipman, D., "Basic local alignment search tool", J. Mol. Biol., 1990, 215(3), 403–410.
10. Hulo N., Sigrist C.J.A., Le Saux V., et al., "Recent improvements to the PROSITE database", Nucleic Acids Research, 2004, 32, D134-137.
11. Sigrist, C.J.A., Cerutti, L., De Castro, E., Langendijk-Genevaux, P.S., Bulliard, V., Bairoch, A., Hulo, N., "PROSITE, a protein domain database for functional characterization and annotation", Nucleic Acids Res., 2010, 38(Database issue), D161–D166.
12. Bateman A., Coin L., Durbin R., et al., "The Pfam protein families database", Nucleic Acids Research, 2004, 32, D138-141.
13. Attwood T.K., Bradley P., Flower D.R., et al., "PRINTS and its automatic supplement, prePRINTS", Nucleic Acids Research, 2003, 31, 400-402.
14. Letunic I., Copley R.R., Schmidt S., et al., "SMART 4.0: towards genomic data integration", Nucleic Acids Research, 2004, 32, D142-144.
15. Marchler-Bauer A., Panchenko A.R., Shoemaker B.A., et al., "CDD: a database of conserved domain alignments with links to domain three-dimensional structure", Nucleic Acids Research, 2002, 30, 281-283.
16. Servant F., Bru C., Carrere S., et al., "ProDom: automated clustering of homologous domains", Briefings in Bioinformatics, 2002, 3, 246-251.
17. Rost B., "Protein structures sustain evolutionary drift", Fold. Des., 1997, 2, S19-24.
18. Taylor W. R., Orengo C. A., "Protein structure alignment", J. Mol. Biol., 1989, 208, 1-22.
19. Holm L, Sander C., "Protein structure comparison by alignment of distance matrices", J. Mol. Biol., 1993, 233, 123-138.
20. Shindyalov H.N., Bourne P.E.,, "Protein structure alignment by incremental combinatorial extension (ce) of the optimal path", Protein Eng., 1998, 9, 739-747.
21. Ortiz, A.R., Strauss C.E., Olmea O., "Mammoth (matching molecular models obtained from theory): An automated method for model comparison", Protein Science, 2002, 11, 2606-21.
22. Lupyan D., Leo-Macias A., Ortiz A.R., "A new progressive-iterative algorithm for multiple structure alignment", Bioinformatics 2005, 21(15), 3255-3263.

23. Leibowitz N., Fligelman Z.Y., Nussinov R., Wolfson H.J, "Automated multiple structure alignment and detection of a common substructure motif", Proteins, 2001, 43(3), 235-245.

24. Panchenko A.R., Kondrashov F., Bryant S., "Prediction of functional sites by analysis of sequence and structure conservation", Protein Science, 2004, 13, 884-892.

25. De S., Krishnadev O., Srinivasan N., et al., "Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different", BMC Struct. Biol., 2005, 5, 15.

26. Jones S., Marin A., Thornton J.M., "Protein domain interfaces: characterization and comparison with oligomeric protein interfaces", Protein Eng., 2000, 13, 77-82.

27. Jones S., Thornton J.M., "Principles of protein-protein interactions", Proc. Natl. Acad. Sci. USA, 1996, 93, 13-20.

28. Jones S., Thornton J.M., "Analysis of protein-protein interaction sites using surface patches", J. Mol. Biol., 1997, 272, 121-32.

29. Gong S., Yoon G., Jang I., et al., "PSIbase: a database of protein structural interactome map (PSIMAP)", Bioinformatics, 2005, 21, 2541-2543.

30. Keskin O., Tsai C.J., Wolfson H., et al., "A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications", Protein Science, 2004,13, 1043-1055.

31. Tsai C.J., Lin S.L., Wolfson H.J., et al., "A dataset of protein-protein interfaces generated with a sequence-orderindependent comparison technique", J. Mol. Biol., 1996, 260, 604-620.

32. Lawrence M.C., Colman P.M., "Shape complementarity at protein/protein interfaces", J. Mol. Biol., 1993, 234, 946-950.

33. Lo Conte L., Chothia C., Janin J., "The atomic structure of protein-protein recognition sites", J. Mol. Biol., 1999, 285, 2177-2198.

34. Sheinerman F.B., Norel R., Honig B., "Electrostatic aspects of protein-protein interactions", Curr. Opin. Struct. Biol., 2000, 10, 153-159.

35. Xu D., Lin S.L., Nussinov R., "Protein binding versus protein folding: the role of hydrophilic bridges in protein associations", J. Mol. Biol., 1997, 265, 68-84.

36. Tuncbag, N., Kar, G., Keskin, O., Gursoy, A., Nussinov, R. "A survey of available tools and web servers for analysis of protein-protein interactions and interfaces", Brief. Bioinform., 2009, 10(3), 217–232.

37. Ezkurdia L., Bartoli L., Fariselli P., Casadio R., Valencia A., Tress M.L., "Progress and challenges in predicting protein-protein interaction sites", Briefings in Bioinformatics, April 2009, 10(3), 233-246.

38. Sharan et al., "Network-based prediction of protein function", Molecular System Biology, 2007, 3, 88.

39. Schwikowski et al., "A network of protein-protein interactions in yeast", Nat. Biotechnol., 2000, 18, 1257-1261.

40. Hishigaki et al., "Assessment of prediction accuracy of protein function from protein-protein interaction data", Yeast, 2001, 18, 523-531.

41. Vazquez et al., "Global protein function prediction from protein-protein interaction networks", Nat. Biotechnol., 2003, 21, 697-700.

42. Deng et al., "Mapping Gene Ontology to proteins based on protein-protein interaction data", Bioinformatics, 2004, 20, 895-902.

43. Letovsky S., Kasif S., "Predicting protein function from protein/protein interaction data: a probabilistic approach", Bioinformatics, 2003, 19, 197-204.

44. Nabieva et al., "Whole-proteome prediction of protein function via graph theoretic analysis of interaction maps", Bioinformatics, 2005, 21, i302-i310.

45. Brohée S., van Helden J., "Evaluation of clustering algorithms for protein protein interaction networks", BMC Bioinformatics, 2006, 7, 488.

46. Kirac M., Ozsoyoglu G., “Protein Function Prediction based on Patterns in Biological Networks”, RECOMB 08, 2008.

47. Kirac M., Ozsoyoglu G., Yang J., “Annotating proteins by mining protein interaction networks”, Bioinformatics, 2006, 22(14), e260-e270.

48. Bader G.D., Donaldson I., Wolting C., Ouellette B.F.F., Pawson T., Hogue C.W.V., “BIND: the Biomolecular Interaction Network Database”, Nucleic Acids Res., 2001, 29(1), 242-245.

49. Xenarios I., Salwínski L., Duan X.J., Higney P., Kim S.M., Eisenberg D., “DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions”, Nucleic Acids Res., 2002, 30(1), 303-305.

50. Güldener U., Münsterkötter M., Oesterheld M., Pagel P., Ruepp A., Mewes H.W., Stümpflen V., “MPact: the MIPS protein interaction resource on yeast”, Nucleic Acids Res., 2006, 34(Database issue), D436-41.

51. Chatr-aryamontri A., Ceol A., Palazzi L.M., Nardelli G., Schneider M.V., Castagnoli L., Cesareni G., “MINT: the Molecular INTeraction database”, Nucleic Acids Res., 2007, 35, D572-D574.

52. Breitkreutz B.J., Stark C., Tyers M., “The GRID: The General Repository for Interaction Datasets”, Genome Biology, 2003, 4(3), R23.

53. Benson A.D., Karsch-Mizrachi I., Lipman J.D., Ostell J., Wheeler L.D., “GenBank”, Nucleic Acids Research, 2006, (34), D16-D20.

54. Kulikova T., et al., “EMBL Nucleotide Sequence Database in 2006”, Nucleic Acids Research, 2007, 35, D16-D20.

55. DNA Databank of Japan (DDBJ), 1986, http://www.ddbj.nig.ac.jp.

56. Bairoch A., Apweiler R., “The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000”, Nucleic Acids Research, 2000, 28, 45-48.

57. Barker C.W. et al., “The Protein Information Resource (PIR)”, Nucleic Acids Research, 2000, 28, 41-44.

58. Bairoch A., “The ENZYME Database in 2000”, Nucleic Acids Research, 2000, 28, 304-305.

59. Wang Y., Geer L., Madej T., Marchler-Bauer A., Zimmerman D., Bryant S., “MMDB: 3D Structure Data in Entrez”, Nucleic Acids Research, 2002, 30, 249-252.

60. Attwood T.K., Beck M.E., Bleasby A.J., Degtyarenko K., Parry Smith D.J., “Progress with the PRINTS Protein Fingerprint Database”, Nucleic Acids Research, 1996, 24, 182-188.

61. Henikoff G.J., Greene A.E., Pietrokovski S., Henikoff S., “Increased Coverage of Protein Families with the BLOCKS Database Servers”, Nucleic Acids Research, 2000, 28, 228-230.

62. Aung Z., Tan K.L, “MatAlign: Precise Protein Structure Comparison by Matrix Alignment”, Journal of Bioinformatics and Computational Biology, 2006, 4, 1197-1216.

63. Pearson W.R., Lipman D.J., “Improved tools for biological sequence comparison”, In: Proc. of Natl. Acad. Sci. USA, 85 (8), 2444-2448, 1988.

64. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J., “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs”, Nucleic Acids Research, 1997, 25(17), 3389-3402.

65. Zhang Z., Schäffer A.A., Miller W., Madden T.L., Lipman D.J., Koonin E.V., Altschul S.F., “Protein sequence similarity searches using patterns as seeds”, Nucleic Acids Research, 1998, 26 (17), 3986-3990.

66. Lindahl E., Elofsson A., “Identification of related proteins on family, superfamily and fold level”, Journal of Molecular Biology, 2000, 295(3), 613-625.

67. Bellman R., “Dynamic Programming”, Princeton University Press, 1957.

68. Taylor W.R., Flores T.P., “Multiple protein structure alignment”, Orengo, CA, Protein Science, 1994, 3(10), 1858-1870.

69. Orengo C.A., Michie A.D., Jones S., Jones D.T., Swindells M.B., Thornton J.M., “CATH: A hierarchical classification of protein domain structures”, Structure, 1997, 5(8), 1093-1108.

70. Aung Z., Tan K.L., "MatAlign: Precise Protein Structure Comparison by Matrix Alignment", J. Bioinformatics and Computational Biology, 2006, 4(6), 1197-1216.

71. Pelta D.A., González J.R., Vega M.M., "A simple and fast heuristic for protein structure comparison", BMC Bioinformatics, 2008, 9(161).

72. Cui C., Wang D., Shi J., "Comparing 3-D Protein Structures Similarity by Using Fractal Features", Proc. 2004 IEEE Computational Systems Bioinformatics Conference (CSB'04), 698-699, 2004.

73. K. Marsolo, P. Srinivasan and K. Ramamohanarao, "Structure-Based Querying of Proteins Using Wavelets", Proc. ACM Fifteenth Conference on Information and Knowledge Management (CIKM2006), Arlington, Virginia, USA, 24-33, 2006.

74. Ye Y., Godzik A., "Flexible structure alignment by chaining aligned fragment pairs allowing twists", Bioinformatics, 2003, 19(2), II246-255.

75. Madej T., Gibrat J.F., Bryant S.H., "Threading a database of protein cores", Proteins, 1995, 23, 356-369.

76. Zhu J., Weng Z., "FAST: a novel protein structure alignment algorithm", Proteins, 2005, 58, 618-627.

77. Kawabata T., "MATRAS: a program for protein 3D structure comparison", Nucleic Acids Research, 2003, 31, 3367-3369.

78. Holm L., Park J., "DaliLite workbench for protein structure comparison", Bioinformatics, 2000, 16, 566-567.

79. Harrison A., Pearl F., Sillitoe I., et al., "Recognizing the fold of a protein structure", Bioinformatics, 2003, 19, 1748-1759.

80. Larsen T.A., Olson A.J., Goodsell D.S., "Morphology of protein-protein interfaces", Structure, 1998, 6, 421-427.

81. Ofran Y., Rost B., "Analysing six types of protein-protein interfaces", J. Mol, Biol., 2003, 325, 377-387.

82. Greer J., Bush B.L., "Macromolecular shape and surface maps by solvent exclusion", Proc. Natl. Acad. Sci. USA, 1978, 75, 303-307.

83. Wodak S.J., Janin J., "Analytical approximation to the accessible surface area of proteins", Proc. Natl. Acad. Sci. USA, 1980, 77, 1736-1740.

84. Kuntz I.D., Blaney J.M., Oatley S.J., et al., "A geometric approach to macromolecule-ligand interactions", J. Mol. Biol., 1982, 161, 269-288.

85. Lee R.H., Rose G.D., "Molecular recognition: Automatic identification of topographic surface features", Biopolymers, 1985, 24, 1613-1627.

86. Connolly M.L., "Solvent-accessible surfaces of proteins and nucleic acids", Science, 1983, 221, 709-713.

87. Jiang F., Kim S., "Soft docking: matching of molecular surface cubes", J. Mol. Biol., 1991, 219, 79-102.

88. Helmer-Citterich M., Tramontano A., "PUZZLE: a new method for automated protein docking based on surface shape complementarity", J. Mol. Biol., 1994, 235, 1021-1031.

89. Salemme F.R., "An hypothetical structure for an intermolecular electron transfer complex of cytochromes c and b5", J. Mol. Biol., 1976, 102, 563-568.

90. Warwicker J., "Investigating protein-protein interaction surfaces using a reduced stereochemical and electrostatic model", J. Mol. Biol., 1989, 206, 381-395.

91. Walls P.H., Sternberg M.J.E., "New algorithm to model protein-protein recognition based on surface complementarity: applications to antibody-antigen docking", J. Mol. Biol., 1992, 228, 277-297.

92. Shoichet B.K., Kuntz I.D., "Matching chemistry and shape in molecular docking", Protein Eng., 1993, 6, 723-732.

93. Vakser I.A., Aflalo C., "Hydrophobic docking: a proposed enhancement to molecular recognition techniques", Proteins, 1994, 20, 320-329.

94. Nooren I.M.A., Thornton J.M., "Structural characterisation and functional significance of transient protein-protein Interactions", J. Mol. Biol., 2003, 325, 991-1018.

95. Ma B., Elkayam T., Wolfson H., et al., "Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces", Proc. of Natl. Acad. Sci. USA, 2003, 100.

96. Chakrabarti P., Janin J., "Dissecting protein-protein recognition sites", Proteins, 2002, 47, 334-343.

97. Bahadur R.P., Chakrabarti P., Rodier F., et al., "Dissecting subunit interfaces in homodimeric proteins", Proteins, 2003, 53, 708-719.

98. Valdar W.S., Thornton J.M., "Protein-protein interfaces: analysis of amino acid conservation in homodimers", Proteins, 2001, 42, 108-124.

99. Caffrey D.R., Somaroo S., Hughes J.D., et al., "Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?", Protein Science, 2004, 13, 190-202.

100. Tseng Y.Y., Liang J., "Estimation of amino acid residue substitution rates at local spatial regions and application in protein function inference: a Bayesian Monte Carlo approach", Mol. Biol. Evol., 2006, 23, 421-436.

101. Valdar W.S., "Scoring residue conservation", Proteins, 2002, 48, 227-241.

102. Glaser F., Pupko T., Paz I., et al., "ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information", Bioinformatics, 2003, 19, 163-164.

103. Glaser F., Rosenberg Y., Kessel A., et al., "The ConSurf-HSSP database: the mapping of evolutionary conservation among homologs onto PDB structures", Proteins, 2005, 58, 610-617.

104. Thorn K.S., Bogan A.A., "ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions, "Bioinformatics, 2001, 17, 284-285.

105. Fischer T.B., Arunachalam K.V., Bailey D., et al., "The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces", Bioinformatics, 2003, 19, 1453-1454.

106. DeLano W.L., "Unraveling hot spots in binding interfaces: progress and challenges", Curr. Opin. Struct. Biol., 2002, 12, 14-20.

107. Kortemme T., Kim D.E., Baker D., "Computational alanine scanning of protein-protein interfaces", Sci. STKE 2004, 2004(219), pl2.

108. Schymkowitz J., Borg J., Stricher F., et al., "The FoldX web server: an online force field", Nucleic Acids Research, 33(Web Server issue), 2005, W382-388.

109. Gonzalez-Ruiz D., Gohlke H., "Targeting protein-protein interactions with small molecules: challenges and perspectives for computational binding epitope detection and ligand finding", Curr. Med. Chem., 2006, 13, 2607-2625.

110. Huo S., Massova I., Kollman P.A., "Computational alanine scanning of the 1:1 human growth hormone-receptor complex", J. Comput. Chem., 2002, 23, 15-27.

111. Rajamani D., Thiel S., Vajda S., et al., "Anchor residues in protein-protein interactions", Proc. Natl. Acad. Sci., 101, USA, 11287-11292, 2004.

112. Hu Z., Ma B., Wolfson H., et al., "Conservation of polar residues as hot spots at protein interfaces", Proteins, 2000, 39, 331-342.

113. Keskin O., Ma B., Nussinov R., "Hot regions in protein-protein interactions: the organization and contribution of structurally conserved hot spot residues", J. Mol. Biol., 2005, 345, 1281-1294.

114. Ma B., Elkayam T., Wolfson H., et al., "Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces", Proc. Natl. Acad. Sci. USA, 100, 2003, 5772-5777.

115. Guney E., Tuncbag N., Keskin O., et al., "HotSprint: database of computational hot spots in protein interfaces", Nucleic Acids Research, 36(Database issue), 2008, D662-666.

116. Darnell S.J., LeGault L., Mitchell J.C., "KFC Server: interactive forecasting of protein interaction hot spots", Nucleic Acids Research, 2008, 36(Web server issue), 265-269.

117. Henrick K., Thornton J.M., "PQS: a protein quaternary structure file server", Trends Biochem. Sci., 1998, 23, 358-361.

118. Valdar W.S., Thornton J.M., "Conservation helps to identify biologically relevant crystal contacts", J. Mol. Biol., 2001, 313, 399-416.

119. Carugo O., Argos P., "Protein-protein crystal-packing contacts", Protein Science, 1997, 6, 2261-2263.

120. Janin J., Rodier F., "Protein-protein interaction at crystal contacts", Proteins, 1995, 23, 580-587.

121. Zhu H., Domingues F.S., Sommer I., et al., "NOXclass: prediction of protein-protein interaction types", BMC Bioinformatics, 2006, 7, 27.

122. Bernauer J., Bahadur R.P., Rodier F., et al., "DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions", Bioinformatics, 2008, 24, 652-658.

123. Lu Y., Wang X., Chen X., Zhao G., "Computational Methods for DNA-binding Protein and Binding Residue Prediction", Protein Peptide Lett., 2013, 20(3), 346–351.

124. Neuvirth H., Raz R., Schreiber G., "ProMate: a structure based prediction program to identify the location of protein-protein binding sites", J. Mol. Biol., 2004, 338, 181-199.

125. Bradford J.R., Westhead D.R., "Improved prediction of protein-protein binding sites using a support vector machines approach", Bioinformatics, 2005, 21, 487-1494.

126. Murakami Y., Jones S., "SHARP2: protein-protein interaction predictions using patch analysis", Bioinformatics, 2006, 22, 1794-1795.

127. Porollo A., Meller J., "Prediction-based fingerprints of protein-protein interactions", Proteins, 2007, 66, 630-645.

128. Zhou H.X., Shan Y., "Prediction of protein interaction sites from sequence profile and residue neighbor list", Proteins, 44, 2001, 336-343.

129. Aytuna A.S., Gursoy A., Keskin O. "Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces", Bioinformatics, 2005, 21(12), 2850-2855.

130. Ogmen U., Keskin O., Aytuna A.S., et al., "PRISM: protein interactions by structural matching", Nucleic Acids Research, 2005, 33(Web Server issue), W331-336.

131. Keskin O., Nussinov R., Gursoy A., "PRISM: protein-protein interaction prediction by structural matching", Methods Mol. Biol., 2008, 484, 505–521.

132. Tuncbag N., Gursoy A., Guney E., et al., "Architectures and functional coverage of protein-protein interfaces", J. Mol. Biol., 2008, 381, 785-802.

133. Davis F.P., Sali A., "PIBASE: a comprehensive database of structurally defined protein interfaces", Bioin¬formatics, 2005, 21, 1901-1907.

134. Gong S., Park C., Choi H., et. al., "A protein domain interaction interface database: InterPare", BMC Bioinformatics, 2005, 6, 207.

135. Teyra J., Doms A., Schroeder M., et al., "SCOWLP: a web-based database for detailed characterization and visualization of protein interfaces", BMC Bioinformatics, 2006, 7, 104.

136. Winter C., Henschel A., Kim W.K., et al., "SCOPPI: a structural classification of protein-protein interfaces", Nucleic Acids Research, 2006, 34 (Database issue), D310-314.

137. Stein A., Russell R.B., Aloy P., "3did: interacting protein domains of known three-dimensional structure", Nucleic Acids Research, 2005, 33 (Database issue), D413-417.

138. Vranic D.V., "3D Model Retrieval", PhD dissertation, Dept. of Computer Science, University of Leipzig, Leipzig, 2004.

139. Daras P., Zarpalas D., Axenopoulos A., Tzovaras D., Strintzis M.G., "Three-Dimensional Shape-Structure Comparison Method for Protein Classification", IEEE/ACM Trans. Computational Biology and Bioinformatics, 2006, 3(3), 193-207.

140. Junichi H., Hideaki U., “Protein dynamics determined by backbone conformation and atom packing”, Protein Engineering, 1997, 10(4), 373-380.

141. Gonzales R.C., Woods R.E., “Digital Image Processing”, 2nd edition, Prentice Hall, New Jersey, 2002, 349-404.

142. Misiti M., Misiti Y., Oppenheim G., Poggi J., “Wavelet Toolbox™ 4 User’s Guide”, The MathWorks Inc., 2008.

143. Murzin A.G., Brenner G.E., Hubbard T., Chothia C., “Scop: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures”, J. Molecular Biology, 1995, 247(4), 1995, 536-540.

144. SCOP Database, http://scop.mrc-lmb.cam.ac.uk/scop, Accessed 21.02.2014.

145. Chandonia J.M., Hon G., Walker N.S., Lo Conte L., Koehl P., Levitt M., Brenner S.E., “The ASTRAL Compendium in 2004”, Nucleic Acids Research, 2004, 32, D189-192.

146. Lee B., Richards F.M., “The interpretation of protein structures: Estimation of static accessibility”, Journal of Molecular Biology, 1971, 55(3), 379–400.

147. Shrake A., Rupley J.A., “Environment and exposure to solvent of protein atoms. Lysozyme and insulin”, Journal of Molecular Biology, 1973, 79(2), 351–371.

148. Hubbard S.J., Thornton J.M., “NACCESS, Computer Program”, Department of Biochemistry and Molecular Biology, University College London, London, UK, 1993.

149. Kyte J., Doolittle R.E., “A simple method for displaying the hydropathic character of a protein”, Journal of Molecular Biology, 1982, 157(1), 105–132.

150. Pintar A., Carugo O., Pongor S., “DPX: for the analysis of the protein core”, Bioinformatics, 2003, 19(2), 313–314.

151. Pintar A., Carugo O., Pongor S., “CX, an algorithm that identifies protruding atoms in proteins”, Bioinformatics, 2002, 18(7), 980–984.

152. Mihel J., Šikić M., Tomić S., Jeren B., Vlahoviček K., “PSAIA – Protein Structure and Interaction Analyzer”, BMC Structural Biology, 2008, 8:21.

153. Quinlan, R., “C4.5: Programs for Machine Learning”, 1st ed. и Morgan Kaufmann Publishers: San Mateo, CA, USA, 1993.

154. John G.H., Langley P., “Estimating Continuous Distributions in Bayesian Classifiers”, In: Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence, Besnard, P., Hanks, S., Eds., 338–345, 1995.

155. Aha D.W., Kibler D., Albert M.K., “Instance-based learning algorithms”, Machine Learning, 1991, 6, 37–66.

156. Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I.H., “The WEKA Data Mining Software: An Update”, SIGKDD Explorations, 2009, 11(1), 10–18.

157. Chothia C., ”The Nature of the Accessible and Buried Surfaces in Proteins ”, J. Mol. Biol., 1976, 105(1), 1–12.

158. Mohamed W.N.H.W., Salleh M.N.M., Omar A.H., “A comparative study of Reduced Error Pruning method in decision tree algorithms”, 2012 IEEE International Conference on Control System, Computing and Engineering, Penang, Malaysia, 392 – 397, 2012.

159. Freund Y., Mason L., “The alternating decision tree learning algorithm”, In: Proceedings of the 7th International Conference on Machine Learning (ICML 1999), Bratko, I., Dzeroski, S., Eds., Morgan Kaufmann: San Francisco, CA, USA, 124–133, 1999.

160. Holmes G., Pfahringer B., Kirkby R., Frank E., Hall M., “Multiclass alternating decision trees”, In: 13th European Conference on Machine Learning, 161–172, 2001.

161. Gama J., “Functional Trees”, Machine Learning, 2004, 55(3), 219–250.

162. Breiman L., ”Random Forests”, Machine Learning, 2001, 45(1), 5-32.

163. Kohavi R., “Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid”, In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Simoudis, E., Han, J., Fayyad, U., Eds., AAAI Press, pp. 202–207, 1996.

164. Friedman N., Geiger D., Goldszmidt M., “Bayesian Network Classifiers”, Machine Learning, 1997, 29(2-3), 131–163.

165. Rosenblatt F., “Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms”, Spartan Books, Washington DC, 1961.

166. Breiman L., “Bagging predictors”, Machine Learning, 1996, 24(2), 123–140.

167. Freund Y., Schapire R.E., “Experiments with a new boosting algorithm”, In: Thirteenth International Conference on Machine Learning, San Francisco, pp. 148-156, 1996.

168. Janikow C.Z., “Fuzzy decision trees: issues and methods”, IEEE Transactions on Systems, Man, and Cybernetics, 1998, 28(1), 1–14.

169. Wang L.X., Mendel J.M., “Generating fuzzy rules by learning from examples”, IEEE Transactions on Systems, Man, and Cybernetics, 1992, 22(6), 1414–1427.

170. Olaru C., Wehenkel L., “A complete fuzzy decision tree technique”, Fuzzy Sets and Systems, 2003, 138(2), 221–254.

171. Suárez A., Lutsko J.F., ”Globally optimal fuzzy decision trees for classification and regression”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1999, 21(12), 1297–1311.

172. Wang X., Chen B., Olan G., Ye F., “On the optimization of fuzzy decision trees”, Fuzzy Sets and Systems, 2000, 112(1), 117–125.

173. Chen Y.-L., Wang T., Wang B.-S., Li Z.-J., “A Survey of Fuzzy Decision Tree Classifier”, Fuzzy Information and Engineering, 2009, 1(2), 149–159.

174. Huang Z.H., Gedeon T.D., Nikravesh M., “Pattern trees induction: a new machine learning method”, IEEE Transaction on Fuzzy Systems, 2008, 16(3), 958–970.

175. Senge R., Hüllermeier E., “Top-Down Induction of Fuzzy Pattern Trees”, IEEE Transactions on fuzzy systems, 2011, 19(2), 241–252.

176. Zahn C.T., “Graph-theoretical methods for detecting and describing gestalt clusters”, IEEE transaction Computer, 1971, C-20(1), 68-86.

177. Zimmermann H.J., “Fuzzy Set Theory and its Applications”, Springer, 4th edition, 2005.

178. Zadeh L.A., “Fuzzy Sets”, Information and Control, 1965, 8, 338-353.

179. Schweizer B., Sklar A., “Associative functions and abstract semigroups”, Publ. Math. Debrecen, 1963, 10, 69-81.

180. Yager R.R., “On ordered weighted averaging aggregation operators in multicriteria decision making”, IEEE Transactions on Systems, Man, and Cybernetics, 1988, 18, 183-190.

181. Chiclana F., Herrera F., Herrera-Viedma E., “The ordered weighted geometric operator: Properties and application”, In: Proc 8th Int Conf on Information Processing and Management of Uncertainty in Knowledge-based Systems, 2000, 985–991.

182. Huang Z.H., Gedeon D.T., Nikravesh M., “Pattern Trees Induction: A New Machine Learning Method”, IEEE Transaction on Fuzzy Systems, 2008, 16(3), 958-970.

183. Pearson K., “On Lines and Planes of Closest Fit to Systems of Points in Space”, Philos. Mag., 1901, 2(11), 559–572.

184. Abdi H., Williams L.J., ”Principal component analysis”, Wiley Interdiscip. Rev. Comput. Stat., 2010, 2(4), 433–459.

185. Kohavi R., John G.H., “Wrappers for Feature Subset Selection”, Artif. Intell., 1997, 97(1), 273–324.

186. Liu H., Setiono R., “Chi2: Feature Selection and Discretization of Numeric Attributes”, In: Proceedings of the 7th IEEE International Conference on Tools with Artificial Intelligence, 388–391, 1995.

187. Fayyad U.M., Irani K.B., "Multi-interval discretization of continuous-valued attributes for classification learning", In: Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI 1993), 1022–1027, 1993.

188. Hunt E.B., Martin J., Stone P., "Experiments in Induction", 1st ed., Academic Press: New York, USA, 1966.

189. Kira K., Rendell L.A., "The feature selection problem: traditional methods and a new algorithm", In: Proceedings of the 10th national conference on Artificial intelligence (AAAI 1992), 129–134, 1992.

190. Kononenko I., "Estimating Attributes: Analysis and Extensions of RELIEF", In: Proceedings of the European Conference on Machine Learning (ECML 1994), 171–182, 1994.

191. Hall M.A., "Correlation-based Feature Selection for Machine Learning", PhD Thesis, University of Waikato, Hamilton, New Zealand, 1999.

192. Pearl J., "Heuristics: Intelligent Search Strategies for Computer Problem Solving", 1st ed. ,Addison-Wesley, Boston, MA, USA, 1984.

193. Goldberg D.E., "Genetic algorithms in search, optimization and machine learning", 1st ed., Addison-Wesley, Boston, MA, USA, 1989.

194. Peng H.C., Long F., Ding C., "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy", IEEE T. Pattern. Anal., 2005, 27(8), 1226–1238.

195. Ding C., Peng H., "Minimum redundancy feature selection from microarray gene expression data", J. Bioinform. Comput. Biol., 2005, 3(2), 185–205.

196. Guyon I., Weston J., Barnhill S., Vapnik V., "Gene selection for cancer classification using support vector machines", Mach. Learn., 2002, 46(3-1), 389–422.

197. Minimum Redundancy Maximum Relevance Feature Selection (mRMR), http://penglab.janelia.org/proj/mRMR/ (Accessed March 12, 2013).

198. Ofran Y., Rost B., "Predicted protein-protein interaction sites from local sequence information", FEBS Lett., 2003, 544(1-3), 236–239.

199. PRINT: Dataset of PRotein Protein INTerfaces. http://prism.ccbb.ku.edu.tr/interface (Accessed August 08, 2013).

200. Capra J.A., Singh M., "Predicting functionally important residues from sequence conservation", Bioinformatics, 2007, 23(15), 1875–1882.

201. Hendlich M., Rippmann F., Barnickel G., "LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins", J. Mol. Graph. Model., 1997, 15(6), 359–363.

202. An J., Totrov M., Abagyan R., "Pocketome via comprehensive identification and classification of ligand binding envelopes", Mol. Cell Proteomics, 2005, 4(6), 752–761.

203. Laskowski R., "SURFNET: a program for visualizing molecular surfaces, cavities and intermolecular interactions", J. Mol. Graph., 1995, 13(5), 323–330.

204. Capra J.A., Laskowski R.A., Thornton J.M., Singh M., Funkhouser, T.A., "Predicting Protein Ligand Binding Sites by Combining Evolutionary Sequence Conservation and 3D Structure", PLoS Comput. Biol., 5(12), 2009.

205. Dessailly, B.H., Lensink, M.F., Orengo, C.A., Wodak S.J., "LigASite a database of biologically relevant binding sites in proteins with known apo-structures", Nucleic Acids Res., 2008, 36 (Database issue), D667–D673.

206. García S., Fernández A., Luengo J., Herrera F., "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power", Inform. Sciences, 2010, 180(10), 2044-2064.

207. Tsoumakas G, Katakis I., Vlahavas I., "Mining multi-label data", Data Mining and Knowledge Discovery Handbook, 2nd ed., O. Maimon, L. Rokach, Eds. Springer, 2010, 667–685.

208. Hunter S., Jones P., Mitchell A. et al., “InterPro in 2011: New developments in the family and domain prediction database”, Nucleic Acids Res., 2012, 40, D306–D312.

209. UniProt Consortium, “Ongoing and future developments at the Universal Protein Resource”, Nucleic Acids Res., 2011, 39, D214–D219.

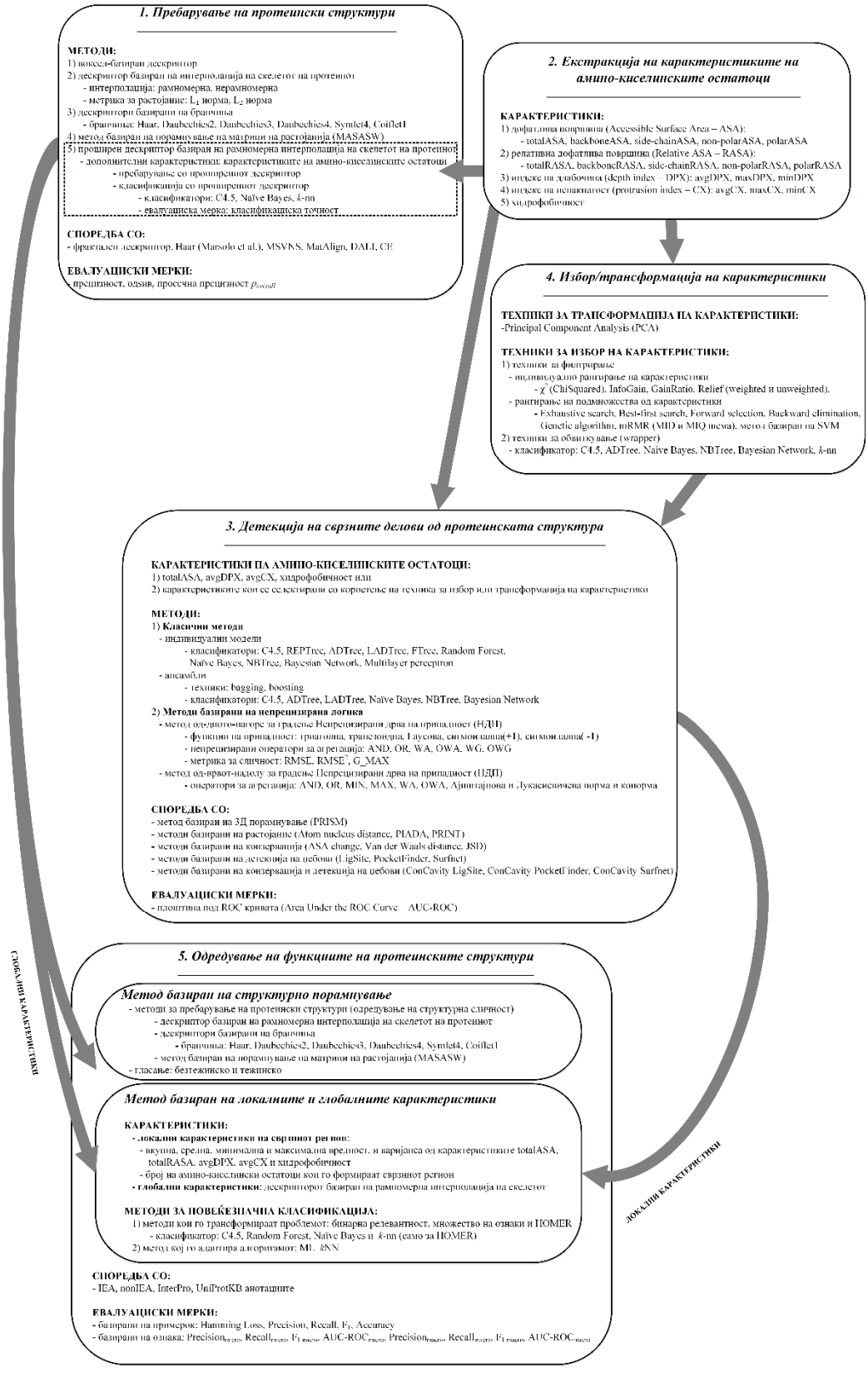
210. Luaces O., Díez J., Barranquero J., José del Coz J., Bahamonde A., “Binary relevance efficacy for multilabel classification”, Progress in Artificial Intelligence, 2012, 1(4), 303-313.

211. Read J., “A pruned problem transformation method for multi-label classification”, In: Proc. 2008 New Zealand Computer Science Research Student Conference (NZCSRS 2008), 2008, 143–150.

212. Tsoumakas G., Katakis I., Vlahavas I., “Effective and efficient multilabel classification in domains with large number of labels”, In: Proc. ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD’08), 2008, 30–44.

213. Zhang M.L., Zhou Z.H., “ML-KNN: A lazy learning approach to multi-label learning”, Pattern Recognition, 2007, 40, 2038–2048.

8. Додаток



Слика Д1 Детален преглед на методите кои се користат во ова истражување.

Се иницијализираат вредностите на сите вокселите на 0;
За секој триаголник T_j со јазли A_j, B_j и C_j

Ако A_j, B_j и C_j припаѓаат на истиот воксел, тогаш $p_j = 1$;
$d_{AB} = (B_j - A_j) / p_j$;
$d_{AC} = (C_j - A_j) / p_j$;
$d_G = (d_{AB} - d_{AC}) / 3$;
$\delta = S_j / (p_j^2 S)$;
за ($x = 0 , . . . , p_j - 2$)
за ($y = 0 , . . . , x$)
inc_att ($A_j + (x - y)d_{AB} + yd_{AC} + d_G , \delta$);
inc_att ($A_j + (x - y)d_{AB} + yd_{AC} + 2d_G , \delta$);
за ($y = 0 , . . . , p_j - 2$)
inc_att ($A_j + (p_j - 1 - y)d_{AB} + yd_{AC} + d_G , \delta$);

// процедурата *inc_att* ги определува индексите (*a,b,c*) на регионот на
// вокселот кој ја содржи точката $G = (g_x , g_y , g_z)$ и го инкрементира u_{abc} за δ
inc_att(G , δ)
$a = \lfloor (g_x - x_{min})/d_x \rfloor$, $b = \lfloor (g_y - y_{min})/d_y \rfloor$, $c = \lfloor (g_z - z_{min})/d_z \rfloor$;
$u_{abc} = u_{abc} + \delta$;

Алгоритам Д1 Апроксимација на вредноста на вокселите.

i=1; *scale*=1;
if $|\hat{f}_{0,0,0}|>0 \Rightarrow$ *scale*=$|\hat{f}_{0,0,0}|$;
for h=1,…k
$f_i = |\hat{f}_{h,0,0}| / scale$; i++;
$f_i = |\hat{f}_{0,h,0}| / scale$; i++;
$f_i = |\hat{f}_{0,0,h}| / scale$; i++;
for x=1,…,h-1
$f_i = |\hat{f}_{0,x,h-x}| / scale$; i++; $f_i = |\hat{f}_{0,x,x-h}| / scale$; i++;
$f_i = |\hat{f}_{x,0,h-x}| / scale$; i++; $f_i = |\hat{f}_{x,0,x-h}| / scale$; i++;
$f_i = |\hat{f}_{x,h-x,0}| / scale$; i++; $f_i = |\hat{f}_{x,x-h,0}| / scale$; i++;
for x=1,…,h-2
for y=1,…,h-1-x
$f_i = |\hat{f}_{x,y,h-x-y}| / scale$; i++; $f_i = |\hat{f}_{x,y,x+y-h}| / scale$; i++;
$f_i = |\hat{f}_{x,-y,h-x-y}| / scale$; i++; $f_i = |\hat{f}_{x,-y,x+y-h}| / scale$; i++;

Алгоритам Д2 Екстракција на протеинскиот воксел-базиран дескриптор.

Карактеристика	ASA					RASA					DPX		CX			хидрофобичност
	total	main-chain	side-chain	polar	non-polar	total	main-chain	side-chain	polar	non-polar	avg	max	avg	max	min	
ChiSquared	1		1		1	1		1					1	1	1	
InfoGain	1		1		1	1		1					1	1	1	
GainRatio	1		1		1	1		1					1	1	1	
Relief unweighted					1			1			1	1	1	1	1	1
Relief weighted							1		1		1	1	1	1	1	1
Exhaustive search	1		1		1	1							1	1	1	1
Forward selection	1		1		1	1							1	1	1	1
Backward elimination	1		1		1	1							1	1	1	1
Best-first search	1		1		1	1							1	1	1	1
Genetic search	1		1		1	1							1	1	1	1
mRMR (MID)		1			1				1	1		1		1	1	1
mRMR (MIQ)	1	1				1		1	1	1		1	1			
SVM	1	1			1	1	1			1				1	1	
Wrapper_C4.5	1	1				1	1			1		1	1			1
Wrapper_ADree		1		1		1	1		1	1	1	1				
Wrapper_NB	1		1		1	1		1			1				1	1
Wrapper_NBTree		1		1		1	1					1	1		1	1
Wrapper_BayesNet		1			1		1		1				1	1	1	1
Wrapper_KNN					1		1		1			1	1	1	1	1

За техниките кои вршат рангирање на карактеристики, прикажани се 8-те најдобро рангирани карактеристики.

Табела Д1 Множества од 8-те најрелевантни карактеристики селектирани со различна техника за избор на карактеристики.

Класификатор	C4.5	NB	NBTree	ADTree	FTree	BayesNet	Од-дното-нагоре НДП триаголна	Од-дното-нагоре НДП трапезоидна	Од-дното-нагоре НДП Гаусова	Од-врвот-надолу НДП
Карактеристики	**Време за тренирање (секунди)**									
сите карактеристики	54	15	284	326	68	19	NA	NA	NA	NA
totalASA, avgDPX, avgCX, хидрофоб.	11	4	33	69	23	4	10	11	13	170
PCA	12	4	39	88	22	4	13	13	14	42
ChiSquared	11	4	34	88	25	4	13	17	13	78
InfoGain	11	4	34	89	25	4	12	17	16	71
GainRatio	10	4	33	72	20	4	13	16	14	39
Relief unweighted	12	4	33	71	25	4	13	15	10	257
Relief weighted	11	4	33	69	22	4	17	11	16	234
Exhaustive/Best-first/ Genetic	25	7	100	157	35	8	NA	NA	NA	NA
Forward selection/ Backward elimination	11	4	34	86	22	4	17	14	16	128
mRMR (MID)	11	4	32	71	23	4	12	16	15	127
mRMR (MIQ)	9	4	31	92	19	4	13	14	11	124
SVM	10	4	35	72	18	5	11	14	11	170
Wrapper_C4.5	11	4	35	69	21	4	14	15	14	40
Wrapper_ADTree	12	4	34	69	21	4	15	11	11	70
Wrapper_NB	9	4	34	72	21	4	17	12	14	189
Wrapper_NBTree	11	4	34	69	25	4	13	18	12	184
Wrapper_BayesNet	11	4	35	82	21	4	21	22	17	122
Wrapper_KNN	11	4	33	69	23	4	11	13	14	189
Карактеристики	**Време за тестирање (секунди)**									
сите карактеристики	13	19	18	14	630	18	NA	NA	NA	NA
totalASA, avgDPX, avgCX, хидрофоб.	5	8	15	10	134	9	138	139	139	7
PCA	4	9	8	10	133	8	146	153	145	4
ChiSquared	5	8	9	9	123	9	136	138	145	4
InfoGain	4	8	9	9	132	8	148	143	142	3
GainRatio	4	7	8	9	36	8	137	147	145	6
Relief unweighted	4	7	8	8	149	8	149	157	148	6
Relief weighted	5	7	8	9	34	8	130	145	145	5
Exhaustive/Best-first/ Genetic	8	12	15	11	396	12	NA	NA	NA	NA
Forward selection/ Backward elimination	5	8	8	9	97	8	141	138	133	6
mRMR (MID)	4	9	8	9	166	8	147	149	147	5
mRMR (MIQ)	5	10	8	9	113	8	141	143	144	3
SVM	5	8	8	9	103	8	138	140	127	5
Wrapper_C4.5	5	8	8	9	104	8	136	137	136	6
Wrapper_ADTree	4	8	8	9	94	8	139	141	139	3
Wrapper_NB	4	8	9	9	114	9	139	139	139	8
Wrapper_NBTree	4	7	9	9	161	8	148	146	145	3
Wrapper_BayesNet	5	9	7	9	128	9	146	145	148	5
Wrapper_KNN	5	7	8	8	149	8	142	138	138	5

Табела Д2 Време потребно за тренирање и тестирање на моделите.

Метод	Бинарна релевантност			Множество на ознаки		
	C4.5	Random Forest	Naïve Bayes	C4.5	Random Forest	Naïve Bayes
Евалуациска мерка	**локални карактеристики**					
Hamming Loss	0.007	0.006	0.116	0.012	0.011	0.011
Precision	0.085	0.067	0.017	0.061	0.084	0.065
Recall	0.034	0.023	0.307	0.063	0.088	0.068
F_1	0.041	0.030	0.032	0.055	0.076	0.059
Accuracy	0.028	0.021	0.017	0.039	0.056	0.042
$Precision_{micro}$	0.156	0.175	0.016	0.057	0.083	0.065
$Recall_{micro}$	0.033	0.025	0.321	0.066	0.091	0.074
$F_{1\ micro}$	0.054	0.043	0.031	0.061	0.087	0.069
$Precision_{macro}$	0.011	0.024	0.013	0.021	0.045	0.022
$Recall_{macro}$	0.004	0.009	0.238	0.019	0.050	0.035
$F_{1\ macro}$	0.005	0.012	0.020	0.016	0.040	0.022
$AUC\text{-}ROC_{micro}$	0.771	0.656	0.756	0.502	0.459	0.540
$AUC\text{-}ROC_{macro}$	0.504	0.541	0.619	0.482	0.483	0.516
Евалуациска мерка	**глобални карактеристики**					
Hamming Loss	0.009	0.005	0.015	0.011	0.010	0.007
Precision	0.153	0.157	0.102	0.113	0.184	0.126
Recall	0.122	0.093	0.182	0.118	0.186	0.095
F_1	0.119	0.105	0.109	0.108	0.176	0.100
Accuracy	0.088	0.091	0.071	0.092	0.159	0.088
$Precision_{micro}$	0.161	0.679	0.097	0.113	0.198	0.179
$Recall_{micro}$	0.142	0.105	0.193	0.130	0.219	0.081
$F_{1\ micro}$	0.151	0.182	0.129	0.121	0.208	0.111
$Precision_{macro}$	0.083	0.224	0.146	0.063	0.176	0.150
$Recall_{macro}$	0.091	0.099	0.139	0.106	0.222	0.065
$F_{1\ macro}$	0.076	0.119	0.120	0.068	0.174	0.077
$AUC\text{-}ROC_{micro}$	0.660	0.716	0.837	0.570	0.469	0.558
$AUC\text{-}ROC_{macro}$	0.540	0.640	0.650	0.565	0.481	0.551
Евалуациска мерка	**локални и глобални карактеристики**					
Hamming Loss	0.009	0.006	0.024	0.011	0.010	0.008
Precision	0.151	0.146	0.074	0.113	0.173	0.120
Recall	0.128	0.082	0.261	0.125	0.184	0.087
F_1	0.122	0.094	0.103	0.111	0.171	0.092
Accuracy	0.089	0.080	0.061	0.096	0.153	0.079
$Precision_{micro}$	0.160	0.688	0.073	0.120	0.185	0.164
$Recall_{micro}$	0.139	0.091	0.275	0.131	0.203	0.075
$F_{1\ micro}$	0.149	0.160	0.116	0.125	0.193	0.103
$Precision_{macro}$	0.078	0.181	0.120	0.066	0.155	0.132
$Recall_{macro}$	0.085	0.078	0.162	0.097	0.196	0.063
$F_{1\ macro}$	0.070	0.095	0.109	0.069	0.153	0.070
$AUC\text{-}ROC_{micro}$	0.665	0.721	0.832	0.564	0.476	0.555
$AUC\text{-}ROC_{macro}$	0.538	0.631	0.649	0.551	0.496	0.558

Табела Д3 Резултати добиени со методите бинарна релевантност и методот множества на ознаки.

Метод	C4.5			Random Forest		
k	3	5	10	3	5	10
Евалуациска мерка	**локални карактеристики**					
Hamming Loss	0.011	0.009	0.009	0.014	0.010	0.009
Precision	0.074	0.089	0.088	0.099	0.114	0.111
Recall	0.077	0.064	0.063	0.146	0.093	0.077
F_1	0.063	0.061	0.060	0.102	0.086	0.076
Accuracy	0.038	0.039	0.038	0.063	0.056	0.050
$Precision_{micro}$	0.073	0.089	0.099	0.090	0.107	0.123
$Recall_{micro}$	0.078	0.061	0.060	0.151	0.096	0.080
$F_{1\ micro}$	0.075	0.073	0.075	0.113	0.101	0.097
$Precision_{macro}$	0.011	0.012	0.011	0.028	0.023	0.023
$Recall_{macro}$	0.012	0.008	0.010	0.037	0.020	0.014
$F_{1\ macro}$	0.010	0.008	0.009	0.026	0.019	0.015
$AUC\text{-}ROC_{micro}$	0.503	0.492	0.521	0.577	0.575	0.623
$AUC\text{-}ROC_{macro}$	0.502	0.522	0.513	0.520	0.511	0.529
Евалуациска мерка	**глобални карактеристики**					
Hamming Loss	0.011	0.011	0.011	0.011	0.009	0.007
Precision	0.116	0.121	0.115	0.198	0.204	0.206
Recall	0.123	0.122	0.119	0.243	0.190	0.158
F_1	0.105	0.109	0.105	0.198	0.178	0.163
Accuracy	0.074	0.081	0.077	0.151	0.141	0.131
$Precision_{micro}$	0.123	0.117	0.112	0.176	0.235	0.291
$Recall_{micro}$	0.144	0.137	0.132	0.266	0.210	0.182
$F_{1\ micro}$	0.133	0.126	0.121	0.212	0.222	0.224
$Precision_{macro}$	0.077	0.074	0.076	0.176	0.191	0.217
$Recall_{macro}$	0.102	0.097	0.097	0.162	0.148	0.133
$F_{1\ macro}$	0.077	0.073	0.075	0.149	0.146	0.146
$AUC\text{-}ROC_{micro}$	0.555	0.557	0.548	0.663	0.657	0.699
$AUC\text{-}ROC_{macro}$	0.549	0.535	0.541	0.629	0.625	0.636
Евалуациска мерка	**локални и глобални карактеристики**					
Hamming Loss	0.011	0.011	0.011	0.012	0.009	0.007
Precision	0.112	0.117	0.114	0.169	0.189	0.197
Recall	0.127	0.113	0.119	0.230	0.176	0.146
F_1	0.102	0.100	0.103	0.176	0.164	0.151
Accuracy	0.069	0.073	0.073	0.127	0.128	0.118
$Precision_{micro}$	0.120	0.108	0.111	0.158	0.221	0.278
$Recall_{micro}$	0.140	0.123	0.129	0.259	0.186	0.166
$F_{1\ micro}$	0.129	0.115	0.119	0.196	0.202	0.208
$Precision_{macro}$	0.072	0.058	0.063	0.141	0.177	0.179
$Recall_{macro}$	0.098	0.081	0.086	0.147	0.117	0.101
$F_{1\ macro}$	0.073	0.058	0.063	0.126	0.127	0.114
$AUC\text{-}ROC_{micro}$	0.559	0.555	0.535	0.655	0.644	0.684
$AUC\text{-}ROC_{macro}$	0.562	0.537	0.540	0.619	0.613	0.608

Табела Д4 Резултати добиени со HOMER методот во комбинација со C4.5 и Random Forest класификаторите. Направени се анализи за различен број на кластери *k*.

Метод	Naïve Bayes			*k*NN		
k	**3**	**5**	**10**	**3**	**5**	**10**
Евалуациска мерка	**локални карактеристики**					
Hamming Loss	0.014	0.022	0.032	0.012	0.011	0.011
Precision	0.039	0.034	0.030	0.086	0.086	0.085
Recall	0.066	0.100	0.123	0.086	0.089	0.086
F_1	0.042	0.044	0.042	0.077	0.078	0.077
Accuracy	0.025	0.025	0.023	0.059	0.060	0.059
$Precision_{micro}$	0.046	0.034	0.027	0.078	0.080	0.078
$Recall_{micro}$	0.073	0.101	0.130	0.091	0.094	0.091
$F_{1\ micro}$	0.057	0.051	0.044	0.084	0.086	0.084
$Precision_{macro}$	0.008	0.010	0.010	0.051	0.055	0.054
$Recall_{macro}$	0.019	0.037	0.051	0.070	0.075	0.070
$F_{1\ macro}$	0.007	0.010	0.010	0.050	0.055	0.052
$AUC\text{-}ROC_{micro}$	0.546	0.576	0.591	0.555	0.557	0.562
$AUC\text{-}ROC_{macro}$	0.539	0.542	0.540	0.525	0.523	0.531
Евалуациска мерка	**глобални карактеристики**					
Hamming Loss	0.010	0.011	0.010	0.008	0.008	0.008
Precision	0.091	0.107	0.098	0.293	0.293	0.295
Recall	0.088	0.118	0.097	0.295	0.294	0.297
F_1	0.076	0.097	0.082	0.284	0.283	0.285
Accuracy	0.050	0.069	0.055	0.259	0.259	0.260
$Precision_{micro}$	0.096	0.108	0.099	0.339	0.338	0.340
$Recall_{micro}$	0.089	0.117	0.089	0.357	0.356	0.358
$F_{1\ micro}$	0.092	0.113	0.094	0.348	0.347	0.349
$Precision_{macro}$	0.056	0.063	0.056	0.344	0.343	0.344
$Recall_{macro}$	0.049	0.056	0.048	0.385	0.385	0.385
$F_{1\ macro}$	0.046	0.048	0.045	0.324	0.323	0.324
$AUC\text{-}ROC_{micro}$	0.568	0.608	0.602	0.670	0.683	0.688
$AUC\text{-}ROC_{macro}$	0.542	0.555	0.556	0.683	0.686	0.688
Евалуациска мерка	**локални и глобални карактеристики**					
Hamming Loss	0.010	0.012	0.013	0.008	0.008	0.008
Precision	0.076	0.093	0.087	0.296	0.297	0.295
Recall	0.092	0.124	0.142	0.301	0.302	0.299
F_1	0.072	0.094	0.092	0.287	0.288	0.285
Accuracy	0.048	0.061	0.057	0.261	0.262	0.261
$Precision_{micro}$	0.091	0.102	0.092	0.323	0.325	0.326
$Recall_{micro}$	0.091	0.138	0.148	0.359	0.360	0.358
$F_{1\ micro}$	0.091	0.117	0.113	0.340	0.341	0.341
$Precision_{macro}$	0.055	0.069	0.072	0.316	0.319	0.316
$Recall_{macro}$	0.047	0.068	0.075	0.390	0.392	0.392
$F_{1\ macro}$	0.043	0.054	0.061	0.313	0.315	0.314
$AUC\text{-}ROC_{micro}$	0.563	0.601	0.615	0.670	0.678	0.682
$AUC\text{-}ROC_{macro}$	0.548	0.568	0.566	0.685	0.686	0.689

Табела Д5 Резултати добиени со HOMER методот во комбинација со Naïve Bayes и *k*-nn класификаторите. Направени се анализи за различен број на кластери *k*.

k	1	2	3	5	10
Евалуациска мерка	**локални карактеристики**				
Hamming Loss	0.006	0.006	0.006	0.006	0.006
Precision	0.004	0.011	0.012	0.007	0.005
Recall	0.002	0.006	0.005	0.003	0.002
F_1	0.002	0.007	0.007	0.004	0.002
Accuracy	0.002	0.006	0.005	0.003	0.002
$Precision_{micro}$	0.085	0.147	0.192	0.156	0.114
$Recall_{micro}$	0.001	0.005	0.005	0.002	0.001
$F_{1\ micro}$	0.003	0.009	0.010	0.005	0.003
$Precision_{macro}$	0.005	0.009	0.010	0.007	0.005
$Recall_{macro}$	0.004	0.003	0.002	0.000	0.000
$F_{1\ macro}$	0.005	0.004	0.003	0.001	0.000
$AUC\text{-}ROC_{micro}$	0.817	0.818	0.819	0.820	0.821
$AUC\text{-}ROC_{macro}$	0.533	0.544	0.547	0.553	0.567
Евалуациска мерка	**глобални карактеристики**				
Hamming Loss	0.008	0.007	0.006	0.006	0.007
Precision	0.295	0.242	0.226	0.177	0.141
Recall	0.275	0.191	0.163	0.114	0.087
F_1	0.272	0.199	0.175	0.126	0.097
Accuracy	0.243	0.178	0.151	0.105	0.078
$Precision_{micro}$	0.338	0.397	0.393	0.359	0.314
$Recall_{micro}$	0.331	0.236	0.199	0.143	0.103
$F_{1\ micro}$	0.335	0.296	0.264	0.204	0.155
$Precision_{macro}$	0.245	0.224	0.182	0.122	0.068
$Recall_{macro}$	0.280	0.204	0.150	0.087	0.042
$F_{1\ macro}$	0.231	0.188	0.146	0.090	0.045
$AUC\text{-}ROC_{micro}$	0.867	0.870	0.870	0.871	0.872
$AUC\text{-}ROC_{macro}$	0.690	0.696	0.699	0.703	0.709
Евалуациска мерка	**локални и глобални карактеристики**				
Hamming Loss	0.008	0.006	0.006	0.006	0.006
Precision	0.299	0.233	0.232	0.189	0.147
Recall	0.282	0.185	0.173	0.127	0.093
F_1	0.276	0.194	0.182	0.138	0.104
Accuracy	0.247	0.175	0.157	0.115	0.085
$Precision_{micro}$	0.324	0.419	0.391	0.397	0.374
$Recall_{micro}$	0.334	0.228	0.210	0.157	0.114
$F_{1\ micro}$	0.329	0.295	0.273	0.225	0.175
$Precision_{macro}$	0.224	0.226	0.209	0.140	0.093
$Recall_{macro}$	0.290	0.197	0.160	0.099	0.050
$F_{1\ macro}$	0.224	0.186	0.158	0.100	0.057
$AUC\text{-}ROC_{micro}$	0.869	0.872	0.873	0.873	0.873
$AUC\text{-}ROC_{macro}$	0.694	0.699	0.702	0.704	0.711

Табела Д6 Резултати добиени со ML_*k*NN методот користејќи различен број на најблиски соседи *k*.