УНИВЕРЗИТЕТ У БЕОГРАДУ
МАТЕМАТИЧКИ ФАКУЛТЕТ

Улфета А. Маровац

Истраживање образаца у одређивању карактеристика протеина

Докторска дисертација

Београд, 2015

UNIVERSITY OF BELGRADE
FACULTY OF MATHEMATICS

Ulfeta A. Marovac

Mining sequential patterns for determination of protein characteristics

Doctoral dissertation

Belgrade, 2015

Подаци о ментору и члановима комисије

Ментор

проф. др Ненад Митић, ванредни професор, Математички факултет, Универзитет у Београду

Чланови комисије

проф. др Ненад Митић, ванредни професор, Математички факултет, Универзитет у Београду

проф. др Гордана Павловић Лажетић, редовни професор, Математички факултет, Универзитет у Београду

др Мирјана Павловић, виши научни сарадник, Институт за општу и физичку хемију, Универзитет у Београду

Датум одбране:

Подаци о докторској дисертацији

Наслов докторске дисертације: Истраживање образаца у одређивању карактеристика протеина

Резиме: Беланчевине или протеини су важни биолошки макромолекули полимерне природе (полипептиди), који се састоје од амино киселина и представљају основну градивну јединицу сваке ћелије. У њихов састав улазе 20+3 различите амино киселине због чега се у биолошким базама података представљају као ниске формиране од 23 различита карактера. Протеини се могу класификовати на основу њихове примарне структуре, секундарне структуре, функција које обављају, итд.

Једна од могућих класификација протеина по функцији је према припадности одређеном кластеру ортологих група ЦОГ[1] (Cluster of Orthhologous Groups - COGs). Ова класификација је заснована на претходном поређењу протеина према сличности по примарној структури, која је најчешће последица хомологије, тј. заједничког (еволуционог) порекла. ЦОГ база података је добијена поређењем познатих или предвиђених протеина комплетно секвенционисаних прокариотских (археа и бактеријских) генома и класификацијом према њиховој ортологији. Протеини су класификовани у 25 категорија, које могу бити распоређене у три основне функционалне групе (протеини одговорни за: (1) садржај и обраду информација, (2) ћелијске процесе и (3) метаболизам), или у групу недовољно окарактерисаних протеина. Класификација протеина према припадности одређеној ЦОГ категорији (*KOG* за еукариотске организме) је важна за боље разумевање биолошких процеса, као и различитих патолошких стања код људи и других организама.

У раду је предложен модел за класификацију протеина у ЦОГ категорије на основу аминокиселинских n-грама (ниски дужине n). Скуп података садржи протеинске секвенце генома из 8 различитих таксономских класа [TKL97] бактерија (*Aquificales*, *Bacteroidia*, *Chlamydiales*, *Chlorobia*, *Chloroflexia*, *Cytophagia*, *Deinococci*, *Prochlorales*) за које постоје информације о класификацији по

[1]Користићемо у даљем тексту скраћеницу ЦОГ за кластере ортологих група а не КОГ да не би дошло до мешања термина са класификацијом еукариотских организама (*euKaryote Orthologous Groups-KOG*)

ЦОГ категоријама. Приказана је нова метода заснована на генерализованим системима једначина Булове алгебре, која се користи за издвајање n-грама који карактеришу протеине одговарајуће ЦОГ категорије. Приказаном методом значајно се смањује број n-грама који се обрађују у односу на претходно коришћене методе n-грамске анализе, тако да се добија на уштеди меморијског простора и времена обраде протеина.

До сада познате методе класификације протеина по функционалним категоријама су вршиле поређење сваког новог протеина (коме треба одредити функцију) са скупом свих протеина који су већ класификовани према функцијама ради одређивања групе која садржи протеине који су најсличнији протеину који се класификује. Предност нове методе у односу на претходне је што не врши секвенца- секвенца поређење већ се у протеину траже обрасци (n-грами дужине до 10) који су карактеристични за одговарајућу ЦОГ категорију. Издвојени обрасци придружени одговарајућој ЦОГ категорији описују секвенце одређене дужине које су се до сада појавиле само у протеинима те ЦОГ категорије али не и у протеинима осталих ЦОГ категорија. На основу предложене методе релизован је предиктор за класификацију протеина према ЦОГ категоријама. Најмања поузданост предвиђања се задаје као улазни параметар предиктора. При тестирању предиктора постигнути су јако добри резултати са највећом поузданошћу класификације од 99%.

Због својих особина и једноставности конструкције модела, предложена метода може да се примени и на сличним областима у којима се проблем решава преко n-грамске анализе секвенци.

Кључне речи: карактеристике протеина, класификација, истраживање секвенцијалних образаца, n-грам, Булова алгебра

Научна област: Рачунарство

Ужа научна област: Истраживање података

УДК број: [004.6:004.832.2+512.563]:547.96(043.3)

Dissertation Data

Doctoral dissertation title: Mining sequential patterns for determination of protein characteristics

Abstract: Proteins are significant biological macromolecules of polymeric nature (polypeptides), which contain amino acids and are basic structural units of each cell. Their contents include 20+3 amino acids and, as a consequence, they are presented in biological databases as sequences formed from 23 different characters. Proteins can be classified based on their primary structure, secondary structure, function etc.

One of possible classifications of proteins by their function is related to their contents in a certain cluster of orthololegous groups (COGs). This classification is based on the previous comparison of proteins by their similarities in their primary structures, which is most often a result of homology, i.e. their mutual (evolutionary) origin. COG database is obtained by comparison of the known and predicted proteins encoded in the completely sequenced prokaryotic (archaea and bacteria) genomes and their classification by orthology. The proteins are classified in 25 categories which can be ordered in three basic functional groups (the proteins responsible for: (1) information storage and processing; (2) cellular processes and signaling; and (3) metabolism), or in a group of poorly characterized proteins. Classification of proteins by their contents in certain COG category (euKaryote Orthologous Groups-KOG for eukaryotic organisms) is significant for better understanding of biological processes and various pathological conditions in people and other organisms.

The dissertation proposed the model for classification of proteins in COG categories based on amino acid n-grams (sequences of n- length). The set of data contains protein sequences of genomes from 8 different taxonomic classes [TKL97] of bacteria (Aquificales, Bacteroidia, Chlamydiales, Chlorobia, Chloroflexia, Cytophagia, Deinococci, Prochlorales), which are known to have been classified by COG categories. The new method is presented, based on the generalized systems of Boolean equations, used for separation of n-grams characteristic for proteins of corresponding COG categories. The presented method significantly reduces the number of processed n-grams in comparison to previously used methods of n-gram analysis,

thus more memory space is provided and less time for protein procession is necessary.

The previously known methods for classification of proteins by functional categories compared each new protein (whose function had to be determined) to the set of all proteins which had already been classified by functions in order to determine the group which contained most similar proteins to the one which was to be classified. In relation to the previous, the advantage of the new method is in its avoidance of sequence-sequence comparison and in search for those patterns (n-grams, up to 10 long) in a protein which are characteristic of the corresponding COG category. The selected patterns are added to a corresponding COG category and describe sequences of certain length, which have previously appeared in that COG category only, not in the proteins of other COG categories.

On the basis of the proposed method, the predictor for determination of the corresponding COG category for a new protein is implemented. Minimal precision of the prediction is one of the predictors arguments. During the test phase the constructed predictor shown excellent results, with maximal precision of 99% reached for some proteins.

According to its properties and relatively simple construction, the proposed method can be applied in similar domains where the solution of problem is based on n-gram sequence analysis.

Keywords: characteristics of proteins, classification, mining sequential patterns, n-gram, Boolean algebra

Scientific field: Computer science

Scientific discipline: Data mining

UDC number: [004.6:004.832.2+512.563]:547.96(043.3)

Предговор

У овом раду се разматра могућност употребе секвенцијалних образаца за одређивање карактеристика протеина. Карактеристике протеина могу бити структурне (нпр. уређеност протеина) и функционалне (нпр. припадност ЦОГ категоријама). Представљен је нов приступ у истраживању секвенцијалних образаца заснован на Буловој алгебри. Издвојени секвенцијални обрасци коришћени су за конструкцију модела за класификацију по ЦОГ категоријама. Тестирањем је утврђено да модел омогућава врло високу прецизност класификације, често преко 90%. Нови метод за предвиђање функционалних карактеристика протеина може се користити као алтернатива познатим методама за класификацију протеина по функцијама (рачунарски захтевним методама) или као њихова допуна.

Овај рад је резултат вишегодишњег испитивања разних могућности у циљу решавања постављеног проблема. На том путу имала сам професионалну и пријатељску подршку од свог ментора проф. др Ненада Митића. За постављање биолошког проблема и разјашњење свих дилема дугујем захвалност др Милошу Бељанском са Института за општу и физичку хемију. Хвала члановима комисије са којима сам и током студија имала сусрета на пруженој подршци. Бити део Математичког факултета није формална припадност већ начин живота и на то сам поносна. Захваљујем се и колегама Државног универзитета у Новом Пазару са катедре за математику и информатику, посебно проф. др Драгићу Банковићу чије су идеје биле покретачка снага овог рада. На крају хвала мом супругу и нашој великој породици на стрпљењу, љубави и мотивацији. Без њих би сваки неуспех био пропаст, а сваки успех кратког даха. Надам се да ће ова порука доћи и до мог оца који ми је био ослонац од првог дана факултета учећи ме да рад и труд нису узалудни. Овом дисертацијом завршавам један од оних задатака које је он започео и посвећујем је својој деци да им буде охрабрење за сваки започети посао.

Садржај

Предговор **vi**

Списак слика **viii**

Списак табела **ix**

1 Увод **1**

1.1 Позадина и мотивација 1

1.2 n-Грамска анализа биолошких макромолекула (ДНК, протеини) . 7

1.3 Проблем истраживања и циљ дисертације 8

1.4 Преглед дисертације 9

2 Модели и методе за одређивање карактеристика протеина **11**

2.1 Постојеће методе за одређивање карактеристика протеина 11

2.2 Истраживање образаца помоћу метода истраживања података . . 14

2.2.1 Методе истраживања података 15

2.2.2 Класификација 16

2.2.3 Правила придруживања 19

2.2.4 Истраживање секвенцијалних образаца 22

2.2.5 Истраживање образаца у биоинформатици 28

3 Модел за одређивање карактеристичних n-грама за ЦОГ-ове протеина **30**

3.1 Модел за одређивање карактеристичних аминокиселинских подниски у протеинима 30

3.1.1 Мотивација 30

3.1.2 Опис методе и имплементација 33

3.2 Конструкција модела за предвиђање ЦОГ категорија протеина . . 44

4 Тестирање и примена модела **49**

4.1 Материјал 49

4.2 Квалитет издвојених образаца 51

4.3 Утицај различитих параметара на квалитет издвојених секвенцијалних образаца 57

4.4 Квалитет модела за класификацију протеина по ЦОГ категоријама 59

4.5 Поређење образаца добијених на различитим фамилијама 63

4.5.1 Резултати предиктора протеина по функционалним категоријама 66

5 Закључак **70**

Додаци **72**

А Улазни скуп организама 72

Б Карактеристике класификационог модела за одређену класу бактерија 78

В Карактеристике класификационог модела за све врсте бактерија . . 81

Г Резултати предиктора на некласификованим протеинима 84

Д Скраћенице коришћене у раду 88

Литература **89**

Биографија **95**

Списак слика

1.1 Шематски приказ примарне, секундарне, терцијарне и кватернарне (комплекс два или више протеина) структуре протеина 3
1.2 Факсимил са преводом апстракта рада *"A genomic perspective on protein families"* [TKL97] у коме је први пут описана ЦОГ база података. 4
1.3 Филогенетско стабло које приказује порекло и однос две основне врсте хомологих секвенци (ортологе и паралоге). Различите боје означавају различите врсте, кружићи протеине у оквиру исте, или различитих врста (боје). 5
1.4 Класификација ЦОГ-ова по ЦОГ категоријама, најбројније категорије 5

2.1 Поравњање секвенци хистон $H1$ протеина (секвенце од 120-те до 180-те амино киселине) у различитим организмима 13
2.2 Илустрација класификације 16
2.3 Систематизација алгоритама правила придруживања 22

3.1 Илустрација изградње модела за класификацију 46

4.1 Број издвојених дескриптора по ЦОГ категоријама за различите дужине n-грама 52
4.2 Број издвојених приближних дескриптора по ЦОГ категоријама за различите дужине n-грама 52

Списак табела

1.1 Једнословне и трословне ознаке амино киселина 2

1.2 ЦОГ категорије и функције протеина који им припадају по групама протеини одговорни за: 1. ћелијске процесе; 2. садржај и обраду информација; 3. метаболизам и 4. недовољно окарактерисани протеини 6

2.1 Матрица конфузије 17

2.2 Матрица конфузије за n класа 19

4.1 Класе бактерија коришћене за израду модела 50

4.2 Подела генома у скупове података за тренинг и тестирање по класама бактерија 51

4.3 Број издвојених дескриптора по класама бактерија 53

4.4 Број издвојених приближних дескриптора по класама бактерија . 54

4.5 Број дескриптора $\#d$ за које су у моделу пронађени карактеристични n-грами у више од 0.5% протеина одговарајуће ЦОГ категорије 55

4.6 Број приближних дескриптора ($\#d*$) за које су у моделу пронађени карактеристични n-грами у више од 0.5% протеина одговарајуће ЦОГ категорије 55

4.7 Дескриптори са највећим бројем појава у моделу 56

4.8 Приближни дескриптори са највећим бројем појава у моделу . . . 56

4.9 Проценат карактеристичних n-грама пронађених у одговарајућој ЦОГ категорији у подацима за тестирање за различиту вредност прага min_σ 58

4.10 Проценат карактеристичних n-грама пронађених у очекиваној ЦОГ категорији у подацима за различите ЦОГ категорије и дужине n-грама 58
4.11 Проценат карактеристичних n-грама пронађених у очекиваној ЦОГ категорији посматрано по различитим ЦОГ категоријама и класама бактерија 60
4.12 Квалитет модела у зависности од минималног броја различитих образаца пронађених у протеину да би протеин био класификован у одговарајућу ЦОГ категорију (h) 61
4.13 Квалитет модела у зависности од минималног броја карактеристичних n-грама пронађених у протеину да би протеин био класификован у одговарајућу ЦОГ категорију (h) 61
4.14 Квалитет модела посебно за сваку ЦОГ категорију 62
4.15 Број различитих образаца груписаних по броју ЦОГ категорија којима су придружени 63
4.16 Број различитих образаца груписаних по броју класа бактерија у којима се појављују 64
4.17 Квалитет класификационог модела направљеног над свим класама бактерија у зависности од минималног броја (h) пронађених карактеристичних n-грама при додели одговарајуће ЦОГ категорије, при чему је апсолутна подршка образаца коришћених у моделу већа од 5 ($min\_\sigma = 5$) 64
4.18 Квалитет класификационог модела направљеног над свим класама бактерија у зависности од минималног броја (h) пронађених карактеристичних n-грама при додели одговарајуће ЦОГ категорије, при чему је апсолутна подршка образаца коришћених у моделу већа од 10 ($min\_\sigma = 10$) 65
4.19 Квалитет модела посебно за сваку ЦОГ категорију 66
4.20 Број нових протеина придружених одговарајућој ЦОГ категорији 68
4.21 Резултати предвиђања ЦОГ категорије за некласификоване протеине 68

A.1 Број протеина по ЦОГ категоријама у скуповима за тренинг и тестирање 72

А.2 Број протеина по ЦОГ категоријама у скуповима за тренинг и тестирање 73
А.3 Подаци о организмима обрађених класа бактерија 74
Б.1 Карактеристике класификационог модела изражене преко прецизности и одзива добијене при класификацији протеина у одговарајућу ЦОГ категорију у зависности од минималног броја различитих карактеристичних n-грама (дескриптора) пронађених у протеину (h) 78
В.1 Карактеристике класификационог модела изражене преко прецизности и одзива добијене при класификацији протеина у одговарајућу ЦОГ категорију за различите вредности min_σ и h када се користе издвојени обрасци из свих обрађених класа бактерија 81
Г.1 Табела са резултатима предвиђања ЦОГ категорије за некласификоване протеине 84
Д.1 Пуни називи скраћеница у раду као и скраћенице и називи њихових оригинала 88

Поглавље 1

Увод

1.1 Позадина и мотивација

Биоинформатика (*bioinformatics*) је област науке која проучава биологију помоћу статистике и рачунарских метода. Она се бави информатичким основама, као и бележењем, организацијом и анализом биолошких података. Иако релативно млада дисциплина, биоинформатика је нашла примену у многим научним дисциплинама укључујући биологију, биохемију, медицину, хемију, итд. Од свог почетка 1980-их година развој биоинформатике је у директној вези са развојем информатике и то не само у развоју специјализованог хардвера и софтвера. Поред тога, биотехнолошки напредак представљања генома, микроорганизама и протеома додатно су допринели развоју биоинформатике.

Потреба за биоинформатиком се јавила као последица експоненцијалног раста количине података. Осим бројности, расте и сложеност, што захтева развој нових начина интерпретације постојећих биолошких података. Основни циљ биоинформатике је трансформација података у знање. Истраживање података (*data mining*) и откривање знања из података (*knowledge discovery from data - KDD*) служе за издвајање нетривијалних, имплицитних, претходно непознатих и потенцијално корисних информација из података.

По својој хемијској природи биолошки макромолекули се могу поделити у три категорије. То су полимери шећера (полисахариди), полимери нуклеотида (нуклеинске киселине) и полимери амино киселина (полипептиди по хемијским везама између две суседне амино киселине). Беланчевине (или протеини) су

линеарни полимери амино киселина, у чији састав улазе 23 различите амино киселине. У биолошким базама података су представљени као ниске формиране од 23 слова.[1]

Број и редослед амино киселина (означен као примарна структура протеина)

Табела 1.1: Једнословне и трословне ознаке амино киселина

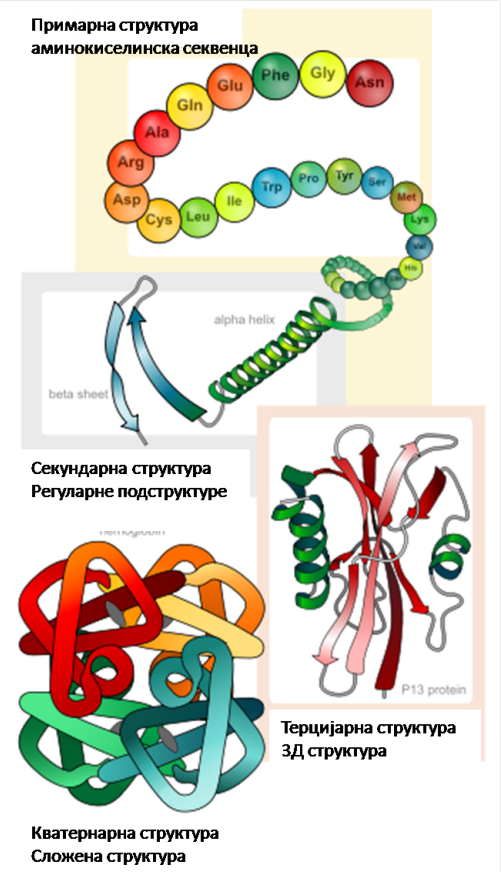
Назив амино киселине	Назив на енглеском језику	Једнословна ознака	Трословна† ознака
Аланин	Alanine	A	Ala
Аспарагин или аспарагинска к.	Asparagine or aspartic acid	B	Asx
Цистеин	Cysteine	C	Cys
Аспарагинска киселина	Aspartic acid	D	Asp
Глутаминска киселина	Glutamic acid	E	Glu
Фенилаланин	Phenylalanine	F	Phe
Глицин	Glycine	G	Gly
Хистидин	Histidine	H	His
Изолеуцин	Isoleucine	I	Ile
Леуцин или изолеуцин	Leucine or Isoleucine	J	Xle
Лизин	Lysine	K	Lys
Леуцин	Leucine	L	Leu
Метионин	Methionine	M	Met
Аспарагин	Asparagine	N	Asn
Пиролизин	Pyrrolysine	O	Pyl
Пролин	Proline	P	Pro
Глутамин	Glutamine	Q	Gln
Аргинин	Arginine	R	Arg
Серин	Serine	S	Ser
Треонин	Threonine	T	Thr
Селеноцистеин	Selenocysteine	U	Sec
Валин	Valine	V	Val
Триптофан	Tryptophan	W	Trp
Неодређена или непозната а.к.	Unspecified or unknown	X	Xaa
Тирозин	Tyrosine	Y	Tyr
Глутамин или глутаминска к.	Glutamine or glutamic acid	Z	Glx
N-формилметионин	N-Formylmethionine		fMet

†Ознака за N-формилметионин је четворословна

варира и одређује њихову просторну (секундарну и терцијарну-3Д) структуру (Слика 1.1). Секундарна структура се може дефинисати као регуларна (или уређена) структура полипептидног ланца. Постоје три основна типа секундарне структуре: завојница, трака и заокрет. Поред уређене структуре, протеини могу да садрже и структурно неуређене делове (*intrinsically disordered/unstructured regions*).

[1]Поред ознака за ове 23 амино киселине јављају се још и ознаке које се користе у случају да амино киселина у протеину није прецизно одређена: B - Аспарагин или Аспарагинска киселина, J - Леуцин или Изолеуцин, Z - Глутамин или Глутаминска киселина, и X - непозната или неодређена АК.

Функција протеина проистиче из њихове структуре и интеракције са другим



Слика 1.1: Шематски приказ примарне, секундарне, терцијарне и кватернарне (комплекс два или више протеина) структуре протеина

молекулима. Они се могу класификовати према различитим критеријумима:

1. према сличности-разлици по својој структури (примарној, секундраној или терцијарној- 3Д),

2. према сличности-разлици по својој молекулској (биохемијској) или биолошкој функцији

3. према сличности-разлици по својој позицији у ћелији-организму итд,

4. према комбинацији два или више критеријума.

Структурална класификација протеина (*Structural Classification of Proteins (SCOP)*-база података, *http://scop.mrc-lmb.cam.ac.uk/scop/*) је једна од најчешће примењиваних. Она се базира на класификацији по структуралним доменима, тј. сличности њихових секвенци и 3Д структура [Mur+95].

Према биолошким функцијама које обављају протеини се могу поделити у 8 категорија: ензими, складишни (резервни) протеини, транспортни протеини, контрактилни протеини, заштитни протеини, хормони, токсини и структурни протеини.

Пфам (*Pfam-protein family*) је класификација протеина - база података (*http://pfam.xfam.org/*) која се заснива на њиховој анотацији и вишеструком поређењу секвенци применом скривених Марковљевих модела [Bat+00].

Кластери ортологих група протеина-гена (*Clusters of Orthologous Groups of proteins/genes (COGs)*) је класификација - база података (*http://www.ncbi.nlm.nih.gov/COG*) која се базира на филогенетској класификацији (хомологији) протеина из комплетних генома бактерија, археа или еукариа (*KOG*) (Слика 1.2). Хомологи протеини (гени) имају

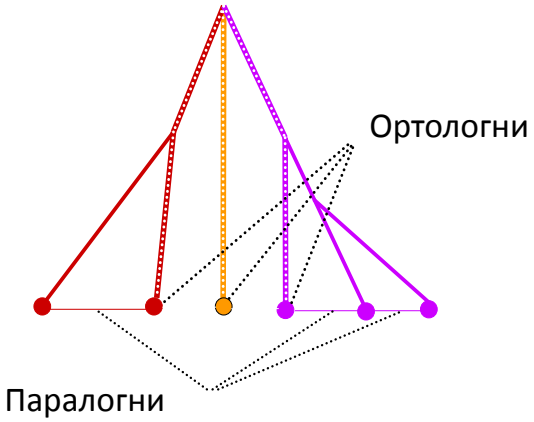
Abstract: In order to extract the maximum amount of information from the rapidly accumulating genome sequences, all conserved genes need to be classified according to their homologous relationships. Comparison of proteins encoded in seven complete genomes from five major phylogenetic lineages and elucidation of consistent patterns of sequence similarities allowed the delineation of 720 clusters of orthologous groups (COGs). Each COG consists of individual orthologous proteins or orthologous sets of paralogs from at least three lineages. Orthologs typically have the same function, allowing transfer of functional information from one member to an entire COG. This relation automatically yields a number of functional predictions for poorly characterized genomes. The COGs comprise a framework for functional and evolutionary genome analysis.

Превод абстракта: Да би се издвојила максимална количина информација из све већег броја секвенционисаних генома, сви конзервирани гени треба да буду класификовани према њиховим хомологим односима. Поређење протеина кодираних у седам комплетних генома од пет главних линија филогенетског стабла и утврђивање конзистентних образаца сличности у секвенци омогућило је одређивање 720 кластера ортологоих група (ЦОГ). Сваки ЦОГ се састоји од појединачних ортологих протеина или ортологе групе паралога од најмање три гране филогенетског стабла. Ортологе обично имају исту функцију, омогућавајући пренос информација од једног функционалног члана на цели ЦОГ. Овај однос аутоматски доприноси броју функционалних предвиђањима за слабо окарактерисане геноме. ЦОГ чини оквир за функционалну и еволутивну анализу генома.

Слика 1.2: Факсимил са преводом апстракта рада *"A genomic perspective on protein families"* [TKL97] у коме је први пут описана ЦОГ база података.

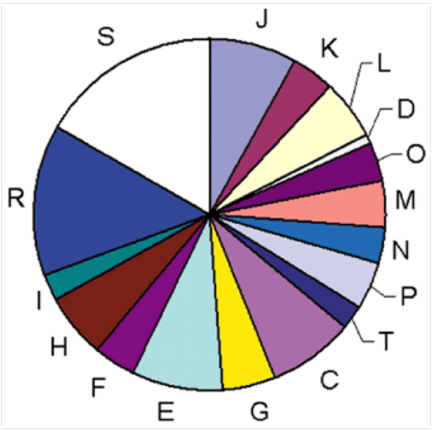
заједничко еволутивно порекло. Постоје два основна типа хомологије секвенци: ортологе и паралоге. Ортологи су протеини (гени) у различитим врстама,

који воде порекло од заједничког претка и имају исту функцију. Паралоги су протеини (гени) у истој врсти, настали дупликацијом гена и имају сличну, али различиту функцију [TKL97; Tat+00] (Слика 1.3). Сваки ЦОГ се састоји



Слика 1.3: Филогенетско стабло које приказује порекло и однос две основне врсте хомологих секвенци (ортологе и паралоге). Различите боје означавају различите врсте, кружићи протеине у оквиру исте, или различитих врста (боје).

од појединачних ортологих протеина или ортологе групе паралога од најмање три гране филогенетског стабла. Они су класификовани у 26 функционалних категорија (Слика 1.4 и Табела 1.2), укључујући и класе којој је предвиђена општа функција (*R, general functional prediction*), као и ЦОГ-ове који нису окарактерисани (*S, no functional prediction*) и који бројчано представљају највећу групу. Функционалне категорије могу бити распоређене у три основне



Слика 1.4: Класификација ЦОГ-ова по ЦОГ категоријама, најбројније категорије

функционалне групе (протеини одговорни за садржај и обраду информација,

ћелијске процесе и метаболизам), или у групу недовољно окарактерисаних протеина.

Табела 1.2: ЦОГ категорије и функције протеина који им припадају по групама протеини одговорни за: 1. ћелијске процесе; 2. садржај и обраду информација; 3. метаболизам и 4. недовољно окарактерисани протеини

ЦОГ	Група	Функција
D	1	Контрола ћелијског циклуса и митозе
M	1	Биогенеза ћелијског зида / мембране / омотача
N	1	Ћелијска покретљивост
O	1	Пост-транслациона модификација, промет протеина, шаперонске функције
T	1	Пренос сигнала
U	1	Интрацелуларни промет и секреција
Y	1	Нуклеарна структура
Z	1	Цитоскелет
V	1	Одбрамбени механизми
W	1	Екстрацелуларне структуре
A	2	РНК обрада и модификација
B	2	Хроматинска структура и динамика
J	2	Транслација
K	2	Преписка
L	2	Умножавање и поправка
C	3	Производња и конверзија енергије
E	3	Метаболизам и транспорт амино киселина
F	3	Метаболизам и транспорт нуклеотида
G	3	Метаболизам и транспорт угљених хидрата
H	3	Метаболизам коензима
I	3	Метаболизам липида
P	3	Транспорт и метаболизам неорганских јона
Q	3	Секундарна структура
R	4	Предвиђена само општа функција
S	4	Непозната функција
N.C.		Није ни у једној групи

Пошто ортологи протеини задржавају исту функцију током еволуције, идентификација ортолога је корисна за поуздано предвиђање функција протеина у секвенцама новообрађених генома. ЦОГ база се периодично ажурира после обраде нових генома. ЦОГ класификација може бити примењена на проблем функционалног обележавања нових секвенци генома

коришћењем Когнитор (*COGnitor*) програма који је доступан на ЦОГ веб страници (*www.ncbi.nlm.nih.gov/COG/cognitor.html*). ЦОГ категорија неког протеина се идентификује поређењем са сваким од гена (секвенце протеина које су кодиране у целим геномима) користећи БЛАСТ (*BLAST*) методу. ЦОГ конструктивна процедура је заснована на једноставном уочавању да су протеини из сваке групе од најмање три протеина из различитих генома, који су више слични једни другима него осталим протеинима генома којима припадају, највероватније у истом скупу ортолога [Koo; Tat+03]. ЦОГ база може да се користи и за откривање недостајућих података у геному, односно гена који су пропуштени током анотације. Постоји могућност издвајања свих ЦОГ-ова који се јављају у патогеним бактеријама. У ширем смислу, ЦОГ систем је погодна платформа за разне еволуционо оријентисане анализе протеинских породица.

1.2 n-Грамска анализа биолошких макромолекула (ДНК, протеини)

Експериментално одређивање просторне структуре и функције протеина је дуг и скуп процес. Захваљујући "геномској револуцији", тј. масовном, релативно брзом и јефтином одређивању примарне структуре генома и протеина, јавља се могућност и потреба за *in silico* карактеризацијом и предвиђањем протеинске структуре и функције. До сада је развијен велики број програма чија поузданост није увек задовољавајућа, тако да постоји стална потреба за њиховим усавршавањем. Одређивање *in silico* структуре и фунције протеина убрзава и скраћује њихову експерименталну потврду. Овако добијени резултати су од великог значаја за боље разумевање биолошких процеса, као и разних патолошких стања код човека и других организама.

Протеини и геномске секвенце се могу посматрати као ниске симбола, које могу бити предмет истраживања секвенци, дисциплине истраживања података. Фаузи (*Faouzi*) и сарадници у [FRM09] су приказали приступ хијерархијског издвајања n-грама (подниске дужине n ниске дужине m, $m \geq n$) из протеина у циљу класификације протеина. Османбејолу (*Osmanbeioglu*) и сарадници су анализирали 970 микроорганизама и издвојили су n-граме који су презаступљени у једном организму и јако ретко заступљени у осталим

организмима што се може искористити у карактеризацији протеома [OG11]. Ганапатирацу (*Ganapathiraju*) и сарадници су показали у својим радовима [Gan+02b; Gan+02a; Gan+04; Pod+07; Gan+12] да се биолошке секвенце могу обрађивати истим методама као природни језик. Различите варијанте метода заснованих на n-грамима успешно су примењиване и за: одређивање сличности секвенци и реконструкцију филогенетског стабла [Wu+92; CCKS05], поређење особина кодирајућих и некодирајућих региона у геному [Man+95], одређивање лингвистичке сложености секвенци [Tro+02], класификацију и надгледано хијерахирско кластеровање геномских секвенци [TJK06], проблем откривања промотера [RBB07], карактеризацију геномских острва [MPLB08; PLMB09], итд.

1.3 Проблем истраживања и циљ дисертације

Основни циљ тезе је да се помоћу истраживања образаца направи модел за одређивање карактеристика протеина (аминокиселинских ниски) на основу садржаја и редоследа амино киселина у њима. Карактеристике могу да буду структуралне, функционалне или комбинација структуралних и функционалних особина. Једна од комбинација структуралних и функционалних карактеристика јесте и припадност некој ЦОГ категорији. Традиционални приступ за класификацију протеина у ЦОГ категорије је заснован на обимном поређењу сличности нових секвенци са скупом већ класификованих секвенци.

У овом раду је испитивана зависност карактеристика протеина од њиховог аминокиселинског састава. Аминокиселински састав протеина је представљен аминокиселинским n-грамима који се јављају у протеинским секвенцама.

Циљ дисертације је прављење новог модела за одређивање припадности протеина кластерима ортологих група (ЦОГ-овима). Циљ је да се одређивање сличности (хомологије) протеина заснује на карактеристичним нискама које представљају "потпис" одређене ЦОГ категорије. Претпоставка је да садржај аминокиселинских ниски у протеинима одређује њихове структуралне и функционалне карактеристике. Додатна претпоставка је да су аминокиселински n-грами циљне карактеристичне ниске. Модел треба да буде такав да омогући примену математичких метода на карактеристичне ниске

у циљу провере и процене коректности добијених резултата. Истовремено, циљ конструкције модела је да се након провере коректности он искористи за предвиђање ЦОГ категорије до сада нераспоређених протеина. Резултати добијени у тези треба да допринесу бољем разумевању биолошких система организације, са крајњом применом у медицини, фармацији, итд.

1.4 Преглед дисертације

Рад је организован у пет поглавља и пет додатка. Прво поглавље је уводно и описује мотивацију и проблем дисертације. Друго поглавље се састоји из две целине:

- опис постојећих метода за одређивање карактеристика протеина,
- опис метода истраживања образаца помоћу техника истраживања података.

У првом делу је акценат на алатима и методама које се користе за одређивање припадности протеина ЦОГ категоријама, док се у другом делу описују методе истраживања података које се користе за класификацију протеина и додатно методе истраживања секвенцијалних образаца као и њихова досадашња примена на карактеризацију протеина.

У трећем поглављу је описан нови приступ n-грамској анализи аминокиселинских секвенци заснован на Буловој алгебри. Изложен је алгоритам за издвајање секвенцијалних образаца из протеина који припадају истој ЦОГ категорији као и алгоритам за класификацију некласификованих протеина по ЦОГ категоријама.

Поглавље четири приказује резултате класификације протеина по ЦОГ категоријама добијене применом предложених метода у поређењу са резултатима који су већ раније добијени познатим методама. У њему су приказане карактеристике модела као и параметри који могу да утичу на њихово побољшање. Као крајњи резултат изложени су резултати класификације протеина који су досадашњим методама остали некласификовани.

Пето поглавље је закључак са предлозима будућег унапређења и проширења

методе. У четвртом поглављу су изложени само примери резултата док се табеле са више информација налазе у додатним поглављима која су организована у пет целина : материјал (Додатак *А*), карактеристике класификационог модела када се обрасци издвајају за сваку класу бактерија посебно (Додатак *Б*), карактеристике класификационог модела на основу здружених образаца добијених из целог материјала (Додатак *В*), резултати класификације за све протеине улазног скупа података којима ранијим методама није придружена ЦОГ категорија (Додатак *Г*), и регистар скраћеница (Додатак *Д*).

Поглавље 2

Модели и методе за одређивање карактеристика протеина

2.1 Постојеће методе за одређивање карактеристика протеина

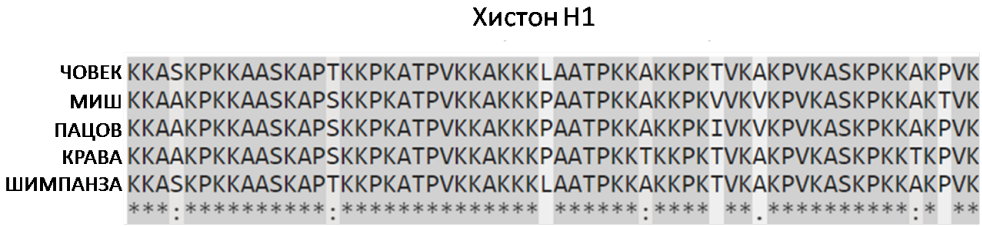
База кластера ортологих протеина (ЦОГ-ова) укључује протеине кодиране у комплетним геномским секвенцама који су класификовани према концепту ортолога. Ортолози су директни еволуциони рођаци који су повезани вертикалним везама, за разлику од паралога који представљају гене истог генома и везани су дупликацијом [Koo05]. Ортологи протеини задржавају исту функцију, док паралоги имају различите функције. Коректна класификација протеина који се налазе у комплетно секвенционираним геномима (у даљем тексту комплетним геномима) кључна је за максималну употребу геномских секвенци за функционално и еволуционарно истраживање. Као што је већ речено у уводу, ЦОГ категорија неког протеина се идентификује поређењем протеина са сваким од претходно класификованих протеина (секвенцама протеина кодираних у комплетним геномима којима је већ придружена ЦОГ категорија) користећи БЛАСТ методу. Протеини који се налазе у различитим геномима али су сличнији међусобно него са осталим протеинима генома којима припадају, јесу ортологи протеини. Овакав начин предвиђање функционише и када је сличност између протеина мала; прецизност одређивања се прилагођава захтевима корисника. Конструкција ЦОГ-а састоји се из следећих корака:

1. Поређење сличности сваке са свим осталим протеинским секвенцама;
2. Издвајање очигледних паралога гена једног генома који су сличнији једни другима него генима других врста;
3. Издвајање тројки ортолога које имају највећу сличност узимајући у обзир издвојене паралоге;
4. Спајање група ортолога у ЦОГ-ове;
5. Провера једног по једног ЦОГ-а. Овом анализом елиминишу се лажно позитивни и идентификују се протеини са мултидоменом. За протеине са мултидоменом понављају се поступци од 1-4 и издвају се скупови који не садрже мултидомен протеине;
6. Провера великих ЦОГ група које садрже много протеина из различитих генома и њихова подела на две или више група.

Једном идентификовани ЦОГ-ови коришћењем претходне процедуре могу се посредством програма Когнитор додати ЦОГ бази. Програм Когнитор се заснива на моделу најбољег слагања.

БЛАСТ

Прве информације о новооткривеним протеинима се добијају претрагом сличности са већ познатим протеинским секвенцама. Утврђивањем сличности могу се добити информације о функцијама протеина. Поравнање секвенци нуклеотидних или аминокиселинских ниски представља се у облику редова матрице (Слика 2.1, преузета са *http://en.wikipedia.org/wiki/Sequence_alignment*) уз евентуално убацивање празнина између редова. Поравнање секвенци користи се и на небиолошким секвенцама, као на пример у природним језицима или финансијским подацима. Постоје разни начини за утврђивање сличности секвенци а најпопуларнији је БЛАСТ [Alt+97]. БЛАСТ метода веома брзо даје резултате као и процену квалитета резултата. Њом се добија очекивани број случајних подударања у датом скупу чиме корисник добија информацију о прецизности. БЛАСТ даје резултате локалних поравнања протеина. Већина протеина садржи један



Слика 2.1: Поравњање секвенци хистон $H1$ протеина (секвенце од 120-те до 180-те амино киселине) у различитим организмима

или више функционалних домена. Исти домени могу да се јаве у протеинима различитих врста. БЛАСТ алгоритам тражи исте или краће домене у секвенцама ради утврђивања сличности. Уколико би БЛАСТ метода тражила поклапање дуж целе секвенце тада би се утвђивало много мање сличности међу секвенцама. Постоје различити типови БЛАСТ методе:

- за поређење нуклеотидних секвенци. Метода је оптимизована за секвенце из исте или сличне врсте и тражи скоро потпуно поклапање (megaBLAST);
- за поређење нуклеотидних секвенци различитих врста (BLASTN);
- за поређење протеинских секвенци (BLASTP);
- за поређење нуклеотидних секвенци преко протеинске базе (BLASTX), уз проверу сва три низа кодирања (reading frame);
- за поређење протеинских секвенци преко нуклеотидне базе (TBLASTN), уз динамичко превођење у сва три низа кодирања (reading frame), итд.

БЛАСТ метода користи корисничке хеуристике које представљају пречице до одговарајућих одговора. Секвенце се не пореде директно из ГенБанк[1] базе већ се преносе у посебан формат у БЛАСТ базу секвенци ради ефикасније претраге. Она дозвољава повећање брзине поређења смањењем прага осетљивости. БЛАСТ програм захтева време пропорционално производу дужина секвенце чија се сличност испитује и базе секвенци са којима се поређење врши. С обзиром да се ради о великим базама података време претраге је велико и рачунари на којима се раде овакве претраге су јако оптерећени, тако да све нове верзије БЛАСТ програма теже оптимизацији времена обраде.

[1]Регистар скраћеница је у додатку Д

БЛАСТ пролази кроз три фазе: "подешавање", "прелиминарна претрага" и "поновно трагање".

- У подешавањима БЛАСТ пролази кроз упит, претражује параметре и базу. Издвајају се парови кратких речи, секвенце одређене дужине на основу упита које задовољавају дефинисан праг поузданости. Оне служе за покретање претраге за сличним подсеквенцама у бази. У прелиминарној претрази, сви кораци се изводе на свакој секвенци у бази. Први корак је скенирање подударности речи са базом и то служи за иницијално поравнање којим се издвајају подударне секвенце без празнина.

- У другом кораку се полазни скуп секвенци без празнина користи за формирање проширења и добијање подударања са празнинама. Само она проширења са празнинама која задовољавају неку оцену квалитета чувају се за даље кораке.

- У последњој фази се може поновити претрага за мотивима при чему се скуп проширења са празнинама узима као иницијални и на њему се врше додатна брисања и уметања да би се добили нови мотиви.

Главни недостатак ових метода је време обраде. Поређење сваке секвенце са сваком секвенцом у свакој претрази сличности траје јако дуго. Зато се јавља потреба за другачијим решењем. Издвајање мотива који се везују за одређену карактеристику протеина своди поређење нове протеинске секвенце са скупом протеинских секвенци за које су претходно одређени карактеристични мотиви, на поређење само са издвојеним мотивима. Издвајање мотива се најчешће врши помоћу метода истраживања података.

2.2 Истраживање образаца помоћу метода истраживања података

Биолози играју кључну улогу у класификацији протеина. Међутим, велика количина биолошких података као што су протеини, ДНК, РНК, итд. захтева коришћење истраживачких алата и техника у циљу помоћи биолозима, углавном зато што је ручна класификација ових података скоро немогућа.

Сарадња информатичара и биолога резултирала је великим бројем нових дисциплина који се баве истраживањем биолошких података у оквиру биоинформатике. Биоинформатика је интердисциплинарна наука о тумачењу биолошких података користећи информационе технологије и информатику. Посебно активна област истраживања у оквиру биоинформатике је примена и развој техника истраживања података за решавање биолошких проблема. Примери за ову врсту истраживања укључују анализу структуре протеина, класификацију гена, класификацију тумора на основу података биохемијске структуре, статистичко моделовање интеракције протеина, итд.

2.2.1 Методе истраживања података

Истраживање података је аутоматизован процес добијања, из великих скупова података, нових информација које нису очигледно уочљиве или експлицитно представљене. Методе истраживања података омогућавају откривање карактеристика посматраног скупа података као и предвиђање исхода посматрања нових података. Методе истраживања података се могу, у општем случају, поделити у две велике групе:

описне методе којима се издвајају скривене везе унутар посматраног скупа података, и

методе предвиђања којима се предвиђају вредности података на основу познатих резултата добијених из других скупова података или историјских података.

Описне методе су кластеровање, сумаризација, правила придруживања и секвенцијална анализа. Методе предвиђања су класификација, регресија, анализа временских серија и предвиђање.

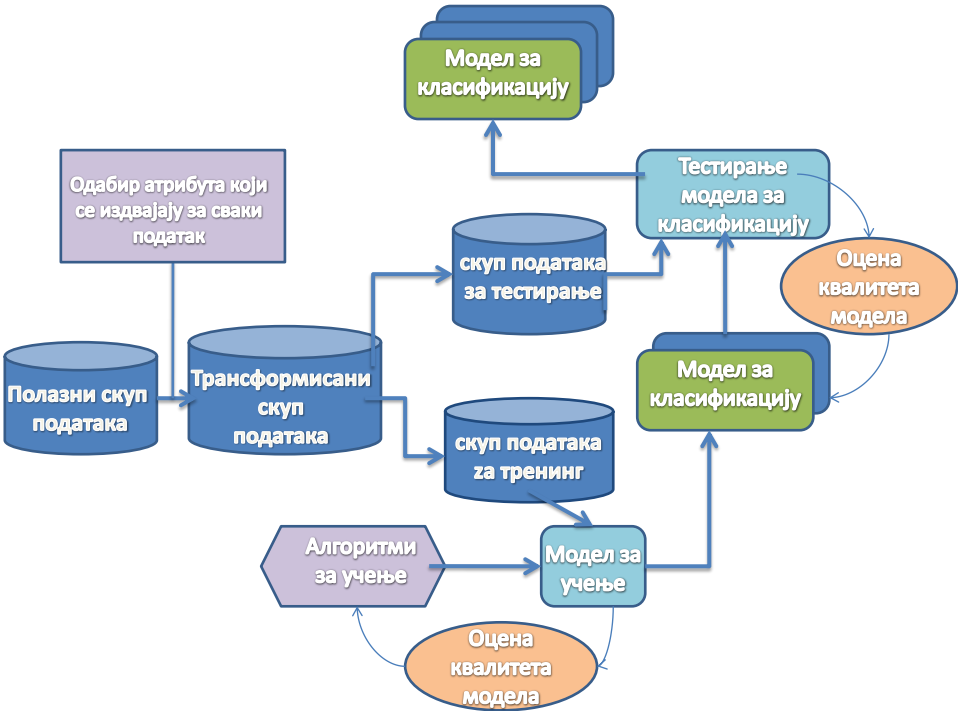
Која ће се метода користити при истрживању података зависи од типа података и од информације за којом се трага. Преглед метода истраживања података може се наћи на пример у [TSV06]. У тексту који следи биће детаљније описане методе које су коришћене у овом раду: класификација, правила придруживања и истраживање секвенцијалних образаца.

2.2.2 Класификација

Класификација је метода којој је циљ придруживање одређеног објекта једној од унапред утврђених класа. Улазни податак у класификацију је скуп слогова (подаци за тренинг). Сваки слог је облика (x, y) где је x скуп атрибута а y специјални атрибут одређен за ознаку класе. Потребно је наћи класификациони модел (функцију) који пресликава сваки скуп атрибута x у једну од предефинисаних ознака класа y.

Дефиниција 1. *Класификација је задатак учења циљне функције f да преслика сваки скуп атрибута x у једну од предефинисаних класа са ознаком y.*

Циљ класификације је доделити некласификоване слогове једној од предефинисаних класа што је могуће прецизније. Улазни подаци се обично деле у два дела: податке за тренинг помоћу којих се формира модел и податке за тестирање који се користе за проверу исправности модела (Слика 2.2). Класификација може бити коришћена и као описна метода и као метода



Слика 2.2: Илустрација класификације

предвиђања. Класификациони модел може дати карактеризацију објеката

који припадају различитим класама. С друге стране, корисна је употреба класификације за одређивање припадности класи непознатих објеката. Технике класификације су:

- методе засноване на дрветима одлучивања,
- методе засноване на правилима,
- неуронске мреже,
- статистички засноване методе,
- методе засноване на подржавајућим векторима.

Свака од техника користи алгоритам учења да би одредила најбољу везу између скупа атрибута и ознаке класе као улазних података. Мерење перформанси израчунавања класификације се заснива на броју слогова из података за тестирање који су тачно односно нетачно придружени одговарајућој класи на основу модела. Ови бројеви се уносе у табелу звану матрица конфузије. Матрица конфузије (Табела 2.1) се састоји од четири вредности f_{ij} које означавају број слогова класе i за које предвиђено да припадају класи j. Тако на пример бројеви f_{00} и f_{11} представљају слогове података којима је тачно одређена класа 0 односно 1, док f_{01} и f_{10} су бројеви слогова из података којима је погрешно придружена класа. Ове информације се користе ради одређивања квалитета модела за предвиђање.

Мере којима се изражава ваљаност модела су *тачност* или *степен грешке*.

Табела 2.1: Матрица конфузије

		Предвиђена класа	
		класа=1	класа=0
Права класа	класа=1	f_{11}	f_{10}
	класа=0	f_{01}	f_{00}

Тачност је однос броја исправно класификованих слогова и укупног броја класификованих слогова:

$$t = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}. \tag{2.1}$$

Степен грешке је однос броја неисправно класификованих слогова и укупног броја класификованих слогова:

$$sg = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}. \tag{2.2}$$

За сваку појединачну класу се могу дефинисати додатне мере: прецизност унутар класе и одзив. *Прецизност* у класи K је однос броја исправно класификованих слогова класе K и укупног броја слогова који су моделом придружени класи K:

$$p = \frac{f_{11}}{f_{11} + f_{01}}. \tag{2.3}$$

Одзив у класи K је однос броја исправно класификованих слогова класе K и укупног броја слогова у класи K:

$$o = \frac{f_{11}}{f_{11} + f_{10}}. \tag{2.4}$$

Ако се матрица конфузије посматра за бинарни класификациони проблем (позитивно/негативно), онда се често користе следеће ознаке:

- TP(*true positive-TP*) или f_{++}, за број позитивних примера коректно предвиђених у класификационом моделу,
- TN(*true negative-TN*) или f_{--}, за број негативних примера коректно предвиђених у класификационом моделу,
- NP(*false positive-FP*) или f_{-+}, за број негативних примера некоректно предвиђених у класификационом моделу, и
- NN(*false negative-FN*) или f_{+-}, за број позитивних примера некоректно предвиђених у класификационом моделу.

У општем случају када постоји више од две класе за које се врши класификација, матрица конфузије се може представити у облику Табеле 2.2, док одговарајуће мере ваљаности (*тачност*, *степен грешке*, *прецизност* и *одзив*) модела за класификацију у том случају имају следећи облик:

$$t = \frac{\sum_{i=1}^{n} f_{ii}}{\sum_{i=1}^{n}\sum_{j=1}^{n} f_{ij}}, \tag{2.5}$$

$$sg = \frac{\sum_{i=1}^{n}\sum_{j=i+1}^{n} f_{ij} + \sum_{j=1}^{n}\sum_{i=j+1}^{n} f_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} f_{ij}}, \tag{2.6}$$

$$p = \frac{f_{ii}}{\sum_{j=1}^{n} f_{ji}},\ i = 1, ..., n, \tag{2.7}$$

$$o = \frac{f_{ii}}{\sum_{j=1}^{n} f_{ij}},\ i = 1, ..., n. \tag{2.8}$$

Табела 2.2: Матрица конфузије за n класа

		Предвиђена класа			
		1	2	...	n
Права класа	1	f_{11}	f_{12}	$\cdots$	f_{1n}
	2	f_{21}	f_{22}	$\cdots$	f_{2n}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	n	f_{n1}	f_{n2}	$\cdots$	f_{nn}

2.2.3 Правила придруживања

У истраживању података учење правилима придруживања [JM01] је популаран и добро истражен метод за откривање интересантних односа између промењљивих у великим базама података. Метода правила придруживања служи за добијање правила придружених одређеној класи података из великих скупова података коришћењем различитих мера за одређивање интересантности добијених правила. Ово су правила типа "ако-тада". Она показују вероватноћу да један догађај уз себе повлачи други догађај. На пример ако се догоди A_1, A_2, ...A_n, тада се често догоди B_1, B_2, ...B_n. Први их је увео Ракеш Агравал са сарадницима ([AIS93]).

Нека је D база података, $X_D = \{i_1, i_2, \ldots, i_d\}$ скуп атрибута (догађаја) у бази D, а нека је $S_D = \{s_1, s_2, \ldots, s_n\}$ скуп слогова из базе D. Ради једноставности посматрајмо сваки слог базе као секвенцу сачињену од нула и јединица, дужине d. Свако поље секвенце представља један атрибут и има вредност 1 ако је атрибут садржан у слогу и 0 ако атрибут није садржан у слогу.

Правило придруживања карактерише мера у којој присуство датог скупа X атрибута у бази D подразумева присуство неког другог скупа посебних атрибута Y. Важна особина скупа атрибута је бројач подршке. Бројач подршке $\sigma(X)$ је број слогова базе који садрже скуп атрибута X (2.9). Математички би се то могло записати као:

$$\sigma(X) = |\{s_i | X \subseteq s_i, s_i \in S_D\}|. \tag{2.9}$$

Правило придруживања је импликативни израз облика $X \rightarrow Y$, где су X и Y дисјунктни скупови, $X \cap Y = \emptyset$. Јачина правила придруживања се рачуна на основу мера *подршке* и *поверења*. Подршка (*support-s*) показује колико често се правило појављује у датој бази података (2.10), док поверење (*confidence-c*) приказује колико често се скуп Y појављује у слоговма који садрже X (2.11).

$$s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{n} \tag{2.10}$$

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} \tag{2.11}$$

Мера која се често користи за оцењивање квалитета правила придруживања је лифт (*lift-l*). Лифт је уведен да би се умањио недостатак поверења као мере која игнорише подршку за атрибуте који се налазе на десној страни правила. За издвојено правило $X \rightarrow Y$ лифт показује колико пута више се заједно појављују ставке са леве и десне стране правила у односу на очекивани број појављивања у случају да су X и Y статистички независни. Лифт се рачуна као:

$$l = \frac{c(X \rightarrow Y)}{s(Y)}. \tag{2.12}$$

Дефиниција 2. *Проблем одређивања правила придруживања се може дефинисати на следећи начин: за дати скуп трансакција(слогова у бази) S_D пронаћи сва правила која имају подршку већу од min_s, и поверење веће од min_c где су min_s, min_c унапред дефинисани прагови за подршку и поверење.*

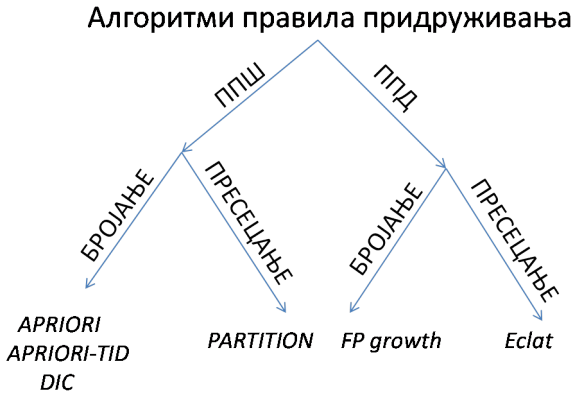
Правила придруживања указују на чињеницу колико се често неки атрибути појављују заједно. Ова метода се често користи код обраде података који прате неке трансакције (нпр. продају, набавку и слично). Такве информације могу бити коришћене као основа за одлуке о маркетиншким активностима као што

су промотивне цене или распоређивање производа. Правила придруживања су данас укључена у многе области као што су претраживање коришћења веба, детекције упада, биоинформатика, итд. Насупрот претраживању низова методе за одређивање правила придруживања обично не узимају у обзир редослед атрибута у трансакцији. Доказ да је (теоријски) могуће решење претходно дефинисаног проблема је дат у облику тзв. методе грубе силе која се састоји из следећих корака:

- издвојити сва могућа правила придруживања
- израчунати подршку и поверење за свако правило
- поткресати (искључити) правила која не задовољавају услов да имају подршку и поверење веће од претходно задатих вредности min_s и min_c.

Правила се сортирају по вредности поверења у опадајућем редоследу. Јачина правила је директно пропорционална његовом поверењу. Јака правила указују на велику повезаност појављивања скупова атрибута X и Y. Висока подршка правила указује да се оно није случајно појавило. У случају јако великих база min_c се поставља јако високо, на пример 80% док је уобичајен да је min_c значајно мање због велике разноликости података. Примена алгоритма грубе силе је рачунарски врло захтеван процес и због тога се не примењује у пракси у случајевима иоле већих база података.

За издвајање правила придруживања развијен је велики број алгоритама. Издвајање правила је са рачунарске стране јако захтеван посао. Зато се тежи смањењу: броја кандидата за правила, броја трансакција и броја поређења. Алгоритме делимо према начину претраге на оне који претрагу врше по ширини ППШ (*BFS -Breadth First Search*) и класу алгоритама која претрагу правила придруживања врши по дубини материјала ППД (*DFS -Depth First Search*). Према начину одређивања вредности подршке скупа атрибута, алгоритми се деле на бројачке (одређују број појава сваке секвенце) и оне који користе пресецање скупова (одређују кандидате на основу броја појава подсеквенци, ако подсеквенца има малу подршку онда ће је имати и одговарајућа секвенца). На слици 2.3 је приказана систематизација алгоритама по ове две поделе коју је у представио Хип (Hipp) ([HGN00]). Неки од најпознатијих алгоритама за одређивање правила придруживања су:



Слика 2.3: Систематизација алгоритама правила придруживања

Априори алгоритам(*Apriori algorithm*) је први дефинисан и представља најпознатији алгоритам за одређивање правила придруживања. Користи стратегију претраге по ширини за бројање скупова ставки и користи априори принцип за одбацивање кандидатских ниски које не задовољавају услове везане за праг подршке и поверења.

Алгоритам партиција (*Partition algorithm*) је заснован на априори алгоритму али користи пресецање скупова за одређивање вредности подршке.

Алгоритам ФП-раста (*FP - frequent pattern*) употребљава компримовану репрезентацију базе података помоћу ФП-дрвета. Када је ФП-дрво конструисано користи се рекурзивни приступ типа "подели па владај" ради проналажења честих скупова ставки.

Еклат (*ECLAT- Equivalence Class Transformation*) је алгоритам за претрагу по дубини и користи пресецање скупова за одређивање вредности подршке.

Једна од врста одређивања правила придруживања је проналажење секвенцијалних образаца. Истраживање секвенцијалних образаца је откривање подсеквенци које су заједничке за више од одређеног процента секвенци у бази података секвенци. О истраживању секвенцијалних образаца биће више у наредном поглављу.

2.2.4 Истраживање секвенцијалних образаца

Секвенце су уређене листе елемената било које врсте. У многим областима (трговини, лингвистици, географији, биологији,...) подаци су обично представљени у облику секвенце. Секвенце могу бити последица природног

временског редоследа између појединих података (историја куповања купца) или поштовања неке физичке структуре (гени у хромозомима). Секвенцијални образац је релативно кратка секвенца која се статистички значајно више (мање) пута појављује у одређеном скупу секвенци. Секвенцијална анализа, односно истраживање секвенцијалних образаца се развило као посебан правац у оквиру истраживања података. Истраживање секвенцијалних образаца се примењује на широку палету проблема почевши од редоследа прегледања веб страница до распореда амино киселина у протеинима, или пак са друге стране од оптерећења великих рачунарских система до природних непогода. Истраживање секвенцијалних образаца је трагање за заједничком подсеквенцом или честим обрасцем у датом скупу секвенци. Не може се на свим проблемима применити исти начин истраживања секвенцијалних образаца. Свака област захтева јединствен модел и решење. У даљем тексту дат је опис неких метода истраживања образаца које су нашле примену у биоинформатици. Основни појмови који се користе у овој методи су ниске (секвенце), подниске и секвенцијални обрасци.

Дефиниција 3. *Ниска је уређена листа елемената* $s =< e_1e_2e_3... >$. *Сваки елемент садржи скуп догађаја ставки* $e_i = \{i_1, i_2, ..., i_k\}$. *Сваком елементу се додељује одређено место или време. Број елемената ниске одређује дужину ниске, у ознаци* $|s|$. *Ниска која садржи* k *догађаја назива се* k*-ниска.*

Дефиниција 4. *Ниска* $a =< a_1a_2...a_n >$ *је садржана у другој нисци* $b =< b_1b_2...b_m >$ *где важи* $m \geq n$, *ако постоје цели бројеви* $i_1 < i_2 < ... < i_n$ *такви да важи* $a_1 \subseteq b_{i_1}, a_2 \subseteq b_{i_2}, \ldots, a_n \subseteq b_{i_n}$. *Тада се за ниску* a *каже да је подниска ниске* b.

Дефиниција 5. *Подршка ниске* a *је количник броја ниски које садрже* a *у односу на укупан број ниски.*

Дефиниција 6. *Секвенцијални образац је честа подниска (тј. подниска чија је подршка већа од унапред одређене минималне подршке* min_s*).*

Циљ истраживања секвенцијалних образаца је проналажење свих подниски у задатој бази ниски које имају подршку већу од унапред утврђене минималне подршке min_s. Велики број истраживача бави се изградњом модела

секвенцијалних образаца и ефективних алгоритама за истраживање ових модела. Већина модела може се сместити у једну од четири категорије: чести обрасци, периодични обрасци, статистички значајни обрасци, и приближни обрасци [WJ05].

Чести обрасци

Број појава образаца може бити значајна метрика у неким појавама. Ако се образац појављује често у неком скупу података то може бити значајно за карактеризацију датих података. Алгоритми истраживања честих образаца се разликују по: начину генерисања кандидата узорака, тестирању кандидата, израчунавању подршке,... Алгоритми за претраживање честих образаца заснивају се на априори принципу (*GSP, SPADE, SPAM*) или на принципу растућег обрасца (*FreeSpan, PrefixSpan*).

Дефиниција 7. *(Априори принцип за секвенцијалне обрасце) Нека је P_i скуп секвенци, P секвенцијални образац, и $s(P)$ подршка обрасца P у скупу P_i. Тада је $s(P) \leq min_{P' \subset P}(s(P'))$ где је P' подобразац од обрасца P.*

Истраживање секвенцијалних образаца увели су Агравал (*Agrawal*) и Срикант(*Srikant*). У раду [AS94] је изложен алгоритам издвајања образаца "Уопштени секвенцијални обрасци" (*Generalized sequential pattern- GSP*) заснован на априори принципу. Овај алгоритам полази од чињенице да свака секвенца дужине 1 може бити образац. Издвајају се секвенце које имају задовољавајућу подршку. Затим се понавља следећи поступак за сваки ниво све док има нових кандидата за обрасце:

- генерисање кандидата за обрасце дужине $k+1$ на основу образаца дужине k коришћењем априори принципа,
- одређује се подршка сваког кандидата за образац и одстрањују они чија је подршка мања од очекиване.

Заки (*Zaki*) је 2001. године унапредио претходни алгоритам и дефинисао алгоритам СПАДЕ (*SPADE*) у коме је редукован број пролазака кроз базу и убрзан процес применом ефикаснијих алгоритама претраге. Поред броја појава свака k-ниска носи са собом и листу објеката у којима се појављује тако да се

формирање образаца кандидата дужине $k+1$ врши преклапањем листи објеката у којима се појављују подобрасци дужине k [Zak01]. Пеи (*Pei*) и Хан (*Han*) су први увели алгоритме истраживања структурних образаца који се заснивају на особини растућих образаца (pattern growth property).

Дефиниција 8. *(Особина растућих образаца) Нека је α чест образац у бази D, и B део базе одређен са α, а β члан у B. Тада је $\alpha \cup \beta$ чест образац у D ако и само ако је β чест образац у B.*

Алгоритам Префикспан (*Prefixspan*) тражи честе подниске P а затим тражи оне обрасце којима је P префикс [Pei+01]. Овим алгоритмом се у првом кораку издвају префикси дужине 1 па се скуп секвенци дели на основу префикса који садрже. Даље се сваки подскуп секвенци (формиран по k-ниски) дели на подскупове на основи проширених префикса дужине $k + 1$. У КлоСпан алгоритму (*CloSpan- Mining Closed Sequential Patterns*) се смањује број образаца без губитка изражајне снаге алгоритма [XJR03]. За смањење броја правила користи се варијанта Априори принципа: у преосталом скупу образаца ни за један образац α не постоји образац β који га садржи и има исту подршку као α. У претходним алгоритмима посматрана је учесталост неког обрасца у подацима али није вођено рачуна где се у податку дати образац појављује. У неким областима учесталост појављивања неког обрасца није битна, више су значајани обрасци који се јављају са неком правилношћу.

Регуларни обрасци

Регуларни обрасци се деле на цикличне и периодичне. Циклични обрасци су представљени у [ORS98] као циклична правила придруживања. Правило придруживања има циклус (l, o) ако се у интервалу дужине l увек јавља на позицији o. Образац је периодичан када се низ карактера појављује у целом свом саставу периодично у некој секвенци, на пример у секвенци $< abcabc >$ $< abc >$ је периодични образац. Периодични обрасци се деле на потпуне и на делимичне. Потпуно периодични су они обрасци који се јављају узастопно у свом целокупном саставу, на пример $< bc >$ је потпуно периодични образац у $< abcbcbc >$. Делимично периодични обрасци се јављају периодично али не у целокупном саставу. На пример $< a * c >$ је делимично периодичан образац за

$< aacabcacc >$ [HDY99].

Обрасци који се појављују периодично без и једног изостанка називају се синхрони обрасци. Асинхрони обрасци су они који се не појављују једни за другим и подједнако у свим деловима. Нека је *minrep*-минималан број узастопних понављања и *maxdis*-максимално растојање између два суседна обрасца, тада кажемо да је валидан сегмент секвенце онај који садржи најмање *minrep* понављања обрасца и задовољава услов да је минимално растојање између суседних образаца *maxdis*.

За модел $A = a_1, a_2, a_3, a_4, a_5, a_6, \ldots$ образац *so* дужине l је $so =< o_1 o_2 o_3 o_4 \ldots o_l >$ где је o_i или * или је из скупа A. Специјализација обрасца је замена * конкретним елементом из A, док је генерализација постављање * уместо неког o_i. Априори принцип важи и код периодичних образаца са истим периодом. Ако је валидан сегмент за неки образац онда је он валидан сегмент и за образац са мање одређених елемената (генералнији образац). На пример валидан сегмент за образац $< a_1 a_2 * >$ је валидан и за $< a_1 * * >$.

Мета-периодични обрасци. Основни обрасци састављени од симбола су само један специјални случај мета података. Код мета података може било где доћи до шума чија варијација трајања задовољава постављене границе. Флексибилност мета образаца отежава процес њиховог откривања. У овом случају фреквенција појављивања обрасца није добра мера за битност узорка.

Статистички значајни обрасци

У многим областима појављивање симбола у секвенцама није равномерно па се често неки симболи појављују значајно чешће од других. Отуда је очекивано да обрасци који садрже симболе који се чешће појављују имају већу шансу да се појаве у секвенци. Појављивање образаца у великој мери није занимљиво ако је то очекивана појава; од интереса је појављивање образаца у већем броју од очекиваног. Брин (*Brin*) је у [BMS97] први дошао до неких значајних образаца применом корелације. За издвајање образаца коришћене су статистичке методе хи-квадрат тест, хеш табеле, и друге [Coh+00][FUM00].

Неизвесност појављивања обрасца је обрнуто пропорционална вероватноћи да се образац појави. Ако се на основу претходног искуства зна да је велика вероватноћа појављивања неког обрасца онда је неизвесност његовог

појављивања мала и његово појављивање није значајно. Што је неизвесност већа то је информација о појављивању обрасца значајнија. Информативност обрасца $so, so \in E$ са вероватноћом појављивања $Prob(so)$ дефинисана је као

$$I(so) = -log_{|E|}Prob(so) \tag{2.13}$$

где $|E|$ представља укупан број могућих образаца у скупу E. Вероватноћа појављивања обрасца може се одређивати на више начина: као равномерна расподела у ком случају је $Prob(so) = |E|^{-1}$, на основу експерименталних резултата као однос броја појава обрасца so у скупу A или на основу неких теоријских очекивања. Претходне методе које су значајност образаца мериле на основу његове подршке у секвенцама подржавале су априори принцип по коме ако је образац значајан онда је и његов подобразац значајан. То није случај са моделима које значајност обрасца мере његовом информативношћу. Јанг(*Yang*) је представио алгоритам који истражује ретке али статистички значајне обрасце (алгоритам InfoMiner, [YWY03]). У овом алгоритму се информативност множи са подршком умањеном за један тако да обрасци који се никада нису поновили нису од интереса. Позиција на којој се образац појављује се не узима у обзир. У неким областима узастопно појављивање образаца је значајније од изолованог појављивања, на пример код понављајућих секвенци у ДНК. Зато је претходни алгоритам допуњен тако да прави разлику између узастопних и разбацаних појављивања образаца (алгоритам InfoMiner+, [YWY02]). Годину дана касније представљен је алгоритам (*STAMP*) [YWY01] који издваја значајне обрасце и низове одговарајућих подсекванци у којима се појављују.

Приближни обрасци

Непрецизност, шумови, одступања су саставни део реалног света па се и истраживање образаца мора усмерити ка приближним обрасцима. Истраживање приближних образаца допушта апроксимације у појављивању као и у структури образаца. У том случају се претражују образци који се не појављују у потпуном саставу већ уз делимичне измене [PTH01],[YFB01].

2.2.5 Истраживање образаца у биоинформатици

Повезивањем информатике и биологије велика количина биолошких података постала је јавно доступна. Многи истраживачи су техникама и алатима истраживања података долазили до корисних информација из велике количине биолошких података. Развијен је велики број програмских система који могу да се примене и за истрживање биолошких података (*IBM Intelligent Miner, Weka, SPSS, KNIME, SGI MineSet, SAS® Enterprise Miner™*,...). Постоји пуно простора за анализу биолошких података и развијање нових метода истраживања података прилагођених биолошким подацима [Wan+05]. Нека од поља истраживања су:

- анализа честих образаца, секвенцијалних образаца или структурних образаца (идентификација копојављивања и корелације биосеквенци или биоструктура),
- класификација и поређење биолошких података,
- разне методе кластер анализе биолошких података,
- рачунарско моделирање биолошких мрежа,
- визуелизација података и визуелно истраживање података.

Поређење секвенци, претраживање сличности и проналажење образаца сматрају се основним приступима секвенцијалне анализе биолошких ниски. Математички модели и основни алгоритми анализе секвенци датирају још из 1960. године када су развијене методе за предвиђање филогенетских односа протеина. У претходном поглављу су описани различити алгоритми за проналажење честих образаца. Међутим нису сви они погодни за анализу биолошких ниски јер они траже правилне обрасце што често није ситуација код биолошких ниски. Тако на пример ДНК секвенцијални обрасци често дозвољавају уметање и брисање неких делова ниске и друге шумове. Откривање честих образаца у биолошким нискама или њиховим тродимензионим структурама је јако значајно за разумевање многих биолошких процеса. Развијени су многи специјализовани алати за претрагу честих образаца (као нпр. већ поменути БЛАСТ).

Тандем рипит (*Tandem repeat - TR*) је често истраживан облик образаца. То су подниске које се јављају више од одређеног броја пута у ДНК ниски. Биолози сматрају да обрасци који се појављују више пута у ДНК ниски имају неко специјално значење. Методе за истраживање ових образца дозвољавају да се они појављују уз одређене измене у ДНК ниски (могу бити нешто краћи, они у којима недостају неки нуклеотиди, односно дужи са убаченим нуклеотидима као шумовима). Направљени су многи алгоритми за тражење оваквих понављајућих секвенци (*Perfect Tandem Repeat - PTR, Approximate Tandem Repeat - ATR, Reputer, Trfinder*). Протеини су дугачке ниске и проналажење заједничког мотива у различитим протеинима је јако корисно. Мотиви се у различитим протеинима могу појављивати уз одређене измене.

Поглавље 3

Модел за одређивање карактеристичних n-грама за ЦОГ-ове протеина

3.1 Модел за одређивање карактеристичних аминокиселинских подниски у протеинима

3.1.1 Мотивација

Као што је раније поменуто истраживања секвенцијалних образаца може се применити и на биолошким секвенцама. Протеини су уређени скупови амино киселина. Амино киселине могу бити посматране као симболи а самим тим и протеини као ниске симбола. Међу најпогоднијим методама за анализу биолошких секвенци су оне које користе n-грамске технике.

Дефиниција 9. *Нека је дата ниска* $x =< a_1 \dots a_{k-1} a_k >$ *дужине* k *дефинисана над датом азбуком* A *кардиналности* $|A| = r$. *n-Грам ниске* x *је подниска ниске* x *дужине* n, *односно сегмент* $n (n \leq k)$ *узастопних симбола из ниске* x.

Уобичајено је да се n-грами дужине 1, 2, 3 и 4 називају моногрми, биграми, триграми и тетраграми, док се за остале вредности n једноставно називају n-грами. Број различитих n-грама L над азбуком A кардиналности r је једнак броју варијација са понављањем n-те класе од r елемената, тј. $L = r^n$. Свака ниска може да се представи као вектор n-грама који се у њој појављују. Записи ниски у облику вектора могу да се користе за њихово поређење.

Дефиниција 10. *n-Грами који се генеришу померањем оквира дужине n дуж секвенце називају се преклапајући n-грами. Померање се врши позицију по позицију, односно симбол по симбол.*

Из секвенце x чија је дужина k може се издвојити $k - n + 1$ преклапајућих n-грама ($< a_1 \dots a_{n-1}a_n >, < a_2 \dots a_n a_{n+1} >, \dots, < a_{n-k+1} \dots a_{k-1}a_k >$). У анализи помоћу n-грама користе се вероватноћа појављивања n-грама у ниски, релативна фреквенција различитих n-грама, као и многе друге статистичке особине n-грама. Анализом садржаја n-грама, могу се генерисати правила о повезаности појављивања неког n-грама и карактеристика ниски.

Поред примене у биоинформатици описане у уводу, анализа n-грама се употребљава за компресију текстуалних података [Wis87], аутоматизовану категоризацију и предвиђање текстуалних података [CT94; AZ07], идентификацију језика на коме су писана документа [Sch91], корекцију грешака при писању [ZPZ81; PS12], придруживање дела аутору [KPC03], издвајање и предвиђање карактеристика система [Mut10], издвајање кључних термина, итд.

Протеини су ниске дефинисане над 20 слова[1] (амино киселина) азбуке A_{ak}={A,E,Q,D,N,L,G,K,S,V,R,T,P,I,M,F,Y,C,W,H}. Број могућих n-грама над азбуком A_{ak} је 20^n. Тако на пример за $n = 2$, број могућих биграма је $L = 20^2 = 400$ и скуп биграма је {<AA>, <AE>, <AQ>, <AD>, <AN>, <AL> , . . . , <HY>, <HC>, <HW>, <HH>}. За секвенцу X = <ANEQALEGGKTГ> дужине $k = 12$, и $n = 4$ постоји $k - n + 1 = 9$ преклапајућих тетраграма у њој:

```
ANEQALEGGKTG
ANEQ
 NEQA
  EQAL
   ...
```

Анализа преклапајућих n-грама свих протеина рађена је са циљем да се одреде n-грами који карактеришу протеине из одређене ЦОГ категорије.

[1]Као што је претходно речено протеини су ниске у којима могу да се јаве 23 амино киселине, али се пиролизин (pyrrolysine), селеноцистеин (selenocysteine) і H-формилметионин (N-Formylmethionine) јако ретко појављују у протеинима. Због тога ће комплетан опис методе бити дат у односу на претходно поменутих 20 амино киселина, али се лако може проширити и за ове три додатне амино киселине.

Алгоритамска сложеност овог процеса је $O(k, n, num_p)$ где је n дужина n-грама, k дужина протеина и num_p број протеинских секвенци.

Главна идеја у конструкцији модела за проналажење карактеристичних n-грама везаних за категорије ЦОГ-ова је издвајање преклапајућих n-грама a_i одговарајуће дужине n за сваки протеин из скупа података за тренирање модела, пребројавање њиховог појављивања и одређивање значајних n-грама за сваку ЦОГ категорију. Значајност појављивања n-грама a_i у ЦОГ категорији COG_k је изражена помоћу три мере:

- бројача подршке (σ) који представља број појава n-грама a_i у ЦОГ категорији COG_k (апсолутна подршка n-грама $\#a_i$),
- подршке (s) која је једнака проценту броја протеина у ЦОГ категорији COG_k који садрже n-грам a_i, и
- поверења (c) које је представљено процентом броја протеина који припадају ЦОГ категорији COG_k и садрже n-грам a_i у односу на број свих протеина који садрже n-грам a_i.

Сваки n-грам може да се карактерише уређеном петорком $(a_i, n, \#a_i, s(a_i), c(a_i))$. Користећи ову карактеристику издвајају се n-грами који имају подршку и поверење веће од унапред дефинисаних min_s и min_c као n-грами карактеристични за одговарајућу ЦОГ категорију. Провера конструисаног модела се врши упаривањем пронађених карактеристичних ниски са тест подацима.

Премда није толико компликована, предложена метода има озбиљне недостатке. Број могућих n-грама дужине 1, 2, ..., n је 20, 400, ..., 20^n, респективно. Обрада и складиштење ових података врло брзо превазилази расположиве техничке могућности, траје дуго и не даје гаранцију да ће овај процес бити ефективан за већи скуп података. То је главни разлог за увођење трансформације n-грама у предложеном алгоритму за издвајање секвенцијалних образаца. Трансформација n-грама се може посматрати као специјални вид димензионалне редукције који омогућава издвајање карактеристичних n-грама из скупа преклапајућих n-грама.

3.1.2 Опис методе и имплементација

Трансформација n-грама

Сваки подскуп азбуке A_{ak} може бити представљен као ниска у којој су амино киселине поређане по утврђеном распореду амино киселина. Тада на пример подскупови {A, E} и {E, A} су исти подскупови и представљају се ниском <AE>.

Дефиниција 11. *Ниске различитих амино киселина у уређеном распореду, чији елементи представљају подскупове азбуке A_{ak} називамо основне аминокиселинске ниске (OAK ниске).*

Постоји бијекција основних аминокиселинских ниски у елементе партитивног скупа A, $\mathbb{P}(A)$. Како је $|\mathbb{P}(A)| = 2^{|A|}$, то је кардиналност скупа основних аминокиселинских ниски једнак 2^{20}. Нека је $x =< x_1 \dots x_{k-1}x_k >$ ниска дужине k, и $S(x) = \{s_1, \dots, s_m\}$ где је $m \leq k$ скуп различитих елемената азбуке A који се појављују у x. На пример ако је

$$x =< AEETKASW >$$

тада је

$$S(x) = \{A, E, T, K, S, W\}.$$

Ниска x може да садржи неку амино киселину (s_i) више од једног пута. Нека је $num(s_i)$ број појављивања амино киселине s_i у x. Тада важи $|x| = num(s_1) + \dots + num(s_m)$. На пример за ниску $x =< AEETKASW >$ важи $num(A) = 2$, $num(E) = 2$, $num(T) = 1$, $num(K) = 1$, $num(S) = 1$, $num(W) = 1$. Нека је so секвенцијални образац ниски над азбуком A дефинисан подскупом азбуке $s \subseteq A$ и дужином n. Образац so се појављује у некој ниски $y = y_1 \dots y_k$ на позицији i ако важи да је $s = S(< y_i \dots y_{i+n-1} >)$, односно да се подниска дужине n почевши од позиције i састоји искључиво од елемената скупа s. Сваком подскупу s скупа A одговара тачно једна OAK ниска s'. Образац so је одређен паром (s', n). На пример, образац so дефинисан паром $(< AEKSTW >, 8)$ се у ниски $y =< BAEETKASWQAEFWAEEETKSM >$ појављује два пута:

$$y =< B\boxed{\text{AEETKASW}}QAEF\boxed{\text{WAEEETKS}}M > .$$

Статистичка значајност обрасца зависи од броја могућих секвенци које

задовољавају дати образац. Одговор на ово питање је дат у следећој теореми.

Теорема 1. *Нека је so секвенцијални образац дефинисан паром* (s', n) *при чему је* $s' =< s_1 \dots s_k >$ *ОАК ниска,* n *природан број, и нека ниска* x *задовољава секвенцијални образац so акко* $|x| = n$ *и* $OAK(x) = s'$ *при чему је ОАК функција која пресликава аргумент ниску у одговарајућу ОАК ниску. Нека је са* $num(s_i)$ *означен број појављивања амино киселине* s_i *у ниски* x*; тада за свако* s_i *важи да је* $num(s_i) \geq 1$ *. Даље важи:*

1. *Број различитих k-торки* $(num(s_1), \dots, num(s_k))$ *је* $\binom{2n-k+1}{n-k}$ *за* $k < n$ *и 1 иначе, где је к дужина ОАК ниске* s'*;*

2. *Број могућих секвенци дужине n које задовољавају секвенцијални образац so једнак је*

$$\sum_{\substack{num(s_1)+\dots+num(s_k)=n \\ (\forall i)(num(s_i)\geq 1)}} \frac{n!}{num(s_1)! \cdots num(s_k)!}. \tag{3.1}$$

Доказ. За разлику од обичних пермутација број појављивања амино киселина $\{s_1, \dots, s_k\}$ у секвенцама које подржавају образац *so* није одређен јединствено већ само условима $num(s_1) + \dots + num(s_k) = n$ и $(\forall i)(num(s_i) \geq 1)$.

1. Стога на k позиција у ниски x морају бити постављене све амино киселине $\{s_1, \dots, s_k\}$ тако да је почетна вредност за k-торку $(num(s_1), \dots, num(s_k))$ једнака $(1, \dots, 1)$. Осталих $n - k$ позиција одређују коначну вредност k-торке је $(num(s_1), \dots, num(s_k))$. Број различитих могућности за k-торку $(num(s_1), \dots, num(s_k))$ једнак је броју комбинација са понављањем од k елемената реда $n - k$, односно $\binom{2n-k+1}{n-k}$ за $k < n$ и 1 иначе.

2. За сваку k-торку $(num(s_1), \dots, num(s_k))$ број могућих n-грама ниске x у којима се елементи скупа $\{s_1, \dots, s_k\}$ појављују $num(s_1), \dots, num(s_k)$ пута респективно добија се као број пермутација од n елемената са понављањем и износи:

$$\frac{n!}{num(s_1)! \cdots num(s_k)!}.$$

Укупан број могућих секвенци дужине n које задовољавају секвенцијални образац *so* једнак је збиру по свим могућим распоредима АК $s_1, \dots, s_k$ у ниски x за које важи да је $num(s_1) + \dots + num(s_k) = n$

□

Пресликавањем функцијом *OAK* број различитих преклапајућих аминокиселинских n-грама који се се јављају у протеину се замењује бројем OAK ниски у које су одговарајући n-грами пресликани и тиме се значајно смањује димензија вектора који описује састав n-грама у протеину. Тада се сваки n-грам a_i пресликава у једну OAK ниску s' која је еквивалентна скупу $s = S(a_i)$ чија је кардиналност од 1 до n у зависности од тога колико различитих амино киселина садржи a_i. Број различитих подскупова скупа A од $n, n-1, \ldots, 1$ елемената (L_{OAK}) је:

$$L_{OAK} = \binom{20}{n} + \binom{20}{n-1} + \cdots + \binom{20}{1} \tag{3.2}$$

јер сваки скуп од r елемената поседује тачно $\binom{r}{n} = \frac{r!}{(r-n)!n!}$ различитих подскупова који имају n елемената. На пример за $n = 3$ број могућих преклапајућих n-грама је 8000 док број могућих OAK ниски је $\binom{20}{3}+\binom{20}{2}+\binom{20}{1} = 1350$.

Теоријска основа методе

Над скупом OAK ниски можемо дефинисати Булову (*George Boole*) алгебру [Rud74]. Нека је дат скуп B са најмање два елемента $\{0, 1\}$ на коме су дефинисане две бинарне операције, у ознаци $\cup$ и $\cdot$, и једна унарна операција, у ознаци $'$.

Дефиниција 12. *На скупу B је дефинисана Булова алгебра у ознаци $(B, \cdot, \cup, ', \mathbf{0}, \mathbf{1})$, ако за све $a, b, c \in B$ важе следеће аксиоме:*

- $a \cup b = b \cup a$ *;* $a \cdot b = b \cdot a$ *(комутативност),*
- $(a \cup b) \cup c = a \cup (b \cup c)$ *;* $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ *(асоцијативност),*
- $c \cup (a \cdot b) = (c \cup a) \cdot (c \cup b)$ *;* $c \cdot (a \cup b) = (c \cdot a) \cup (c \cdot b)$ *(дистрибутивност),*
- $(a \cup \mathbf{0}) = a$ *;* $(a \cdot \mathbf{1}) = a$ *(неутрални елемент),*
- $(a \cup a') = \mathbf{1}$ *;* $(a \cdot a') = \mathbf{0}$ *(комплемент).*

У Буловој алгебри $(B, \cdot, \cup, ', \mathbf{0}, \mathbf{1})$ можемо увести бинарну релацију $\leq$ ("мање или једнако") као:

Дефиниција 13. *За свако $x, y \in B$, $x \leq y$ акко $x \cup y = y$, и кажемо да y садржи x.*

Лема 1. *У Буловој алгебри $(B, \cdot, \cup, ', \mathbf{0}, \mathbf{1})$ за свако $x, y \in B$ важи $x \leq y$ акко $y' \cdot x = \mathbf{0}$.*

Ако је $B = \mathbb{P}(A_{ak})$, тада је $(B, \cdot, \cup, ', \mathbf{0}, \mathbf{1})$ Булова алгебра где је $\mathbf{0}$ празна ОАК ниска дужине 0, а $\mathbf{1}$ је ОАК ниска која садржи све амино киселине тј. скуп A. Операција уније ($\cup$) примењена на две ОАК ниске представља ОАК ниску која садржи све амино киселине из оба операнда, док операција пресека ($\cdot$) даје ОАК ниску у којој се налазе амино киселине које се појављују у оба операнда. По договору, знак $\cdot$ може бити изостављен. Комплемент ($'$) даје ОАК ниску која се састоји од свих амино киселина из A које нису садржане у операнду. Бинарна операција $\leq$ примењена на две ОАК ниске је тачна ако су амино киселине прве ОАК ниске све садржане у скупу амино киселина друге ОАК ниске.

Циљ истраживања је издвајање секвенцијалних образаца аминокиселинских ниски који су карактеристични за само једну ЦОГ категорију. Како секвенцијални обрасци могу бити представљени као ОАК ниске, треба издвојити ОАК ниске које се налазе у протеинима само једне ЦОГ категорије. Нека су b_j ($j \in 1, \ldots, k_1$) ОАК ниске издвојене из протеина који припадају ЦОГ категорији COG_k, а e_i ($i \in 1, \ldots, k_2$) ОАК ниске који су издвојене из протеина који припадају другим ЦОГ категоријама различитим од категорије COG_k, и нека је x тражени секвенцијални образац категорије COG_k. Тада је x садржано у b_j бар за једно j, што може да се запише у Буловој алгебри као $(\exists j)(x \leq b_j)$ или као еквивалентан израз $(\exists j)(b_j' \cdot x = \mathbf{0})$. Како образац треба да буде карактеристика само категорије COG_k, то x не сме бити садржано у e_i ни за једно i ($i \in 1, \ldots, k_2$), односно важи израз $\neg(\exists i)(e_i' \cdot x = \mathbf{0})$, или еквивалентан израз $(\forall i)(e_i' \cdot x \neq \mathbf{0})$.

Дефиниција 14. *Пресликавање f скупа B^n (где је B Булова алгебра) у скуп B у ознаци*

$$f : B^n \to B$$

се назива Булова функција.

Свака Булова функција се може приказати помоћу Буловог израза.

Теорема 2. *Функција $f : B^n \to B$ је Булова ако и само ако може бити написана у канонској дисјунктивној форми*

$$f(X) = \bigcup_A f(A)X^A \tag{3.3}$$

где је $X = (x_1, \ldots, x_n) \in B^n$, $A = (a_1, \ldots, a_n) \in \{0,1\}^n$, $X^A = x_1^{a_1} \cdots x_n^{a_n}, x_i^0 = x_i'$ и $x_i^1 = x_i$.

Нека је $f : B^n \to B$ Булова функција. Релација $f(X) = \mathbf{0}$ се назива Булова једначина.

Претходне теореме наведене су без доказа; њихов доказ се може наћи у [Rud74]. Такође, из истог извора су преузете и претходно наведене дефиниције.

Дефиниција 15. *Генерализовани системи Булових једначина у ознаци ГСБЈ (The generalized systems of Boolean equations - GSBE) над Буловом алгебром дефинисани су рекурзивном формом:*

(i) свака Булова једначина $f(X) = \mathbf{0}$ је ГСБЈ;

(ii) негација, логичка конјункција и логичка дисјункција било које ГСБЈ је ГСБЈ;

(iii) свака ГСБЈ се добија применом (i) и (ii) коначано много пута.

Проблем решавања ГСБЈ се своди на решавање њених појединачних случајева.

Дефиниција 16. *Елементарна ГСБЈ је или Булова једначина $f(X) = 0$ или систем у форми:*

$$f_1(X) \neq \mathbf{0} \ \wedge \cdots \wedge \ f_k(X) \neq \mathbf{0} \tag{3.4}$$

или

$$g(X) = \mathbf{0} \wedge f_1(X) \neq \mathbf{0} \ \wedge \cdots \wedge \ f_k(X) \neq \mathbf{0}. \tag{3.5}$$

Ако је k=1 кажемо да је ГСБЈ атомска. Атомска ГСБЈ у форми $f(X) \neq \mathbf{0}$ зове се Булова неједначина.

Теорема 3. *Скуп решења сваке ГСБЈ је унија скупа решења елементарних ГСБЈ.*

У складу са претходним дефиницијама, теоремама и Буловим изразима који одређују ниску x која представља секвенцијални образац карактеристичан за одређену ЦОГ категорију, x се може представити као решење генерализованог система Булових једначина:

$$(b'_1 x = \mathbf{0} \vee \cdots \vee b'_{k1} x = \mathbf{0}) \wedge (e'_1 x \neq \mathbf{0} \wedge \cdots \wedge e'_{k2} x \neq \mathbf{0}). \tag{3.6}$$

Решење овог система представљају ОАК ниске дужине од 1 до n које су садржане у пројекцијама преклапајућих n-грама дужине n у ОАК ниске, при чему се такве ниске јављају само у протеинима категорије COG_k. Проблем решавања генерализованих система Булових једначина није у потпуности решен. Неки резултати се могу наћи у [Rud74; Rud01]. Банковић је дао сва решења за: Булове неједначине [Ban07], систем Булове једначине и Булове неједначине [Ban10], као и за систем две Булове неједначине [BM15]. Маровац је разматрала системе од k Булових неједначина у [Mar15] и дисјункцију Булових једначина у [Mar]; добијени резултати су приказани у теоремама које следе.

Дефиниција 17. *Нека су* $f, F_1, \ldots, F_n : B^n \to B$ *Булове функције и* $F = (F_1, \ldots, F_n)$. *Формула*

$$X = F(T),$$

или у скаларној форми

$$x_i = F_i(t_1, \ldots, t_n) \quad (i = 1, \ldots, n),$$

представља опште решење Булове једначине $f(X) = 0$ *акко за свако* $X \in B^n$

$$f(X) = 0 \Leftrightarrow (\exists T) X = F(T).$$

Теорема 4. *[Mar15] Нека су* $f_1, \ldots, f_k : B^n \to B$ *Булове функције. Тада*

$$f_1(X) \neq \mathbf{0} \wedge \cdots \wedge f_k(X) \neq \mathbf{0} \Leftrightarrow$$

$$(\exists p_1) \cdots (\exists p_k)(\exists T)(p_1 \neq \mathbf{0} \wedge \cdots \wedge p_k \neq \mathbf{0} \wedge X = \Phi(p_1, \ldots, p_k, T)$$

$$\wedge \prod_A f_1(A) \leq p_1 \leq \bigcup_A f_1(A)$$

$$\wedge\, p_1 \prod_A (f'_1(A) \cup f_2(A)) \cup p'_1 \prod_A (f_1(A) \cup f_2(A))$$

$$\leq p_2 \leq p_1 \bigcup_A (f_1(A) f_2(A)) \cup p'_1 \bigcup_A (f'_1(A) f_2(A))$$

$$\ldots$$

$$\bigcup_{C_{k-1}\in\{0,1\}^{k-1}} {p_1}^{c_1}\cdots{p_{k-1}}^{c_{k-1}}\prod_A(f_1^{c'_1}(A)\cup\cdots\cup f_{k-1}^{c'_{k-1}}(A)\cup f_k(A))$$
$$\leq p_k \leq \bigcup_{C_{k-1}\in\{0,1\}^{k-1}} {p_1}^{c_1}\cdots p_{k-1}^{c_{k-1}}\bigcup_A(f_1^{c_1}(A)\cdots f_{k-1}^{c_{k-1}}(A)f_k(A))),$$

где $X=\Phi(p_1,\ldots,p_k,T)$ *представља опште решење једначине*

$$(f_1(X)+p_1)\cup\ldots\cup(f_k(X)+p_k)=0.$$

Теорема 5. *[Mar] Нека су* $f_1,\ldots,f_k : B^n\to B$ *Булове функције. Тада*

$$f_1(X)=\mathbf{0}\vee\cdots\vee f_k(X)=\mathbf{0}\Leftrightarrow$$
$$(\exists T)(\exists s_1)\cdots(\exists s_{k-1})\big((s_i=0\vee s_i=1)\wedge\cdots\wedge(s_{k-1}=0\vee s_{k-1}=1)\wedge$$
$$X=s'_1\Phi_1(T)\cup\cdots\cup s_1\cdots s_{k-2}s'_{k-1}\Phi_{k-1}(T)\cup s_1\cdots s_{k-1}\Phi_k(T)\big)$$

где за свако $i\in\{1,\ldots,k\}$, $\Phi_i(T)$ *представља опште решење једначине* $f_i(X)=0$.

Применом ових теорема системи ($b'_1x=\mathbf{0}\vee\cdots\vee b'_{k1}x=\mathbf{0}$) и ($e'_1x\neq\mathbf{0}\wedge\cdots\wedge e'_{k2}x\neq\mathbf{0}$) се могу трансформисати у једначине. На тај начин се систем (3.6) своди на систем од две једначине чије је решавање тривијално. Решења система (3.6) су секвенцијални обрасци који карактеришу једну ЦОГ категорију.

Имплементација

Програм за издвајање ОАК секвенци и одређивање образаца каратеристичних за одређену ЦОГ категорију имплементиран је у програмском језику *C*. Улазни скуп података чине протеинске секвенце преузете из НЦБИ банке података. НЦБИ база протеинских секвенци (http://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/, датотеке означене суфиксом .ptt) је колекција секвенци из више извора ГенБанк, РефСик, ..., као и евиденција из Свиспрот, ПДБе, ... Информације о филогенетској класификацији протеина налазе се у ЦОГ бази података (http://www.ncbi.nlm.nih.gov/COG/), генерисане поређењем нових протеина и већ класификованих протеинских секвенци кодираних у комплетним геномским секвенцама прокариотских организама. ЦОГ база података садржи информације и о припадности протеина ЦОГ категоријама (кластерима ЦОГ-ова). За формирање улазног скупа података реализовани су програми

за издвајање протеинских секвенци, као и информација о припадајућим ЦОГ категоријама протеина из одабране групе прокариотских организама (бактерија и археа). Сваки улазни податак у моделу садржи инфомације о:

- идентификацији протеина:
  - ⋆ ГИ (The GI number, GenInfo Identifier) - представља јединствен низ цифара који је додељен свакој секвенци,
  - ⋆ запис о обради од стране НЦБИ, и
  - ⋆ ПИД (PID - Pathway Interaction Database, http://pid.nci.nih.gov), ознаку за ГИ протеинских секвенци у ГенБанк.
- организму коме припада протеин;
- комплетној аминокиселинској секвенци протеина.

Из сваке аминокиселинске ниске издвојени су преклапајући n-грами и пресликани у ОАК ниске. Ради лакше обраде података помоћу рачунара ОАК ниске су представљене као бинарне речи дужине 20, тј. помоћу 20-цифрених бинарних бројева. Свака позиција у бинарној речи дужине 20 редом од најмање тежине до највеће представља по једну амино киселину из скупа A (1 представља присуство, а 0 одсуство те амино киселине). Тако је амино киселина:

- A представљена са константом $A = (2^0)_{10} = (00000000000000000001)_2$,
- E представљена са константом $E = (2^1)_{10} = (00000000000000000010)_2$,
- Q представљена са константом $Q = (2^2)_{10} = (00000000000000000100)_2$,
- D представљена са константом $D = (2^3)_{10} = (00000000000000001000)_2$,
- ...
- W представљена са константом $W = (2^{18})_{10} = (01000000000000000000)_2$,
- H представљена са константом $H = (2^{19})_{10} = (10000000000000000000)_2$.

Тада се на пример ОАК секвенца <AQDH> представља са 20-цифреним бинарним бројем $(10000000000000001101)_2$.

Булове операције унија и пресек две ОАК ниске су имплементиране помоћу логичких операција на нивоу бита "или" (|) и "и" (&) програмског језика *C*, респективно. Тако на пример нека су дати скупови амино киселина a и b представљени ОАК нискама $a =$ <AQDH> и $b =$ <AEDW>. Тада је :

$$\begin{aligned} a = &\ (10000000000000001101)_2 \\ b = &\ (01000000000000001011)_2 \\ a|b = &\ (11000000000000001111)_2 =< AEDWH > \\ a\&b = &\ (00000000000000001001)_2 =< AD > \end{aligned}$$

Комплемент ОАК ниске се добија одузимањем 20-цифреног бинарног броја којим је представљена ОАК ниска од бинарног еквивалента броју $(2^{20} - 1)_{10}$ који се састоји од 20 јединица и представља скуп од свих 20 амино киселина (**1**). Празан скуп се представља нулом.

Статистички модел

Скуп решења система (3.6) представља све секвенцијалне обрасце који карактеришу једну ЦОГ категорију, па и оне који су се случајно појавили и немају више појава у ЦОГ категорији. Секвенцијалном обрасцу треба придружити меру колико је он значајан за протеине одговарајуће категорије. Издвајање ОАК ниски које представљају секвенцијалне обрасце за одређену ЦОГ категорију може бити извршено и коришћењем статистичких мера. За неки n-грам a_i одговарајућа ОАК ниска b_j се добија као вредност функције OAK која пресликава аргумент у ОАК ниску дужине од 1 до 20, зависно од броја различитих амино киселина које су садржане у a_i. Значајност појављивања ОАК ниске b_j у ЦОГ категорији COG_k је изражена преко подршке, поверења и бројача подршке у поређењу са предефинисаним вредностима $(min\_s, min\_c, min\_\sigma)$. Бројач подршке је апсолутна подршка ОАК ниске, односно број n-грама a_i ($\#a_i$) пронађених у обрађеним протеинима таквих да је $OAK(a_i) = b_j$. Подршком је изражен проценат протеина у којима се b_j појавио у односу на укупан број протеина у категорији COG_k, док поверење представља

проценат протеина који припадају категорији COG_k из скупа протеина који садрже ОАК ниску b_j. ОАК секвенце које имају поверење и подршку већу од постављених прагова за ове вредности су кандидати за секвенцијалне обрасце за одговарајућу ЦОГ категорију. Секвенцијални обрасци који карактеришу само једну појединачну ЦОГ категорију COG_k, су такође и решења система (3.6) и називамо их *дескриптори*.

Дефиниција 18. *Дескриптор d_{jkn} је дефинисан једначином $d_{jkn} = b_j$ ако постоји четворка $T_{jkn} = (n, COG_k, b_j, \#a_i)$ таква да је поверење од b_j једнако 100%, а подршка већа од унапред дефинисане вредности min_s. За сваки дескриптор d_{jkn} дефинише се скуп D_{jkn} као $D_{jkn} = \{a_i | b_j = OAK(a_i) \wedge d_{jkn} = b_j\}$, којим је представљен скуп n-грама који се пресликавају у $b_j(d_{jkn})$.*

Алгоритам 1 - издвајање скупа секвенцијалних образаца придружених одговарајућој ЦОГ категорије протеина

Користећи дескрипторе као карактеристичне обрасце за појединачну ЦОГ категорију може се дефинисати алгоритам за издвајање скупа секвенцијалних образаца карактеристичних за одговарајућу ЦОГ категорију. Алгоритам се састоји од 7 корака:

1. Постављање почетних вредности:
   (a) почетна вредност дужине n-грама се поставља на 0 ($n = 0$),
   (b) задаје се максимална вредност дужине n-грама max_n,
   (c) задају се вредности прага за подршку s (min_s) и поверења c (min_c) за ОАК ниске.
2. Увећати n за 1 и за сваки протеин P издвојити скуп преклапајућих n-грама (a_i) одређене дужине n, $O_n = \{a_i | a_i \in P, |a_i| = n\}$.
3. За сваки протеин P одредити скуп ОАК ниски $B_n = \{b_j | b_j = OAK(a_i), a_i \in O_n\}$.
4. На основу добијених скупова формирају се четворке $T_{jkn} = (n, COG_k, b_j, \#a_i)$, где је n дужина обрађених преклапајућих n-грама, COG_k је ЦОГ категорија, $b_j \in B_j$ је добијена ОАК ниска из преклапајућих

n-грам протеина категорије COG_k, и $\#a_i$ је број секвенци a_i за које важи $(b_j = OAK(a_i) \wedge a_i \in P \wedge P \in COG_k)$.

5. Из скупа конструисаних четворки $T_{jkn} = (n, COG_k, b_j, \#a_i)$ издвајају се оне код којих су подршка и поузданост за ОАК ниску b_j веће од предефинисаних вредности min_s и min_c. На основу издвојеног скупа четворки се за ЦОГ категорију COG_k одређује вектор од l карактеристичних секвенцијалних образаца издвајањем првих l ОАК ниски b_j из скупа четворки T_{jkn} за одговарајућу COG_k, при чему је скуп четворки T_{jkn} уређен у опадајућем поретку у односу на пар (поверење, подршка). Из скупа пронађених секвенцијалних образаца одабрати оне који су дескриптори.

6. Ако је $n < max_n$ вратити се на корак 2.

7. Објединити скуп свих дескриптора добијених у кораку 5 за различите вредности n.

Издвојени скуп дескриптора се користи за даљу изградњу модела за класификацију протеина по ЦОГ категоријама. Класификациони модел се конструише на основу чињенице да је сваки дескриптор придружен само једној ЦОГ категорији. Нов протеин који још увек није класификован по ЦОГ категоријама би се класификовао у ЦОГ категорију COG_k ако садржи n-грам из скупа D_{jkn} за неко j и n.
Секвенцијални обрасци не морају да буду потпуно одређени. Тачније, могу да буду још општији са приближним обрасцима. Генерализација секвенцијалних образаца у овом случају представља замену једне конкретне амино киселине из ОАК ниске са џокером (променљивом која може представљати било коју амино киселину).

Нека је so^* приближни секвенцијални образац дефинисан паром (b_j^*, n) при чему је b_j^* унија ОАК ниске b_j и џокера у ознаци $*$, и n природан број. Ниска x задовољава секвенцијални образац so^* акко је $|x| = n$ таква да се њених $n-1$ карактера пресликава у ОАК ниску b_j, док се један карактер пресликава у $*$ и може бити било која амино киселина. Нека је са OAK^* означено пресликавање које пресликава ниску x у одговарајућу ниску b_j^*.

Дефиниција 19. *Приближни дескриптор d^*_{jkn} n-грама дужине n је комбиновани секвенцијални образац који се састоји од ОАК ниске b_j и једног џокера у ознаци $*$ $(b_j, *)$, такав да постоји четворка $T^*_{jkn} = (n, COG_k, b^*_j, \#a_i)$ где је n дужина обрађених преклапајућих n-грама, COG_k је ЦОГ категорија, $b_j \in B_j$ је добијена ОАК ниска из преклапајућих n-грама протеина категорије COG_k, $\#a_i$ број секвенци a_i за које важи $(b^*_j = OAK^*(a_i) \wedge a_i \in P \wedge P \in COG_k)$, поверење од b^*_j је једнако 100% а подршка већа од унапред дефинисане вредности min_s.*

Дефиниција 20. *n-Грами описани дескрипторима (приближним дескрипторима) називају се карактеристични n-грами.*

Алгоритам 1 се може искористити и за издвајање скупа приближних дескриптора. Дескриптори и приближни дескриптори заједно чине скуп образаца. Квалитет издвојених правила (образац, ЦОГ категорија), односно специјалније (карактеристични n-грам, ЦОГ категорија) проверава се на скупу података за тестирање. Прецизност издвојених образаца (po) се одређује као однос броја карактеристичних n-грама који су пронађени у протеинима очекиване ЦОГ категорије (придружене обрасцем) и укупног броја пронађених карактеристичних n-грама ($TP/(TP + NP)$).

3.2 Конструкција модела за предвиђање ЦОГ категорија протеина

Предвиђање структуре и функција протеина помоћу биоинформатичких алата укључује претрагу: сличних секвенци у протеинима, вишеструког појављивања секвенци у протеину, идентификацију и карактеризацију домена, прогнозу секундарне структуре протеина, изградњу тродимензионалних модела, итд. Употребом ових метода на одговарајући начин, могу се обезбедити вредни показатељи структуре протеина и њихове функције. Први корак једне методе за предвиђање структуре или функције протеина је да се утврди да ли секвенца протеина или њен део има било какве сличности са секвенцама протеина познатих структура и фукција које се могу наћи у доступним базама података (на пример у ГенБанк). Овакве претраге су рачунарски јако захтевне и треба их оптимизовати и наћи алтернативу поређења сваке нове протеинске

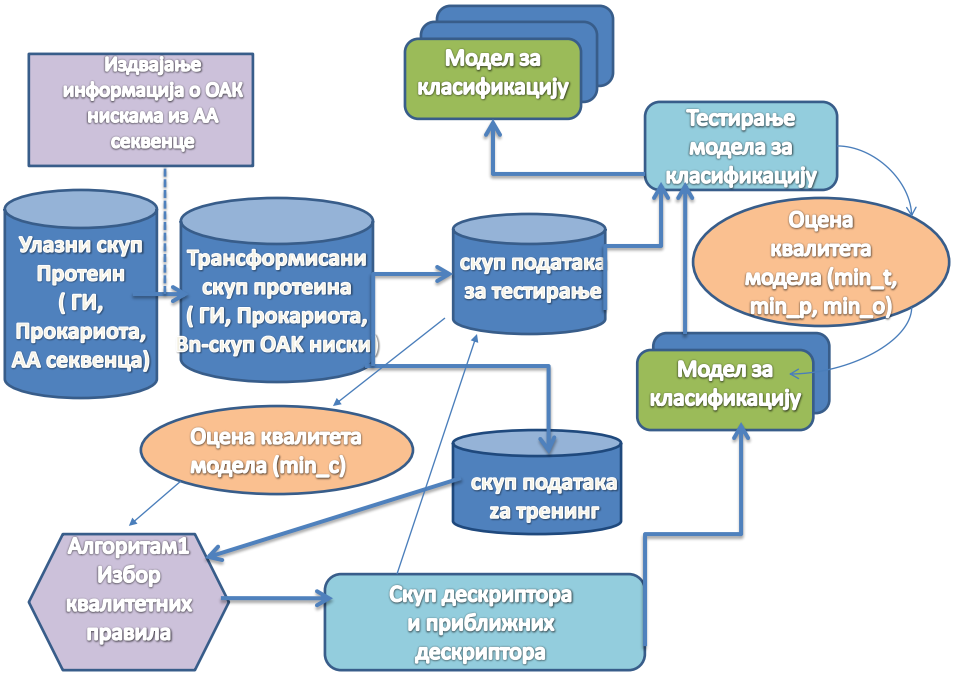
секвенце са скупом свих познатих секвенци. Описане методе у претходним поглављима омогућавају прављење модела којим се поређење секвенце са секвенцом на коме се заснивају претходни методи замењује поређењем секвенце са скупом издвојених ниски које су карактеристичне за одређену ЦОГ категорију протеина. У даљем тексту је приказана структура предиктора за класификацију која користи модел учења претраживањем секвенцијалних образаца. Илустрација схеме изградње модела за класификацију протеина по ЦОГ категоријама је приказана на Слици 3.1. Модел је иницијално тестиран за класу *Chlamydiales*. Опис модела и резултати тестирања су приказани у раду [ММ]. Квалитет конструисаног класификационог модела зависи од

- квалитета издвојених образаца представљеног са прецизношћу образаца po,
- резултата поређења класификације са резултатима познатих метода на тест подацима представљених бројем протеина из података за тестирање који су тачно односно нетачно придружени одговарајућој ЦОГ категорији користећи матрицу конфузије и мере тачност (t), прецизност (p) и одзив (o).

Скуп образаца је класификован на основу мера поверења c, подршке s и бројача подршке σ. Корисник поставља вредност прагова минималне подршке min_s док је поверење по дефиницији образаца 100%. Праг бројача подршке min_σ се одређује према захтеваном квалитету издвојених образаца које корисник уноси као праг прецизности издвојених правила min_po. Протеину се придружује одређена ЦОГ категорија COG_k ако је број карактеристичних n-грама које протеин садржи и који су придружени овој категорији већи од унапред одређеног прага h. Вредност прага h се добија на основу захтеване тачности. Опис поступка изградње модела за класификацију приказан је у Алгоритму 2.

Алгоритам 2 - изградња модела за класификацију протеина по функционалним категоријама ЦОГ-ова

Поступак изградње модела за класификацију протеина по функционалним категоријама ЦОГ-ова се састоји из следећих корака:



Слика 3.1: Илустрација изградње модела за класификацију

1. *Избор скупа података над којима ће бити направљен модел и постављање почетних вредности*

   (a) избор скупа прокариотских организама (бактерија и археа) из НЦБИ базе протеина над којима ће бити направљен модел, као и информација о класификацији њихових протеина у ЦОГ категорије;

   (b) задају се вредности за прагове подршке min_s и прецизности min_po обрасца;

   (c) задају се вредности за очекивани квалитет класификације (min_t, min_p, min_o);

   (d) постављају се почетне вредност за min_σ и h на 1.

2. *Трансформација улазних података у податке са траженим скупом атрибута*

   За сваки протеин улазни податак се претвара у скуп атрибута са информацијама о врсти и броју ОАК секвенци које садрже. Поступак за добијање ових информација је описан у претходном поглављу и добија се

пресликавањем комплетне аминокиселинске ниске протеина у скуп ОАК ниски Bn.

3. *Подела скупа података*

   Скуп улазних података се дели на два дела у одговарајућем односу (скуп за тренинг и скуп за тестирање).

4. *Издвајање правила придруживања над скупом података за тренирање*

   Правила придруживања укључују секвенцијалне обрасце (дескриптори и приближни дескриптори) који су добијени помоћу алгоритма 1, примењеног на скупу података за тренинг.

5. *Оптимизација скупа издвојених правила*

   Скуп издвојених података може бити превелик и такав модел носи са собом последице везане за време обраде као и квалитет добијених информација. Зато је уведен услов који треба да задовоље издвојени обрасци а то је минимална подршка min_c и минимална апсолутна подршка min_σ.

6. *Оцена квалитета издвојених правила*

   Квалитет издвојених правила се проверава на тест подацима израчунавањем мере прецизности po, за коју је очекивано да је већа од прага min_po.

7. *Условни скок на корак 4*

   Ако правила не задовољавају минималну прецизност треба поновити поступак од корака 4 и при томе повећати праг за бројач подршке min_σ за 1.

8. *Дефинисање правила класификације*

   Свакој ЦОГ категорији претходним корацима додељен је скуп образаца који су за њу карактеристични $SD(COG_k)$ где је $SD(COG_k) = \{so|(so = d_{jkt} \wedge (\exists T_{jkn})(T_{jkn} = (n, COG_k, b_j, \#a_i) \wedge d_{jkt} = b_j \wedge \#a_i \geq min\_\sigma)) \vee (so = d^*_{jkt} \wedge (\exists T^*_{jkn})(T^*_{jkn} = (n, COG_k, b^*_j, \#a_i) \wedge d^*_{jkt} = (b_j, *) \wedge \#a_i \geq min\_\sigma))\}$. Да би новом протеину била додељена ЦОГ категорија COG_k он мора да садржи више од унапред одређеног прага h карактеристичних n-грама описаних скупом $SD(COG_k)$.

9. *Тестирање класификационог модела*
   Квалитет модела за класификацију се проверава на скупу података за тестирање. Израчунавају се мере квалитета тачност (t), прецизност (p) и одзив (o).

10. *Условни скок на корак 8*
    Захтеве за квалитет модела поставља корисник и они се изражавају кроз минималне прагове за мере: тачност min_t, прецизност min_p и одзив min_o. Ако модел не задовољава тражени квалитет поновити поступак од корака 8 са повећаном вредности прага h за 1.

11. *Модел за класификацију*
    Добијени модел се користи за предвиђање функционалних карактеристика до сада некласификованих (нових) протеина.

На основу дефинисаног модела имплементиран је предиктор за класификацију некласификованих протеина у одговарајуће ЦОГ категорије. Алгоритам који је коришћен у имплементцији је приказан у даљем тексту.

Алгоритам 3 - предиктор

1. Прикупљају се информације о новом протеину.

2. Врши се трансформација улазних података (практично аминокиселинске секвенце протеина) у скуп ОАК ниски.

3. Уноси се жељени праг квалитета $min\_po, min\_t, min\_p, min\_o$ на основу кога се издваја скуп правила придруживања као и класификациони модел који задовољава ове мере квалитета.

4. Помоћу направљеног модела врши се поређење ОАК ниски протеина који се класификује.

5. Као резултат се бира ЦОГ категорија која са највећом поузданошћу може да се придружи протеину.

Поглавље 4

Тестирање и примена модела

У овом поглављу биће приказани резултати тестирања претходно описаног модела, као и примене предиктора заснованог на моделу на протеине који нису били претходно класификовани.

4.1 Материјал

Разноврсност живог света на нашој планети је изузетно велика, па се потреба за груписањем организама јавила веома давно. Класификација бактерија је разврставање бактерија у таксономске групе на основу сличних особина и њихове повезаности (сродности). Бактерије као ћелијски организми су класификовани у разделе а затим у класе. Предложени модел је тестиран на 8 различитих класа бактерија. Класе бактерија над којима су изведени експерименти су приказане у Табели 4.1.

Представљени модел за одређивање функционалне категорије протеина (ЦОГ категорије) је класификациони модел. У процесу конструкције и провере класификационог модела скуп улазних података се дели на скуп за тренинг и скуп за тестирање. Скуп података за тренинг се користи у процесу формирања модела, док се скуп податка за тестирање користи у процесу тестирања конструисаног модела. При формирању предложеног модела за класификацију протеина, на основу података за тренинг се издвајају карактеристике протеина који припадају истој ЦОГ категорији (секвенцијални обрасци).

Табела 4.1: Класе бактерија коришћене за израду модела

Надраздео (*superphylum*)	Раздео (*phylum*)	Класа(*class*)
Aquificae	Aquificae	Aquificales
Bacteroidetes/Chlorobi group	Bacteroidetes	Bacteroidia
Chlamydiae/Verrucomicrobia group	Chlamydiae	Chlamydiales
Bacteroidetes/Chlorobi group	Chlorobi	Chlorobia
Chloroflexi	Chloroflexi	Chloroflexia
Bacteroidetes/Chlorobi group	Bacteroidetes	Cytophagia
Deinococcus-Thermus	Deinococcus-Thermus	Deinococci
Cyanobacteria	Cyanobacteria	Prochlorales

На основу добијених правила придруживања *секвенцијални образац - ЦОГ категорија* формира се класификациони модел. Исправност модела проверава се над подацима за тестирање, поређењем добијених резултата класификације и вредности до којих су биолози дошли методама поравнања. Овом провером утврђује се прецизност класификације протеина представљеним моделом. Утврђивање ваљаности модела засновано је на мерама дефинисаним за квалитет правила придруживања и класификације приказаним у претходним поглављима.

Први важан корак у формирању класификационог модела је прављење одговарајућег скупа података за тренинг и тестирање. Мањи скуп података за тренинг даје већа одступања у моделу, док са друге стране превелики скуп података за тренинг доводи у питање поузданост информације о ваљаности модела добијене над малим скупом података за тестирање. Најчешће су поделе података на тренинг скуп и скуп за тестирање у односу $2:1$ или $70:30$. Постоје класификациони модели код којих је довољан и мањи број података у скупу за тренинг (нпр. код методе подржавајућих вектора подела је у односу $1:1$). Како је приказани модел за класификацију протеина по ЦОГ категоријама заснован на дрвету одлучивања подаци су подељени у односу оријентационо $70:30$ према броју генома у фамилијама (уз заокруживање материјала у групе на цео број генома). За скуп података узети су само они организми који имају протеине класификоване по различитим ЦОГ категоријама (које нису у N.C.). При подели података на тренинг и тест скуп вођено је рачуна и да подела протеина класификованих по ЦОГ категоријама буде приближно у истом односу. Тежило

се да у истом односу буду и укупне дужине протеина који припадају скуповима за тренинг и тестирање. Приказ бројчаног стања генома, протеина као и укупне дужине одговарајућих протеина у скуповима за тренинг и тестирање налази се у Табели 4.2.

Табела 4.2: Подела генома у скупове података за тренинг и тестирање по класама бактерија

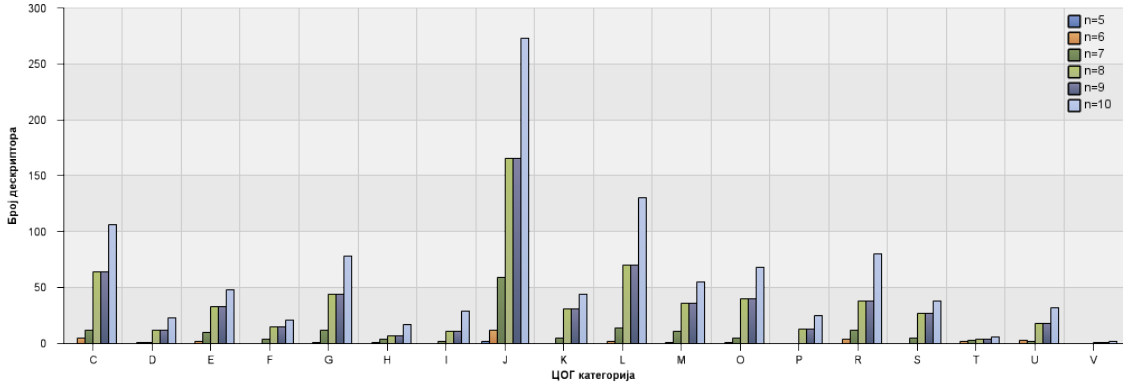
класа	тренинг /тест	Број генома	Број протеина	Укупна дужина протеина
Aquificales	тренинг	7	7299	2432659
Aquificales	тест	2	3254	1091142
Bacteroidia	тренинг	19	24855	10131912
Bacteroidia	тест	9	11642	4686917
Chlamydiales	тренинг	20	12943	4909775
Chlamydiales	тест	8	5348	2019885
Chlorobia	тренинг	8	14493	5195710
Chlorobia	тест	4	6200	2276214
Chloroflexia	тренинг	6	14568	5749788
Chloroflexia	тест	2	6427	2490958
Cytophagia	тренинг	14	18130	7086127
Cytophagia	тест	8	7938	3106526
Deinococci	тренинг	23	21629	7415660
Deinococci	тест	9	9298	318509
Prochlorales	тренинг	8	11387	3936637
Prochlorales	тест	4	5230	1793545

Број протеина по категоријама за сваку класу као и списак организама који припадају скупу тренинг/тест података приказан је у додатку A.

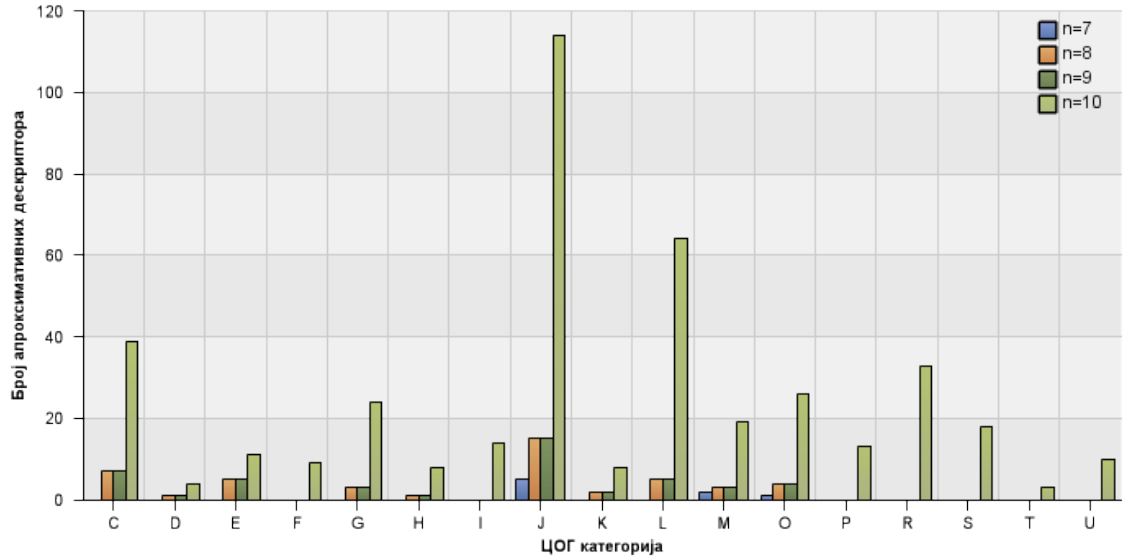
4.2 Квалитет издвојених образаца

Применом приказаног алгоритма за издвајање образаца (Алгоритам 1) на тренинг скуповима за сваку од одабраних класа бактерија посебно, добијен је скуп образаца (дескриптора и приближних дескриптора) за сваку ЦОГ категорију и одређену класу бактерија. Издвојени су дескриптори за $n \in \{5, \ldots, 10\}$ и приближни дескриптори за $n \in \{7, \ldots, 10\}$ за сваку од ЦОГ категорија. За краће n-граме није било могуће издвојити ОАК ниске које

карактеришу искључиво једну ЦОГ категорију. Број издвојених образаца, као што је очекивано, расте са бројем n, дужином n-грама који се описују обрасцима. На Сликама 4.1 и 4.2 приказан је број издвојених образаца за класу *Chlamydiales*. У осталим класама бактерија је слична ситуација.



Слика 4.1: Број издвојених дескриптора по ЦОГ категоријама за различите дужине n-грама



Слика 4.2: Број издвојених приближних дескриптора по ЦОГ категоријама за различите дужине n-грама

Број образаца по класама бактерија и ЦОГ категоријама приказан је у Табелама 4.3 и 4.4. Обрасцима је дефинисан скуп карактеристичних n-грама за сваку класу бактерија и ЦОГ категорију. Проценат издвојених

Табела 4.3: Број издвојених дескриптора по класама бактерија

ЦОГ	Aquificales	Bacteroidia	Chlamydiales	Chlorobia	Chloroflexia	Cytophagia	Deinococci	Prochlorales
A	50	0	220	0	2	0	46	0
B	0	0	97	0	0	0	0	0
C	35180	7947	19724	26050	16735	11100	20553	15722
D	2551	1890	5764	2040	1787	1221	1446	3022
E	16546	11801	19499	13749	16250	15857	23826	26896
F	8822	3874	5882	5177	3840	3728	6796	7306
G	10477	34852	17302	12011	21383	21409	23934	16312
H	14718	8071	11983	14935	9023	6728	6028	20946
I	5969	3981	8522	6209	5863	5953	8403	6055
J	23349	6665	28909	11842	8198	7430	13704	20539
K	4842	5563	3730	4642	7022	7292	6279	4339
L	16067	12748	24962	20462	14570	11979	14180	18370
M	20909	21038	26930	23562	16592	18470	9976	23347
N	1933	52	443	88	53	65	202	0
O	11445	5344	11958	7553	7934	6932	9205	11823
P	12405	15004	9108	11335	6265	11105	7958	8871
Q	1672	898	756	1939	6807	3078	3593	4196
R	22441	23448	26685	32110	36343	31225	27506	30672
S	14023	11801	19687	21163	19662	18693	15852	19177
T	7139	8587	2813	7092	14239	10851	6784	2464
U	1990	2449	6330	2337	2239	1548	2083	2274
V	2453	5722	2163	6085	5638	5770	3676	4465
Z	0	0	0	0	10	0	0	12

образаца по различитим ЦОГ категоријама у односу на укупан број образаца у једној класи бактерија не одступа значајно по различитим класама бактерија. Број издвојених образаца за ЦОГ категорију COG_k је директно зависан од броја протеина у ЦОГ категорији COG_k и укупног броја протеина у класи. Значајност издвојених образаца зависи од броја различитих протеина у којима су пронађени карактеристични n-грами који задовољавају образац као и укупног броја тих карактеристичних n-грама.

Мере значајности образаца придружених ЦОГ категоријама су поверење и подршка. Дефиницијом дескриптора и приближних дескриптора одређено је да је њихово поверење увек 100%. Број издвојених правила (образац, ЦОГ категорија) мора се редуковати из два разлога:

1) са сложеношћу модела повећава се и време обраде и обрнуто;

2) бољи резултати класификације се добијају када се избаце слабија правила.

Табела 4.4: Број издвојених приближних дескриптора по класама бактерија

ЦОГ	Aquificales	Bacteroidia	Chlamydiales	Chlorobia	Chloroflexia	Cytophagia	Deinococci	Prochlorales
A	34	0	42	0	0	0	4	0
B	0	0	10	0	0	0	0	0
C	12972	118	4158	4558	3320	1523	4992	3662
D	697	35	1202	270	331	93	220	618
E	4663	123	4403	2216	3155	2345	5413	6572
F	2824	68	1321	1225	817	472	1728	1617
G	3270	814	3718	2044	4226	3328	5763	3994
H	4404	125	2875	2751	1778	772	1219	5310
I	2054	64	1856	1064	1227	722	2116	1487
J	7485	109	6039	1750	1620	1038	3361	4849
K	1501	81	726	854	1395	1049	1539	993
L	4201	220	5074	3949	3303	1738	3482	4387
M	5573	434	6106	3929	2983	2542	2005	5193
N	268	0	76	6	17	1	64	0
O	3561	120	2760	1319	1374	921	2312	2527
P	3148	164	1782	1648	1047	833	1595	1842
Q	663	13	220	295	1354	406	758	924
R	6020	419	5546	6217	7193	4340	6443	6975
S	4072	257	4184	3586	3857	2601	3703	4383
T	1897	131	498	1186	2468	1175	1507	480
U	357	29	1270	323	302	190	465	488
V	598	89	449	752	962	586	740	783

У Табелама 4.5 и 4.6 приказан је број образаца са подршком већом од 0.5% по ЦОГ категоријама збирно за све класе бактерија. Мала подршка је оправдана бројем n-грама који се могу појавити, на пример број могућих n-грама дужине 5 је $20^5 = 3200000$. За ЦОГ категорије које имају мали број протеина сви обрасци који су се појавили бар у једном протеину имају подршку већу од 0.5%. Зато се при избору образаца уводи услов да је апсолутна подршка обрасца већа од 1 (бројач подршке $\sigma > 1$). Апсолутна подршка обрасца која представља

број пронађених карактеристичних n-грама који задовољавају одговарајући образац, битна је јер правило које се није ни једном поновило може бити случајно.

Табела 4.5: Број дескриптора $\#d$ за које су у моделу пронађени карактеристични n-грами у више од 0.5% протеина одговарајуће ЦОГ категорије

ЦОГ	A	B	C	D	E	F	G	H	I	J	K	L
$\#d$	34	78	8952	5539	4651	6686	8195	5148	7839	4799	4209	8403

ЦОГ	M	N	O	P	Q	R	S	T	U	V	Z
$\#d$	3620	259	6774	4920	3097	1810	4418	2192	3863	4373	2

Табела 4.6: Број приближних дескриптора ($\#d*$) за које су у моделу пронађени карактеристични n-грами у више од 0.5% протеина одговарајуће ЦОГ категорије

ЦОГ	A	B	C	D	E	F	G	H	I	J	K
$\#d*$	20	10	2379	1125	1178	1686	2152	1406	1995	1288	954

ЦОГ	L	M	N	O	P	Q	R	S	T	U	V
$\#d*$	2057	865	33	1736	1128	739	426	1105	476	616	853

Због различитог броја протеина по ЦОГ категоријама мора се водити рачуна о апсолутној подршци па се уводи праг за бројач подршке обрасца min_σ. Овај број се одређује експериментално за сваку класу на основу позузданости модела добијеног овом рестрикцијом. Утврђено је да овај број није мањи од 5 за све обрађене класе.

У Табелама 4.7 и 4.8 приказани су дескриптори и приближни дескриптори са највећом апсолутном подршком у тренинг подацима. Тако је на пример пронађено 119 n-грама дужине $n = 9$ састављених искључиво од амино киселина из скупа {G, K, Y} у протеинима ЦОГ категорије D бактерије класе *Bacteroidia* и то у 13 различитих протеина што чини 4 % од укупног броја протеина. Приближни дескриптор одређен паром (OAK, n)=("NGVRYWH", 10) потврђен је са 73 карактеристична n-грама у 23 протеина класе *Chlamydiales* који припадају само ЦОГ категорији J, што чини 1.2% укупног броја протеина

ове категорије у класи *Chlamydiales*.

Табела 4.7: Дескриптори са највећим бројем појава у моделу

Класа	ЦОГ	ОАК	n	$\#n$ -грама	#гена	#гена у ЦОГ-у	подршка
Bacteroidia	D	GKY	9	119	13	316	4.11%
Deinococci	L	AQKRTYWH	9	111	56	1233	4.54%
Deinococci	L	AQKRTYWH	8	111	56	1233	4.54%
Deinococci	L	AQLKRTYWH	10	110	55	1233	4.46%
Deinococci	L	EQKRTMF	10	108	54	1233	4.38%
Bacteroidia	D	GKY	10	108	12	316	3.8%
Prochlorales	O	GM	8	85	8	656	1.22%
Chlamydiales	K	ENKRTPIFY	10	80	16	425	3.76%
Prochlorales	O	GM	9	77	8	656	1.22%
Cytophagia	L	QTFYCH	8	72	18	992	1.81%

Табела 4.8: Приближни дескриптори са највећим бројем појава у моделу

Класа	ЦОГ	ОАК	n	$\#n$ -грама	#гена	#гена у ЦОГ-у	подршка
Chlamydiales	J	NGVRYWH	10	73	23	1877	1.23%
Cytophagia	L	QLGKRTMCW	10	68	21	992	2.12%
Chlorobia	L	AQKVRIFYW	10	63	21	995	2.11%
Chlamydiales	C	ANLGPMFYC	10	62	16	751	2.13%
Chlamydiales	J	DLGSPMFWH	10	59	16	1877	0.85%
Deinococci	L	EQNLGKRMH	10	56	56	1233	4.54%
Chlamydiales	L	ASRIMCWH	10	56	13	1125	1.16%
Deinococci	L	DNRPMYCH	10	54	54	1233	4.38%
Cytophagia	L	QSRTMCW	9	53	48	992	4.84%
Chlamydiales	G	ENGRPIYCW	10	52	13	617	02.1%

Издвајањем образаца дужине 10 за класу *Chlamydiales* може се уочити да су неки дескриптори подскупови приближних дескриптора. На пример,

дескриптор d_1 одређен паром (OAK, n)=("AENLGYCWH", 10) који је придружен категорији F са поверењем 100% и подршком 6.4% је садржан у приближном дескриптору d_2 дефинисаним са (OAK, n) = ("AENLGY-CWH", 10) који такође има поверење 100% и подршку 6.4%. Дескриптор d_1 описује све n-граме дужине 10 у којима се појављују све амино киселине из скупа $A_1 = \{'A','E','N','L','G','Y','C','W','H'\}$ и ниједна друга амино киселина. Приближним дескриптором d_2 описују се сви n-грами дужине 10 у којима се на девет позиција појављују све амино киселине из скупа $A2 = \{'A','E','N','L','G','Y','C','W','H'\}$ и ни једна друга, а на једној произвољној позицији може се наћи било која амино киселина. Провером на тест подацима утврђена је 100% прецизност за оба обрасца. Стога, дескриптор d_1 може бити искључен из скупа образаца придружених ЦОГ категорији F јер га садржи приближни дескриптор d_2 са истом прецизношћу, поверењем и подршком. У циљу смањења броја образаца придружених једној категорији могу се користи и ове методе при чему треба увек водити рачуна о прецизности модела који се овом рестрикцијом добија.

4.3 Утицај различитих параметара на квалитет издвојених секвенцијалних образаца

Када се дефинише скуп правила може се проверити исправност модела поређењем резултата класификације и већ унапред познатих информација о ЦОГ категоријама. Квалитет правила одређује квалитет модела.

Претходно је указано да квалитет правила зависи од мера подршке (s), поверења (c), и од апсолутне подршке (σ), тј. броја понављања тог правила. Поверење свих издвојених правила је 100%. Због разноликог броја протеина по категоријама подршка која представља процентуалан број протеина у ЦОГ категорији у којима се правило јавило постављена је на 0.5% да би биле обухваћене и категорије са великим бројем протеина.

Квалитет издвојених правила (образаца, ЦОГ категорија), односно специјалније (карактеристични n-грам, ЦОГ категорија) проверава се на скупу података за тестирање рачунањем прецизности издвојених правила *ро*. У Табели 4.9 приказан је утицај вредности параметра min_σ на квалитет

класификационог модела.

Табела 4.9: Проценат карактеристичних n-грама пронађених у одговарајућој ЦОГ категорији у подацима за тестирање за различиту вредност прага min_σ

min_σ	Aquificales	Bacteroidia	Chlamydiales	Chlorobia	Chloroflexia	Cytophagia	Deinococci	Prochlorales
1	53%	74%	58%	67%	65%	55%	74%	89%
2	62%	77%	59%	71%	75%	61%	77%	91%
3	68%	78%	59%	73%	75%	63%	78%	92%
4	74%	81%	64%	75%	76%	63%	81%	93%
5	76%	82%	66%	75%	76%	64%	82%	94%
6	78%	84%	68%	77%	76%	60%	83%	94%
7	79%	84%	71%	79%	85%	66%	83%	94%
8	79%	85%	74%	77%	82%	64%	84%	95%
9	79%	86%	75%	78%	78%	65%	85%	95%
10	77%	88%	75%	79%	77%	62%	84%	95%
11	77%	90%	76%	79%	77%	62%	83%	96%
12	79%	91%	78%	81%	86%	58%	82%	96%
13	89%	92%	79%	81%	100%	58%	82%	96%
14	91%	93%	80%	82%	100%	57%	79%	97%
15	91%	93%	81%	82%	100%	57%	78%	96%

Табела 4.10: Проценат карактеристичних n-грама пронађених у очекиваној ЦОГ категорији у подацима за различите ЦОГ категорије и дужине n-грама

n	Aquificales	Bacteroidia	Chlamydiales	Chlorobia	Chloroflexia	Cytophagia	Deinococci	Prochlorales
5	73.68%	-	77.77%	62.16%	85.71%	80%	72.72%	91.66%
6	70.00%	71.26%	74.77%	71.39%	-	87.50%	83.63%	89.25%
7	78.65%	72.50%	75.77%	74%	95.00%	72.22%	83.93%	92.08%
8	81.60%	75.64%	80.81%	75.69%	87.24%	80.25%	84.85%	92.68%
9	79.01%	78.50%	80.72%	77.79%	87.26%	78.68%	83.29%	93.68%
10	80.85%	79.38%	81.36%	78.60%	90.21%	79.14%	83.13%	93.71%

Већина класа бактерија задовољава прецизност већу од 64% за $min\_\sigma = 5$. Стога ће се у даљем раду ова вредност користити као праг апсолутне

подршке обрасца. У Табели 4.10 приказана је прецизност издвојених правила за различите вредности дужина карактеристичних n-грама. Поузданост очекивања да ће се карактеристични n-грами описани обрасцима наћи у ЦОГ категорији којој је образац придружен је већа од 71% за скоро све дужине карактеристичних n-грама и класа бактерија. Посматрано по различитим дужинама n-грама може се уочити да квалитет образаца варира и са дужином n-грама који се описују.

Број протеина одговарајуће ЦОГ категорије који припадају једној класи такође директно утиче на квалитет образаца који се добијају. Обрасци који су придружени ЦОГ категоријама са мањим бројем протеина имају мању прецизност. Тако, у Табели 4.11 је приказана прецизност издвојених правила посебно за сваку ЦОГ категорију и класу бактерија. Из табеле се може уочити да за ЦОГ категорије A, B, N, Q, T, U, V, Z у више од пола класа број пронађених карактеристичних n -грама у очекиваној ЦОГ категорији (ЦОГ категорији која му је придружена одговарајућим обрасцем) је мањи од две трећине укупног броја пронађених карактеристичних n-грама. Ако упоредимо број протеина по ЦОГ категоријама (додатак A) може се видети да су ово уједно и ЦОГ категорије са најмањим бројем протеина. Међутим, ни ово није нерешив проблем и може се исправити приликом одређивања правила класификовања елемената.

Прецизност модела зависи и од минималног броја карактеристичних n -грама придружених ЦОГ категорији COG_k који је потребан да буде пронађен у протеину да би протеин био класификован у ЦОГ категорију COG_k. На овај начин случајан проналазак карактеристичног n-грама у неодговарајућој ЦОГ категорији биће игнорисан. Такође проблем ЦОГ категорија са малим бројем протеина може се на овај начин неутралисати.

4.4 Квалитет модела за класификацију протеина по ЦОГ категоријама

У претходном поглављу је описан квалитет правила придружених одговарајућој ЦОГ категорији за одређену класу бактерија. Утврђени су параметри од којих

Табела 4.11: Проценат карактеристичних n-грама пронађених у очекиваној ЦОГ категорији посматрано по различитим ЦОГ категоријама и класама бактерија

ЦОГ	Aquificales	Bacteroidia	Chlamydiales	Chlorobia	Chloroflexia	Cytophagia	Deinococci	Prochlorales
A	-	-	-	-	-	-	0%	-
B	0%	-	-	-	-	0%	0%	-
C	86%	75%	85%	89%	94%	83%	85%	98%
D	59%	77%	91%	73%	29%	93%	82%	94%
E	72%	73%	63%	73%	84%	68%	76%	93%
F	80%	80%	70%	82%	72%	44%	81%	96%
G	45%	83%	78%	71%	64%	57%	71%	96%
H	75%	59%	48%	76%	84%	57%	63%	94%
I	83%	53%	65%	66%	90%	51%	87%	95%
J	94%	90%	94%	89%	94%	95%	92%	98%
K	75%	33%	81%	60%	46%	64%	66%	96%
L	65%	75%	80%	66%	64%	73%	74%	94%
M	44%	60%	62%	52%	54%	26%	40%	82%
N	0%	3%	3%	0%	0%	0%	0%	0%
O	74%	57%	86%	65%	85%	74%	82%	96%
P	44%	47%	51%	62%	56%	50%	49%	96%
Q	0%	7%	0%	13%	8%	29%	34%	50%
R	44%	59%	62%	59%	67%	55%	66%	95%
S	32%	53%	62%	51%	79%	41%	63%	91%
T	13%	8%	41%	21%	48%	34%	36%	90%
U	18%	66%	72%	27%	69%	26%	38%	88%
V	23%	45%	22%	36%	61%	20%	42%	98%
Z	-	-	-	-	0%	-	-	-

зависи квалитет издвојених правила. За сваку ЦОГ категорију COG_k издвојен је скуп карактеристичних секвенцијалних образаца ($S_k = SD(COG_k)$) који задовољавају унапред одређене минималне вредности за поверење, подршку и бројач подршке.

Квалитет модела за класификацију зависи од квалитета издвојених правила али и од начина на који ће бити дефинисана класификација. Једно пронађено правило у протеину из скупа (S_k) није довољно да би протеину била придружена одређена ЦОГ категорија COG_k. Да би избегли случајно појављивање карактеристичног n-грама, услов за придруживање протеина ЦОГ категорији је да се у њему појављује више n -грама карактеристичних за ту ЦОГ категорију.

Табела 4.12: Квалитет модела у зависности од минималног броја различитих образаца пронађених у протеину да би протеин био класификован у одговарајућу ЦОГ категорију (h)

h	TP	NP	p	o
1	974	302	76%	18%
2	696	59	92%	13%
3	523	23	96%	10%
4	433	1	100%	8%
5	351	1	100%	7%
6	291	0	100%	5%
7	241	0	100%	5%
8	205	0	100%	4%
9	171	0	100%	3%
10	137	0	100%	3%

Табела 4.13: Квалитет модела у зависности од минималног броја карактеристичних n-грама пронађених у протеину да би протеин био класификован у одговарајућу ЦОГ категорију (h)

h	TP	NP	p	o
1	1215	494	71%	23%
2	956	160	86%	18%
3	784	73	91%	15%
4	688	29	96%	13%
5	596	15	98%	11%
6	507	8	98%	9%
7	446	8	98%	8%
8	390	4	99%	7%
9	346	3	99%	6%
10	296	0	100%	6%

У Табелама 4.12 и 4.13 су приказане карактеристике класификационог модела за класу *Chlamydiales* (p—прецизност, o—одзив) који протеину додељује ЦОГ категорију COG_k ако је у њему нађено најмање h образаца (карактеристичних n-грама) из скупа (S_k). За остале класе резултати се налазе у додатку *Б*.

Повећањем прага h добија се све већа прецизност и брзо се достиже вредност 100%, уз опадање броја протеина који се овим поступком класификују, и самим тим добија се мањи одзив. Зависно од захтева корисника, прецизност модела се може регулисати променом параметра h.

Табела 4.14: Квалитет модела посебно за сваку ЦОГ категорију

ЦОГ	NP	TP	NN	TN	t	p	o
A	0	0	0	5301	100%	-	-
B	0	0	6	5295	100%	-	0%
C	4	75	236	4986	95%	95%	24%
D	0	12	59	5230	99%	100%	17%
E	2	31	359	4909	93%	94%	8%
F	0	17	126	5158	98%	100%	12%
G	0	44	250	5007	95%	100%	15%
H	0	9	242	5050	95%	100%	4%
I	0	14	207	5080	96%	100%	6%
J	6	190	441	4664	92%	97%	30%
K	1	24	153	5123	97%	96%	14%
L	4	77	328	4892	94%	95%	19%
M	2	32	366	4901	93%	94%	8%
N	0	0	84	5217	98%	-	0%
O	5	50	191	5055	96%	91%	21%
P	0	9	189	5103	96%	100%	5%
Q	0	47	37	5217	99%	100%	56%
R	1	26	597	4677	89%	96%	4%
S	0	10	337	4954	94%	100%	3%
T	1	20	132	5148	97%	95%	13%
U	0	1	208	5092	96%	100%	0%
V	0	0	65	5236	99%	-	0%

Анализом резултата утврђено је да протеини одређене категорије врло често садрже и више карактеристичних n-грама који припадају истом обрасцу, док се то у протеинима осталих категорија ређе дешава, и углавном су то усамљене појаве карактеристичних n-грама. Отуда је у даљем приказу алгоритма коришћено ограничење само броја карактеристичних n-грама. У свим класама прецизност је већа од 90% уколико се се придруживање ЦОГ категорије

протеину врши тек када се у њему нађу више од 4 карактеристична n-грама. Стога је у приказаном класификационом моделу узето да је 4 минималан број неопходних карактеристичних n-грама да би протеин био придружен одговарајућој ЦОГ категорији. У Табели 4.14 приказане су вредности параметара квалитета класификационог модела (тачност (t), прецизност(p) и одзив (o)) за класификацију протеина из тест скупа података класе бактерија *Chlamydiales* за сваку ЦОГ категорију.

4.5 Поређење образаца добијених на различитим фамилијама

Претходно описани резултати представљају одвојене калсификационе моделе за сваку класу бактерија. Посебни модели по класама бактерија имају оправдање, јер је еволуцијом дошло до спецификације протеина који припадају одговарајућој бактерији, те су и сами ортологи протеини који припадају истој класи бактерија сличнији међусобно него њима ортологим протеинима из других класа бактерија. Међутим, због малог броја протеина у једној класи бактерија који припадају истом кластеру ортологих протеина, добија се мањи број образаца који су карактеристични за одговарајућу ЦОГ категорију. Удруживање добијених образаца може бити корисно за конструкцију модела који може да класификује већи скуп некласификованих протеина.

Табела 4.15: Број различитих образаца груписаних по броју ЦОГ категорија којима су придружени

#ЦОГ категорија	1	2	3	%противуречних
#дескриптора	62758	1166	18	2%
#п_дескриптора	16093	321	5	2%

Анализом је утврђено да у 98% образаца нема противуречности у различитим класама бактерија. Односно, постоји мање од 2% образаца који су у једној класи бактерија придружени једној ЦОГ категорији а у другој класи бактерија другој ЦОГ категорији (Табела 4.15), при чему је већина таквих образаца

протиувречна јер се налази у две категорије.[1] С друге стране број истих образаца који су издвојени у више класа бактерија и придружени једној ЦОГ категорији је такође мали. Број различитих образаца који се јављају у 1, 2, ..., 5 класа приказан је у Табели 4.16, уз напомену да је овде реч о обрасцима који су на тренинг подацима из одговарајуће класе бактерија потврђени са више од 5 карактеристичних n-грама ($min\_\sigma = 5$).

Табела 4.16: Број различитих образаца груписаних по броју класа бактерија у којима се појављују

#класа	1	2	3	4	5
#дескриптора	62137	558	56	3	4
#п_дескриптора	15987	100	6	0	0

Класификациони модел направљен над свим класама бактерија даје већи број тачно класификованих протеина али и већи број нетачно класификованих протеина.

Табела 4.17: Квалитет класификационог модела направљеног над свим класама бактерија у зависности од минималног броја (h) пронађених карактеристичних n-грама при додели одговарајуће ЦОГ категорије, при чему је апсолутна подршка образаца коришћених у моделу већа од 5 ($min\_\sigma = 5$)

h	TP	NP	p	o
1	3399	4341	44%	64%
2	2929	3783	44%	55%
3	2544	3158	45%	48%
4	2218	2583	46%	42%
5	2000	2082	49%	38%
6	1831	1613	53%	35%
7	1697	1274	57%	32%
8	1567	983	61%	30%
9	1490	801	65%	28%
10	1414	619	70%	27%

[1] Већина, али не сви. Наиме, поједини протеини се по својој функцији класификују у већи број цогова (од 2 до 5) тако да постоје и обрасци који истовремено карактеришу више од једне ЦОГ категорије.

Табела 4.18: Квалитет класификационог модела направљеног над свим класама бактерија у зависности од минималног броја (h) пронађених карактеристичних n-грама при додели одговарајуће ЦОГ категорије, при чему је апсолутна подршка образаца коришћених у моделу већа од 10 ($min\_\sigma = 10$)

h	TP	NP	p	o
1	1969	3042	39%	37%
2	1606	1862	46%	30%
3	1364	1214	53%	26%
4	1199	720	62%	23%
5	1063	498	68%	20%
6	963	304	76%	18%
7	865	221	80%	16%
8	784	151	84%	15%
9	717	112	86%	14%
10	670	86	89%	13%

Резултати класификације над тест подацима класе *Chlamydiales* за различите вредности прага h, броја пронађених карактеристичних n-грама при додељивању протеинима одговарајуће ЦОГ категорије су приказани у Табелама 4.17 и 4.18. Скуп правила представља скуп образаца из свих класа бактерија који су у одговарајућој класи потврђени са више од унапред одређеног броја карактеристичних n-грама (min_σ) и нису се јавили као карактеристике неких других ЦОГ категорија у другим класама. У Табели 4.17 је $min\_\sigma = 5$ а у Табели 4.18 је $min\_\sigma = 10$.

Поређењем резултата из претходних табела са резултатима у Табели 4.12 и Табели 4.13 може се закључити да нема значајног побољшања у квалитету модела. Метода класификације заснована на обрасцима из само једне класе бактерија има већи степен поузданости док се са комбинацијом свих образаца могу класификовати још неки додатни протеини. У табели 4.19 приказани су резултати квалитета класификације протеина из скупа за тестирање класе *Chlamydiales* по ЦОГ категоријама. Вредности прагова минималног бројача подршке и пронађених карактеристичних n-грама у класификованим протеинима су постављени на $min\_\sigma = 10$, $h = 10$. Вредности параметара за одређивање квалитета класификационог модела за сваку од класа појединачно приказани су у додатку B.

Табела 4.19: Квалитет модела посебно за сваку ЦОГ категорију

ЦОГ	NP	TP	NN	TN	t	p	o
A	0	0	0	5301	100%	-	-
B	0	0	6	5295	100%	-	-
C	50	87	224	4940	95%	64%	28%
D	4	21	50	5226	99%	84%	30%
E	41	43	347	4870	93%	51%	11%
F	6	23	120	5152	98%	79%	16%
G	0	62	232	5007	96%	100%	21%
H	12	19	232	5038	95%	61%	8%
I	0	25	196	5080	96%	100%	11%
J	23	205	426	4647	92%	90%	32%
K	3	31	146	5121	97%	91%	18%
L	28	87	318	4868	93%	76%	21%
M	40	47	351	4863	93%	54%	12%
N	0	0	84	5217	98%	-	0%
O	9	58	183	5051	96%	87%	24%
P	13	19	179	5090	96%	59%	10%
Q	0	0	84	5217	98%	-	0%
R	95	61	562	4583	88%	39%	10%
S	27	40	307	4927	94%	60%	12%
T	0	11	141	5149	97%	100%	7%
U	0	22	187	5092	96%	100%	11%
V	0	2	63	5236	99%	100%	3%

4.5.1 Резултати предиктора протеина по функционалним категоријама

Један од основних циљева рада је конструкција модела који би омогућио одређивање ЦОГ категорије протеина којима до сада (другим методама) категорија није одређена. Представљени предиктор (Алгоритам 3) прави класификациони модел (Алгоритам 2) према постављеним захтевима, које дефинише корисник, за квалитет модела. Предиктор се заснива на класификационом моделу при чему су у скуп правила укључени само обрасци из класе организама којој припада протеин који треба класификовати, или скуп образаца може да садржи обрасце из свих обрађених класа организама. Обрасци морају да задовоље одговарајући квалитет да би и сам модел могао са одговарајућом поузданошћу да придружи функционалну категорију протеину.

Предиктор је примењен на све некласификоване протеине, укупно њих 99, из обрађених класа бактерија. За ове протеине је одређено да јесу у некој ЦОГ категорији, али методама које су коришћене у НЦБИ није до тренутка скидања материјала била одређена конкретна ЦОГ категорија, већ су уместо ознаке категорије имали '*'. Расподела протеина по класама организама је била следећа: 1 протеин класе *Aquificales*, 15 протеина класе *Bacteroidia*, 9 протеина класе *Chlamydiales*, 14 протеина класе *Chloroflexia*, 4 протеина класе *Chlorobia*, 33 протеина класе *Cytophagia*, 10 протеина класе *Deinococci* и 13 протеина класе *Prochlorales*. Над протеинима су примењена три класификациона модела имплементирана у облику три различите варијанте предиктора:

- Предиктор 1 ($P1$) је заснован на класификационом моделу који за скуп образаца узима све неконтрадикторне обрасце из свих обрађених класа организама чија је апсолутна подршка најмање 5.

- Предиктор 2 ($P2$) је заснован на класификационом моделу који за скуп образаца узима све неконтрадикторне обрасце из свих обрађених класа организама чија је апсолутна подршка најмање 10.

- Предиктор 3 ($P3$) је заснован на класификационом моделу који за скуп образаца узима само обрасце из класе организама којој припада протеин кога треба класификовати, при чему је апсолутна подршка образаца најмање 5.

За одређивање ЦОГ категорије издвојеним некласификованим протеинима (горе наведеног скупа од 99 протеина) коришћена су сва три предиктора. Као резултат узето је најпрецизније предвиђање у унији резултата сва три предиктора. Укупно је класификовано 83 протеина у 14 различитих категорија од тога 27 са позданошћу већом од 70%. Неким протеинима су додељене две категорије што није необично јер се и у досадашњим резултатима могу пронаћи протеини који припадају двема категоријама. Табела 4.20 приказује број протеина придружених одговарајућој ЦОГ категорији.

Табела 4.20: Број нових протеина придружених одговарајућој ЦОГ категорији

ЦОГ категорија	J	R	C	M	G	S	L	O	E	Q	P	F	H	V
Број придружених протеина	16	16	12	10	10	8	7	3	2	2	1	1	1	0

Различити предиктори су за 14 протеина донели различите одлуке о класификацији, за 67 протеина се два предиктора слажу и за 18 протеина сва три предиктора су донела исту одлуку. Неслагање предиктора не мора бити потпуно јер врло често за више ЦОГ категорија постоје обрасци пронађени у протеину али једним предиктором најпрецизније је придружена једна, другим друга категорија.

Табела 4.21: Резултати предвиђања ЦОГ категорије за некласификоване протеине

		предиктор1		предиктор2		предиктор3		додељена категорија	
заснован на обрасцима из:		свих класа		свих класа		одговарајуће класе			
min_σ		5		10		5			
PID	класа	ЦОГ	p	ЦОГ	p	ЦОГ	p	ЦОГ	p
124024617	Prochlorales	P	51%	P	86%	P	99%	P	99%
53712922	Bacteroidia	S	51%	S	86%	S	99%	S	99%
29345725	Bacteroidia	S	51%	R	86%	S	99%	S	99%
313203003	Bacteroidia	R	51%	R	86%	M	99%	M	99%
297620987	Chlamydiales	C	51%	J	57%	J	96%	J	96%
72382573	Prochlorales	J	51%	J	86%	C	96%	C	96%
163845851	Chloroflexia	R	51%	R	86%	N.C	0%	R	86%
29347912	Bacteroidia	J	51%	J	86%	N.C	0%	J	86%
313677233	Cytophagia	J	51%	J	86%	N.C	0%	J	86%
159903319	Prochlorales	R	45%	R	80%	N.C	0%	R	80%
163846415	Chloroflexia	J	51%	J	78%	N.C	0%	J	78%

Табела 4.21 приказује резултате сва три предиктора и крајње одлуке за неколико изабраних протеина. Детаљнији подаци о свим обрађеним некласификованим протеинима се налазе у додатку *Г*. Из табеле се може уочити да је прецизност највећа када се класификациони модел заснива на

обрасцима само из одговарајуће класе бактерија. Међутим свега 9 протеина је класификовано помоћу овог предиктора ($P3$). Даље је по прецизности од два преостала предиктора која се заснивају на обрасцима свих класа бољи онај чији обрасци имају апсолутну подршку најмање 10 ($P2$). Њиме је класификовано 28 протеина, док је предиктором $P1$ који има највећи одзив класификовано 83 протеина, али са најмањом прецизношћу (38%-51%). Помоћу ових класификација једном делу протеина се може придружити одређена ЦОГ категорија са релативно великом прецизношћу. Резултати класификације којима протеину није са задовољавајућом прецизношћу придружена конкретна ЦОГ категорија могу да служе као полазни подаци који се могу допунити методама поравнања са изабраним протеинима (протеинима којима су придружене ЦОГ категорије).

Поглавље 5

Закључак

У раду је представљен нови поступак за издвајање n-грама чија појава је специфична за протеине који припадају одређеним ЦОГ категоријама, као и модел за предвиђање ЦОГ категорије којој припада протеин на основу издвојених n-грама. Најважнији резултати који представљају научни допринос ове дисертације су:

- нова метода за анализу n-грамског садржаја протеинских секвенци заснована на Буловој алгебри, која захтева мање меморијских ресурса и краће време обраде од класичних метода n-грамске анализе;
- нова метода за функционалну класификацију протеина која одређује ЦОГ категорију протеина без поређења са другим протеинским секвенцама;
- део протеина из скупа изабраних класа прокариота које нису класификоване досадашњим методама је придружен одговарајућим ЦОГ категоријама;
- допринос теорији генерализованих система Булових једначина (решење система k Булових неједначина и решење дисјункције Булових једначина).

Познати алгоритми n-грамске анализе врше статистичку анализу преклапајућих n-грама. Ово су једноставни алгоритми али њихова обрада података траје дуго и складиштење захтева велике меморијске ресурсе. Приказани алгоритам је вид димензионе редукције заснован на пресликавању скупа преклапајућих n-грама у скуп ОАК ниски чија је димензија значајно

мања. Ова трансформација омогућава брзо издвајање скупа карактеристичних n-грама за одговарајућу ЦОГ категорију. Добијени секвенцијални обрасци су показали висок степен прецизности над подацима за тестирање. Како обрасци у протеинима нису увек правилни, у томе је већа предност представљене методе јер дозвољава шумове у обрасцима.

Над издвојеним скупом образаца карактеристичних за одговарајућу ЦОГ категорију направљен је модел за функционалну класификацију протеина. Време обраде познатих метода за функционалну класификацију протеина пропорционално је производу дужине секвенце коју треба класификовати и укупне дужине скупа већ класификованих секвенци. Представљени метод захтева време за класификацију пропорционално производу дужине секвенце коју треба класификовати и укупне дужине скупа карактеристичних n-грама (при чему је $n \in \{5, 6, ..., 10\}$). Утврђено да се најпрецизније класификација врши када се за модел узимају протеини из једне класе.

Предиктор заснован на предходном моделу примењен је на скуп некласификованих протеина из улазног скупа бактерија. Од 99 некласификованих протеина, новим моделом је 27 протеина класификовано у ЦОГ категорије са прецизношћу већом од 70%. За 60 протеина од преосталих из скупа некласификованих протеина је придружена ЦОГ категорија али са мањом прецизношћу па се може користити у комбинацији са резултатима других метода. Квалитет добијених резултата указује на то да сличан принцип може да се примени и на одређивање других карактеристика протеина.

Теореме до којих се дошло решавајући проблем издвајања секвенцијалних образаца представљају значајан допринос теорији генерализованих система Булових једначина. Њима је представљен алгоритам за решавање дисјункције Булових једначина и система Булових неједначина који су до сада били отворени проблеми.

Даљи рад на развоју модела биће усмерен на повећање прецизности модела који се конструише коришћењем протеина из различитих класа, и укључивање могућности за класификацију протеина који се налазе у више од једне ЦОГ категорије.

A Улазни скуп организама

Табела A.1: Број протеина по ЦОГ категоријама у скуповима за тренинг и тестирање

Класа	Chloroflexia		Cytophagia		Deinococci		Prochlorales	
Категорија	тренинг	тест	тренинг	тест	тренинг	тест	тренинг	тест
A	1	0	0		9	0	0	0
B	8	3	0		17	6	0	0
C	624	327	1214	640	751	311	1277	512
D	89	36	316	158	276	71	185	84
E	530	239	1635	729	924	390	1034	421
F	253	108	712	378	344	143	436	175
G	249	106	2033	700	617	294	601	249
H	474	207	1292	640	665	251	990	405
I	199	86	645	342	573	221	415	180
J	671	272	1603	878	1877	631	1072	414
K	241	109	1527	711	425	177	583	256
L	453	173	1752	877	1125	405	995	378
M	583	252	2582	1116	865	398	1343	562
N	272	90	71	46	238	84	49	36
O	377	160	862	475	620	241	596	270
P	357	168	1582	701	501	198	805	351
Q	77	35	229	124	110	84	182	88
R	717	351	2897	1375	1240	623	1717	759
S	469	218	1483	723	757	347	1153	506
T	299	145	1154	463	326	152	427	261
U	290	127	572	274	618	209	346	167
V	66	42	694	292	61	65	287	124
Z	0	0	0	0	0	0	0	2
W	0	0	0	0	4	0	0	0
збирно	7299	3254	24855	11642	12943	5301	14493	6200

Табела A.2: Број протеина по ЦОГ категоријама у скуповима за тренинг и тестирање

Класа	Aquificales		Bacteroidia		Chlamydiales		Chlorobia	
Категорија	тренинг	тест	тренинг	тест	тренинг	тест	тренинг	тест
A	3	0	9	0	6	2	0	0
B	10	6	3	2	15	7	0	0
C	839	422	849	385	1288	628	733	344
D	129	58	163	65	276	112	139	65
E	1132	532	1214	577	2156	934	1063	512
F	308	149	374	158	702	297	400	202
G	945	455	1154	617	1485	596	600	247
H	659	327	742	309	996	471	910	442
I	477	212	642	257	796	350	329	158
J	658	322	949	335	1391	612	1105	531
K	1006	400	1369	557	1241	513	372	162
L	808	327	992	365	1233	431	593	272
M	962	443	1583	705	987	455	861	377
N	69	34	68	25	194	88	58	25
O	533	234	669	255	859	378	656	321
P	642	322	1013	542	1092	454	490	230
Q	359	131	391	173	465	181	221	93
R	2084	880	2400	1117	2901	1203	1275	588
S	1176	425	1580	683	1886	780	888	367
T	1261	562	1228	471	925	466	254	109
U	203	70	285	139	399	194	266	120
V	298	114	446	201	331	145	172	65
Z	7	2	7	0	3	1	2	0
W	0	0	0	0	0	0	0	0
збирно	14568	6427	18130	7938	21627	9298	11387	5230

Табела А.3: Подаци о организмима обрађених класа бактерија

тренинг скуп генома класе "Aquificales"			
НЦБИ ознака	ИМЕ ОРГАНИЗМА	GC садржај	ДУЖИНА
NC_000918	Aquifex aeolicus VF5	43.3	1551335
NC_001880	Aquifex aeolicus VF5	43.3	39456
NC_010730	Sulfurihydrogenibium sp. YO3AOP1	32.02	1838442
NC_011126	Hydrogenobaculum sp. Y04AAS1	34.84	1559514
NC_012438	Sulfurihydrogenibium azorense Az-Fu1	32.75	1640877
NC_012439	Persephonella marina EX-H1	37.08	53682
NC_013894	Thermocrinis albus DSM 14484	46.93	1500577

тест скуп генома класе "Aquificales"			
НЦБИ ознака	ИМЕ ОРГАНИЗМА	GC садржај	ДУЖИНА
NC_012440	Persephonella marina EX-H1	37.08	1930284
NC_013799	Hydrogenobacter thermophilus TK-6	44	1743135

тренинг скуп генома класе "Bacteroidia"			
НЦБИ ознака	ИМЕ ОРГАНИЗМА	GC садржај	ДУЖИНА
NC_002950	Porphyromonas gingivalis W83	48.28	2343476
NC_003228	Bacteroides fragilis NCTC 9343	43.11	5205140
NC_004663	Bacteroides thetaiotaomicron VPI-5482	42.86	6260361
NC_004703	Bacteroides thetaiotaomicron VPI-5482	42.86	33038
NC_006297	Bacteroides fragilis YCH46	43.2	33716
NC_006347	Bacteroides fragilis YCH46	43.2	5277274
NC_006373	Bacteroides uniformis	40.23	10276
NC_006873	Bacteroides fragilis NCTC 9343	43.11	36560
NC_009614	Bacteroides vulgatus ATCC 8482	42.2	5163189
NC_009615	Parabacteroides distasonis ATCC 8503	45.05	4811379
NC_010729	Porphyromonas gingivalis ATCC 33277	48.35	2354886
NC_011561	Candidatus Azobacteroides pseudotrichonymphae genomovar. CFP2	32.94	37111
NC_011562	Candidatus Azobacteroides pseudotrichonymphae genomovar. CFP2	32.94	31893
NC_011563	Candidatus Azobacteroides pseudotrichonymphae genomovar. CFP2	32.94	4149
NC_011564	Candidatus Azobacteroides pseudotrichonymphae genomovar. CFP2	32.94	37560
NC_011565	Candidatus Azobacteroides pseudotrichonymphae genomovar. CFP2	32.94	1114206
NC_014033	Prevotella ruminicola 23	47.68	3619559
NC_014370	Prevotella melaninogenica ATCC 25845	40.98	1796408
NC_014371	Prevotella melaninogenica ATCC 25845	40.98	1371874
NC_014734	Paludibacter propionicigenes WB4	38.85	3685504

тест скуп генома класе "Bacteroidia"			
НЦБИ ознака	ИМЕ ОРГАНИЗМА	GC садржај	ДУЖИНА
NC_014933	Bacteroides helcogenes P 36-108	44.71	3998906
NC_015160	Odoribacter splanchnicus DSM 20712	43.35	4392288
NC_015164	Bacteroides salanitronis DSM 18170	46.49	4242803
NC_015165	Bacteroides salanitronis DSM 18170	46.49	40303
NC_015166	Bacteroides salanitronis DSM 18170	46.49	6277
NC_015168	Bacteroides salanitronis DSM 18170	46.49	19280
NC_015311	Prevotella denticola F0289	50.36	2937589
NC_015501	Porphyromonas asaccharolytica DSM 20707	52.46	2186370
NC_015571	Porphyromonas gingivalis TDC60	48.34	2339898

тренинг скуп генома класе "Chlamydiales"			
НЦБИ ознака	ИМЕ ОРГАНИЗМА	GC садржај	ДУЖИНА
NC_000117	Chlamydia trachomatis D/UW-3/CX	41.3	1042519
NC_000922	Chlamydophila pneumoniae CWL029	40.57	1230230
NC_002179	Chlamydophila pneumoniae AR39	40.57	1229853
NC_002182	Chlamydia muridarum str. Nigg	40.3	7501
NC_002491	Chlamydophila pneumoniae J138	40.58	1226565
NC_002620	Chlamydia muridarum str. Nigg	40.3	1072950
NC_003361	Chlamydophila caviae GPIC	39.18	1173390
NC_004552	Chlamydophila abortus S26/3	39.86	1144377
NC_004720	Chlamydophila caviae GPIC	39.18	7966
NC_005043	Chlamydophila pneumoniae TW-183	40.57	1225935
NC_005861	Candidatus Protochlamydia amoebophila UWE25	34.71	2414465
NC_007429	Chlamydia trachomatis A/HAR-13	41.26	1044459
NC_007430	Chlamydia trachomatis A/HAR-13	41.26	7510
NC_007899	Chlamydophila felis Fe/C-56	39.34	1166239
NC_007900	Chlamydophila felis Fe/C-56	39.34	7552
NC_010280	Chlamydia trachomatis L2b/UCH-1/proctitis	41.32	1038863
NC_010287	Chlamydia trachomatis 434/Bu	41.32	1038842
NC_012686	Chlamydia trachomatis B/Jali20/OT	41.29	1044352
NC_012687	Chlamydia trachomatis B/TZ1A828/OT	41.3	1044282
NC_014225	Waddlia chondrophila WSU 86-1044	43.73	2116312

тест скуп генома класе "Chlamydiales"			
НЦБИ ознака	ИМЕ ОРГАНИЗМА	GC садржај	ДУЖИНА
NC_014226	Waddlia chondrophila WSU 86-1044	43.73	15593
NC_015217	Chlamydia psittaci 6BC	39.02	7553
NC_015408	Chlamydophila pecorum E58	41.07	1106197
NC_015470	Chlamydia psittaci 6BC	39.02	1171660
NC_015702	Parachlamydia acanthamoebae UV-7	39.02	3072383
NC_015710	Simkania negevensis Z	41.6	132038
NC_015713	Simkania negevensis Z	41.6	2496337
NC_015744	Chlamydia trachomatis L2c	41.32	1038313

тренинг скуп генома класе "Chlorobia"			
НЦБИ ознака	ИМЕ ОРГАНИЗМА	GC садржај	ДУЖИНА
NC_002932	Chlorobium tepidum TLS	56.52	2154946
NC_007512	Chlorobium luteolum DSM 273	57.33	2364842
NC_007514	Chlorobium chlorochromatii CaD3	44.27	2572079
NC_008639	Chlorobium phaeobacteroides DSM 266	48.35	3133902
NC_009337	Chlorobium phaeovibrioides DSM 265	52.99	1966858
NC_010803	Chlorobium limicola DSM 245	51.31	2763181
NC_010831	Chlorobium phaeobacteroides BS1	48.92	2736403
NC_011027	Chlorobaculum parvum NCIB 8327	55.79	2289249

тест скуп генома класе "Chlorobia"			
НЦБИ ознака	ИМЕ ОРГАНИЗМА	GC садржај	ДУЖИНА
NC_011026	Chloroherpeton thalassium ATCC 35110	45.03	3293456
NC_011059	Prosthecochloris aestuarii DSM 271	50.1	2512923
NC_011060	Pelodictyon phaeoclathratiforme BU-1	48.07	3018238
NC_011061	Prosthecochloris aestuarii DSM 271	50.1	66772

тренинг скуп генома класе "Chloroflexia"			
НЦБИ ознака	ИМЕ ОРГАНИЗМА	GC садржај	ДУЖИНА
NC_009523	Roseiflexus sp. RS-1	60.38	5801598
NC_009767	Roseiflexus castenholzii DSM 13941	60.69	5723298
NC_009972	Herpetosiphon aurantiacus DSM 785	50.89	6346587
NC_009973	Herpetosiphon aurantiacus DSM 785	50.89	339639
NC_009974	Herpetosiphon aurantiacus DSM 785	50.89	99204
NC_011830	Chloroflexus aggregans DSM 9484	56.42	4684931

тест скуп генома класе "Chloroflexia"

НЦБИ ознака	ИМЕ ОРГАНИЗМА	GC садржај	ДУЖИНА
NC_010175	Chloroflexus aurantiacus J-10-fl	56.69	5258541
NC_012032	Chloroflexus sp. Y-400-fl	56.68	5268950

тренинг скуп генома класе "Cytophagia"

НЦБИ ознака	ИМЕ ОРГАНИЗМА	GC садржај	ДУЖИНА
NC_008255	Cytophaga hutchinsonii ATCC 33406	38.84	4433218
NC_010830	Candidatus Amoebophilus asiaticus 5a2	35.04	1884364
NC_013037	Dyadobacter fermentans DSM 18053	51.54	6967790
NC_013730	Spirosoma linguale DSM 74	50.14	8078757
NC_013731	Spirosoma linguale DSM 74	50.14	189452
NC_013732	Spirosoma linguale DSM 74	50.14	146936
NC_013733	Spirosoma linguale DSM 74	50.14	36434
NC_013734	Spirosoma linguale DSM 74	50.14	9965
NC_013735	Spirosoma linguale DSM 74	50.14	8651
NC_013736	Spirosoma linguale DSM 74	50.14	7683
NC_013737	Spirosoma linguale DSM 74	50.14	7308
NC_014655	Leadbetterella byssophila DSM 17132	40.41	4059653
NC_014750	Marivirga tractuosa DSM 4126	35.51	4916
NC_014759	Marivirga tractuosa DSM 4126	35.51	4511574

тест скуп генома класе "Cytophagia"

НЦБИ ознака	ИМЕ ОРГАНИЗМА	GC садржај	ДУЖИНА
NC_013738	Spirosoma linguale DSM 74	50.14	6072
NC_015693	Runella slithyformis DSM 19594	46.42	106999
NC_015694	Runella slithyformis DSM 19594	46.42	66926
NC_015695	Runella slithyformis DSM 19594	46.42	38784
NC_015703	Runella slithyformis DSM 19594	46.42	6568739
NC_015704	Runella slithyformis DSM 19594	46.42	93527
NC_015705	Runella slithyformis DSM 19594	46.42	44754
NC_015914	Cyclobacterium marinum DSM 745	38.14	6221273

тренинг скуп генома класе "Deinococci"

НЦБИ ознака	ИМЕ ОРГАНИЗМА	GC садржај	ДУЖИНА
NC_000958	Deinococcus radiodurans R1	66.61	177466
NC_000959	Deinococcus radiodurans R1	66.61	45704
NC_001263	Deinococcus radiodurans R1	66.61	2648638
NC_001264	Deinococcus radiodurans R1	66.61	412348
NC_005835	Thermus thermophilus HB27	69.4	1894877
NC_005838	Thermus thermophilus HB27	69.4	232605
NC_006461	Thermus thermophilus HB8	69.49	1849742
NC_006462	Thermus thermophilus HB8	69.49	256992
NC_006463	Thermus thermophilus HB8	69.49	9322
NC_008010	Deinococcus geothermalis DSM 11300	66.47	574127
NC_008025	Deinococcus geothermalis DSM 11300	66.47	2467205
NC_009939	Deinococcus geothermalis DSM 11300	66.47	205686
NC_012526	Deinococcus deserti VCD115	62.97	2819842
NC_012527	Deinococcus deserti VCD115	62.97	324711
NC_012528	Deinococcus deserti VCD115	62.97	396459
NC_012529	Deinococcus deserti VCD115	62.97	314317
NC_013946	Meiothermus ruber DSM 1279	63.38	3097457
NC_014212	Meiothermus silvanus DSM 9946	62.71	3249394
NC_014213	Meiothermus silvanus DSM 9946	62.71	347854
NC_014214	Meiothermus silvanus DSM 9946	62.71	124421
NC_014221	Truepera radiovictrix DSM 17093	68.13	3260398
NC_014753	Oceanithermus profundus DSM 14977	69.81	135351
NC_015161	Deinococcus proteolyticus MRP	65.63	2147060

тест скуп генома класе "Deinococci"			
НЦБИ ознака	ИМЕ ОРГАНИЗМА	GC садржај	ДУЖИНА
NC_014761	Oceanithermus profundus DSM 14977	69.81	2303940
NC_014958	Deinococcus maricopensis DSM 21211	69.82	3498530
NC_014974	Thermus scotoductus SA-01	64.88	2346803
NC_014975	Thermus scotoductus SA-01	64.88	8383
NC_015162	Deinococcus proteolyticus MRP	65.63	195800
NC_015163	Deinococcus proteolyticus MRP	65.63	97188
NC_015169	Deinococcus proteolyticus MRP	65.63	314518
NC_015170	Deinococcus proteolyticus MRP	65.63	132270
NC_015387	Marinithermus hydrothermalis DSM 14884	68.07	2269167

тренинг скуп генома класе "Prochlorales"			
НЦБИ ознака	ИМЕ ОРГАНИЗМА	GC садржај	ДУЖИНА
NC_005042	Prochlorococcus marinus subsp. marinus str. CCMP1375	36.44	1751080
NC_005071	Prochlorococcus marinus str. MIT 9313	50.73	2410873
NC_005072	Prochlorococcus marinus subsp. pastoris str. CCMP1986	30.79	1657990
NC_007335	Prochlorococcus marinus str. NATL2A	35.12	1842899
NC_007577	Prochlorococcus marinus str. MIT 9312	31.21	1709204
NC_008816	Prochlorococcus marinus str. AS9601	31.32	1669886
NC_008817	Prochlorococcus marinus str. MIT 9515	30.79	1704176
NC_008820	Prochlorococcus marinus str. MIT 9303	50	2682675

тест скуп генома класе "Prochlorales"			
НЦБИ ознака	ИМЕ ОРГАНИЗМА	GC садржај	ДУЖИНА
NC_008819	Prochlorococcus marinus str. NATL1A	34.97	1864731
NC_009091	Prochlorococcus marinus str. MIT 9301	31.33	1641879
NC_009840	Prochlorococcus marinus str. MIT 9215	31.14	1738790
NC_009976	Prochlorococcus marinus str. MIT 9211	38	1688963

Б Карактеристике класификационог модела за одређену класу бактерија

Табела Б.1: Карактеристике класификационог модела изражене преко прецизности и одзива добијене при класификацији протеина у одговарајућу ЦОГ категорију у зависности од минималног броја различитих карактеристичних n-грама (дескриптора) пронађених у протеину (h)

Aquificales									
карактеристични n-грами					дескриптори				
h	*TP*	*NP*	*p*	*o*	*h*	*TP*	*NP*	*p*	*o*
1	482	235	67%	15%	1	349	161	68%	11%
2	362	81	82%	11%	2	231	27	90%	7%
3	282	50	85%	9%	3	154	9	94%	5%
4	237	21	92%	7%	4	111	1	99%	3%
5	200	15	93%	6%	5	87	1	99%	3%
6	161	5	97%	5%	6	75	0	100%	2%
7	136	3	98%	4%	7	53	0	100%	2%
8	115	3	97%	4%	8	42	0	100%	1%
9	96	2	98%	3%	9	32	0	100%	1%
10	79	1	99%	2%	10	26	0	100%	1%

Bacteroidia									
карактеристични n-грами					дескриптори				
h	*TP*	*NP*	*p*	*o*	*h*	*TP*	*NP*	*p*	*o*
1	1609	768	0.676904	14%	1	1300	543	71%	11%
2	1167	209	0.84811	10%	2	841	81	91%	7%
3	876	127	0.87338	8%	3	594	50	92%	5%
4	690	42	0.942623	6%	4	457	3	99%	4%
5	511	28	0.948052	4%	5	364	3	99%	3%
6	399	12	0.970803	3%	6	261	0	100%	2%
7	319	11	0.966667	3%	7	213	0	100%	2%
8	243	7	0.972	2%	8	166	0	100%	1%
9	208	5	0.976526	2%	9	142	0	100%	1%
10	176	3	0.98324	2%	10	121	0	100%	1%

Chlamydiales									
карактеристични n-грами					дескриптори				
h	*TP*	*NP*	*p*	*o*	*h*	*TP*	*NP*	*p*	*o*
1	1215	494	71%	23%	1	974	302	76%	18%
2	956	160	86%	18%	2	696	59	92%	13%
3	784	73	91%	15%	3	523	23	96%	10%
4	688	29	96%	13%	4	433	1	100%	8%
5	596	15	98%	11%	5	351	1	100%	7%
6	507	8	98%	9%	6	291	0	100%	5%
7	446	8	98%	8%	7	241	0	100%	5%
8	390	4	99%	7%	8	205	0	100%	4%
9	346	3	99%	6%	9	171	0	100%	3%
10	296	0	100%	6%	10	137	0	100%	3%

Chlorobia									
карактеристични n-грами					дескриптори				
h	*TP*	*NP*	*p*	*o*	*h*	*TP*	*NP*	*p*	*o*
1	1430	870	62%	23%	1	1248	626	67%	20%
2	1131	337	77%	18%	2	956	155	86%	15%
3	935	150	86%	15%	3	783	57	93%	13%
4	806	61	93%	13%	4	664	15	98%	11%
5	697	39	95%	11%	5	564	11	98%	9%
6	614	19	97%	10%	6	494	6	99%	8%
7	543	12	98%	9%	7	443	5	99%	7%
8	487	6	99%	8%	8	374	2	99%	6%
9	430	6	99%	7%	9	334	2	99%	5%
10	389	5	99%	6%	10	298	2	99%	5%

Chloroflexia									
карактеристични n-грами					дескриптори				
h	*TP*	*NP*	*p*	*o*	*h*	*TP*	*NP*	*p*	*o*
1	503	117	0.81129	8%	1	299	65	82%	5%
2	319	36	0.898592	5%	2	166	12	93%	3%
3	246	22	0.91791	4%	3	86	4	96%	1%
4	168	8	0.954545	3%	4	40	2	95%	1%
5	128	4	0.969697	2%	5	26	2	93%	0%
6	94	2	0.979167	1%	6	22	0	100%	0%
7	74	2	0.973684	1%	7	14	0	100%	0%
8	48	2	0.96	1%	8	8	0	100%	0%
9	38	0	1	1%	9	8	0	100%	0%
10	32	0	1	0%	10	8	0	100%	0%

Cytophagia									
карактеристични n-грами					дескриптори				
h	*TP*	*NP*	*p*	*o*	*h*	*TP*	*NP*	*p*	*o*
1	280	152	0.648148	4%	1	171	8	96%	2%
2	176	28	0.862745	2%	2	95	3	97%	1%
3	128	13	0.907801	2%	3	59	0	100%	1%
4	89	5	0.946809	1%	4	36	0	100%	0%
5	68	2	0.971429	1%	5	22	0	100%	0%
6	44	2	0.956522	1%	6	15	0	100%	0%
7	30	1	0.967742	0%	7	13	0	100%	0%
8	22	1	0.956522	0%	8	9	0	100%	0%
9	19	1	0.95	0%	9	8	0	100%	0%
10	15	0	1	0%	10	6	75	7%	0%

Deinococci									
карактеристични n-грами					дескриптори				
h	*TP*	*NP*	*p*	*o*	*h*	*TP*	*NP*	*p*	*o*
1	1055	353	75%	13.3%	1	832	219	79%	10%
2	781	116	87%	9.8%	2	571	40	93%	7%
3	607	65	90%	7.6%	3	414	21	95%	5%
4	493	19	96%	6.2%	4	317	3	99%	4%
5	391	13	97%	4.9%	5	246	2	99%	3%
6	333	7	98%	4.2%	6	205	0	100%	3%
7	279	6	98%	3.5%	7	166	0	100%	2%
8	222	3	99%	2.8%	8	142	0	100%	2%
9	193	2	99%	2.4%	9	112	0	100%	1%
10	159	2	99%	2.0%	10	85	0	100%	1%

Prochlorales									
карактеристични n-грами					дескриптори				
h	*TP*	*NP*	*p*	*o*	*h*	*TP*	*NP*	*p*	*o*
1	1410	83	94%	27%	1	299	65	82%	6%
2	1091	24	98%	21%	2	166	12	93%	3%
3	870	15	98%	17%	3	86	4	96%	2%
4	710	6	99%	14%	4	40	2	95%	1%
5	594	3	99%	11%	5	26	2	93%	0%
6	482	2	100%	9%	6	22	0	100%	0%
7	401	2	100%	8%	7	14	0	100%	0%
8	323	1	100%	6%	8	8	0	100%	0%
9	274	1	100%	5%	9	8	0	100%	0%
10	233	0	100%	4%	10	8	0	100%	0%

TP -тачно придружени ЦОГ категорији
NP -нетачно придружени ЦОГ категорији
p-прецизност
o-одзив

B Карактеристике класификационог модела за све врсте бактерија

Табела B.1: Карактеристике класификационог модела изражене преко прецизности и одзива добијене при класификацији протеина у одговарајућу ЦОГ категорију за различите вредности min_σ и h када се користе издвојени обрасци из свих обрађених класа бактерија

Aquificales									
min_σ=5					min_σ=10				
h	*TP*	*NP*	*p*	*o*	*h*	*TP*	*NP*	*p*	*o*
1	1488	2570	37%	46%	1	490	1310	27%	15.1%
2	1135	2078	35%	35%	2	260	607	30%	8.0%
3	878	1639	35%	27%	3	172	317	35%	5.3%
4	698	1195	37%	21%	4	113	126	47%	3.5%
5	562	904	38%	17%	5	84	84	50%	2.6%
6	465	633	42%	14%	6	68	32	68%	2.1%
7	385	475	45%	12%	7	54	27	67%	1.7%
8	322	324	50%	10%	8	43	15	74%	1.3%
9	277	253	52%	9%	9	39	13	75%	1.2%
10	231	170	58%	7%	10	33	7	83%	1.0%

Bacteroidia									
min_σ=5					min_σ=10				
h	*TP*	*NP*	*p*	*o*	*h*	*TP*	*NP*	*p*	*o*
1	6809	10163	40%	58%	1	2365	6892	26%	20.3%
2	5352	9250	37%	46%	2	1327	4140	24%	11.4%
3	4255	8267	34%	37%	3	821	2440	25%	7.1%
4	3389	7086	32%	29%	4	530	1321	29%	4.6%
5	2700	5998	31%	23%	5	386	848	31%	3.3%
6	2153	4938	30%	18%	6	299	421	42%	2.6%
7	1796	4088	31%	15%	7	239	287	45%	2.1%
8	1541	3275	32%	13%	8	201	142	59%	1.7%
9	1262	2653	32%	11%	9	174	106	62%	1.5%
10	1080	2080	34%	9%	10	145	55	73%	1.2%

Chlamydiales

min_σ=5					min_σ=10				
h	TP	*NP*	*p*	*o*	*h*	*TP*	*NP*	*p*	*o*
1	3399	4341	44%	64%	1	1969	3042	39%	37%
2	2929	3783	44%	55%	2	1606	1862	46%	30%
3	2544	3158	45%	48%	3	1364	1214	53%	26%
4	2218	2583	46%	42%	4	1199	720	62%	23%
5	2000	2082	49%	38%	5	1063	498	68%	20%
6	1831	1613	53%	35%	6	963	304	76%	18%
7	1697	1274	57%	32%	7	865	221	80%	16%
8	1567	983	61%	30%	8	784	151	84%	15%
9	1490	801	65%	28%	9	717	112	86%	14%
10	1414	619	70%	27%	10	670	86	89%	13%

Chlorobia

min_σ=5					min_σ=10				
h	TP	*NP*	*p*	*o*	*h*	*TP*	*NP*	*p*	*o*
1	3341	5260	39%	54%	1	1293	3169	29%	20.9%
2	2675	4558	37%	43%	2	797	1678	32%	12.9%
3	2207	3806	37%	36%	3	525	929	36%	8.5%
4	1829	3029	38%	30%	4	375	457	45%	6.0%
5	1550	2410	39%	25%	5	280	297	49%	4.5%
6	1291	1844	41%	21%	6	226	143	61%	3.6%
7	1117	1500	43%	18%	7	185	113	62%	3.0%
8	967	1090	47%	16%	8	152	61	71%	2.5%
9	844	872	49%	14%	9	122	47	72%	2.0%
10	737	677	52%	12%	10	110	18	86%	1.8%

Chloroflexia

min_σ=5					min_σ=10				
h	TP	NP	p	o	h	TP	NP	p	o
1	3444	5529.00	38%	54%	1	988	3398	23%	15.4%
2	2722	5121.00	35%	42%	2	522	1721	23%	8.1%
3	2184	4579.00	32%	34%	3	334	939	26%	5.2%
4	1777	3849.00	32%	28%	4	221	468	32%	3.4%
5	1439	3251.00	31%	22%	5	145	307	32%	2.3%
6	1229	2672.00	32%	19%	6	105	152	41%	1.6%
7	1038	2254.00	32%	16%	7	75	117	39%	1.2%
8	887	1800.00	33%	14%	8	56	60	48%	0.9%
9	778	1457.00	35%	12%	9	46	52	47%	0.7%
10	666	1171.00	36%	10%	10	42	28	60%	0.7%

Cytophagia

min_σ=5					min_σ=10				
h	TP	NP	p	o	h	TP	NP	p	o
1	3942	6753.00	37%	50%	1	1143	4252	21%	14.4%
2	2945	6029.00	33%	37%	2	550	2323	19%	6.9%
3	2196	5207.00	30%	28%	3	290	1364	18%	3.7%
4	1658	4305.00	28%	21%	4	180	687	21%	2.3%
5	1248	3536.00	26%	16%	5	124	429	22%	1.6%
6	990	2744.00	27%	12%	6	95	195	33%	1.2%
7	749	2199.00	25%	9%	7	75	132	36%	0.9%
8	588	1692.00	26%	7%	8	64	66	49%	0.8%
9	457	1350.00	25%	6%	9	54	44	55%	0.7%
10	370	997.00	27%	5%	10	47	21	69%	0.6%

Deinococci

min_σ=5					min_σ=10				
h	TP	NP	p	o	h	TP	NP	p	o
1	3831	7360	34%	41%	1	1185	2796	30%	12.7%
2	2801	5975	32%	30%	2	645	1124	36%	6.9%
3	2116	4694	31%	23%	3	405	561	42%	4.4%
4	1627	3460	32%	17%	4	249	227	52%	2.7%
5	1251	2546	33%	13%	5	188	155	55%	2.0%
6	1003	1765	36%	11%	6	140	52	73%	1.5%
7	806	1346	37%	9%	7	107	43	71%	1.2%
8	660	922	42%	7%	8	90	16	85%	1.0%
9	538	716	43%	6%	9	80	12	87%	0.9%
10	471	522	47%	5%	10	71	8	90%	0.8%

Prochlorales

min_σ=5					min_σ=10				
h	TP	NP	p	o	h	TP	NP	p	o
1	3649	4411	45%	70%	1	1721	2399	42%	32.9%
2	3287	3971	45%	63%	2	1202	1204	50%	23.0%
3	2970	3490	46%	57%	3	849	621	58%	16.2%
4	2672	2916	48%	51%	4	643	283	69%	12.3%
5	2445	2451	50%	47%	5	497	169	75%	9.5%
6	2240	1948	53%	43%	6	402	84	83%	7.7%
7	2077	1646	56%	40%	7	334	59	85%	6.4%
8	1931	1323	59%	37%	8	261	32	89%	5.0%
9	1792	1066	63%	34%	9	207	16	93%	4.0%
10	1650	843	66%	32%	10	182	6	97%	3.5%

TP -тачно придружени ЦОГ категорији
NP -нетачно придружени ЦОГ категорији
p-прецизност
o-одзив

Г Резултати предиктора на некласификованим протеинима

Табела Г.1: Табела са резултатима предвиђања ЦОГ категорије за некласификоване протеине

тип предиктора		предиктор1				предиктор2				предиктор3				додељена
min_σ		5				10				5				категорија
заснован на обрасцима из:		свих класа				свих класа				одговарајуће класе				
PID	класа	*m*	*#m*	*ЦОГ*	*p*	*m*	*#m*	*ЦОГ*	*p*	*m*	*#m*	*ЦОГ*	*p*	*ЦОГ*	*p*
225851239	Aquificales	7	1	L	42%	4	1	*	0%	0	24	*	0%	L	42%
53712922	Bacteroidia	112	1	S	51%	40	1	S	86%	86	1	S	99%	S	99%
29345725	Bacteroidia	72	1	S	51%	21	1	R	86%	33	1	S	99%	S	99%
313203003	Bacteroidia	96	1	R	51%	52	1	R	86%	35	1	M	99%	M	99%
29347912	Bacteroidia	19	1	J	51%	12	1	J	86%	0	24	*	0%	J	86%
150008213	Bacteroidia	10	1	E	51%	4	1	*	0%	0	24	*	0%	E	51%
60680604	Bacteroidia	9	1	J	48%	2	2	*	0%	0	24	*	0%	J	48%
313204703	Bacteroidia	7	1	S	42%	4	1	*	0%	1	1	*	0%	S	42%
313204789	Bacteroidia	7	1	M	42%	1	1	*	0%	0	24	*	0%	M	42%
150004398	Bacteroidia	6	1	G	40%	2	1	*	0%	0	24	*	0%	G	40%
53712651	Bacteroidia	5	1	M	38%	1	2	*	0%	0	24	*	0%	M	38%
53713960	Bacteroidia	5	1	R	38%	2	3	*	0%	0	24	*	0%	R	38%
60682164	Bacteroidia	5	1	R	38%	2	3	*	0%	0	24	*	0%	R	38%
302346946	Bacteroidia	4	3	*	0%	1	1	*	0%	2	1	*	0%	*	0%
313202883	Bacteroidia	4	1	*	0%	1	2	*	0%	0	24	*	0%	*	0%
313204577	Bacteroidia	4	1	*	0%	2	1	*	0%	0	24	*	0%	*	0%

тип предиктора		предиктор1				предиктор2				предиктор3				додељена категорија	
min_σ		5				10				5					
заснован на обрасцима из:		свих класа				свих класа				одговарајуће класе					
PID	класа	*m*	*#m*	*ЦОГ*	*p*	*m*	*#m*	*ЦОГ*	*p*	*m*	*#m*	*ЦОГ*	*p*	*ЦОГ*	*p*
338174222	Chlamydiales	6	1	L	40%	3	1	*	0%	2	1	*	0%	L	40%
338733187	Chlamydiales	5	1	S	38%	2	1	*	0%	3	1	*	0%	S	38%
46447209	Chlamydiales	57	1	S	51%	22	1	O	86%	3	1	*	0%	O	86%
46446611	Chlamydiales	3	2	*	0%	3	1	*	0%	1	2	*	0%	*	0%
297620987	Chlamydiales	11	1	C	51%	5	1	J	57%	5	1	J	96%	J	96%
15618718	Chlamydiales	12	1	J	51%	0	24	*	0%	2	2	*	0%	J	51%
15836342	Chlamydiales	12	1	J	51%	0	24	*	0%	2	2	*	0%	J	51%
33242169	Chlamydiales	12	1	J	51%	0	24	*	0%	2	2	*	0%	J	51%
62185319	Chlamydiales	11	1	R	51%	1	1	*	0%	3	2	*	0%	R	51%
145219874	Chlorobia	8	1	S	45%	3	1	*	0%	2	2	*	0%	S	45%
78189635	Chlorobia	5	1	L	38%	0	24	*	0%	4	1	*	0%	L	38%
189345562	Chlorobia	3	1	*	0%	2	2	*	0%	1	1	*	0%	*	0%
21673455	Chlorobia	3	2	*	0%	0	24	*	0%	0	24	*	0%	*	0%
163845851	Chloroflexia	47	1	R	51%	34	1	R	86%	0	24	*	0%	R	86%
163846415	Chloroflexia	16	1	J	51%	8	1	J	78%	0	24	*	0%	J	78%
222524182	Chloroflexia	16	1	J	51%	8	1	J	78%	0	24	*	0%	J	78%
159896723	Chloroflexia	14	2	J,C	51%	4	1	*	0%	1	1	*	0%	J,C	51%
159899498	Chloroflexia	10	1	L	51%	1	2	*	0%	0	24	*	0%	L	51%
159897018	Chloroflexia	8	1	O	45%	2	4	*	0%	0	24	*	0%	O	45%
159897834	Chloroflexia	8	1	C	45%	1	4	*	0%	0	24	*	0%	C	45%
159899201	Chloroflexia	8	1	C	45%	4	1	*	0%	0	24	*	0%	C	45%
219848992	Chloroflexia	7	1	G	42%	3	3	*	0%	1	1	*	0%	G	42%
159897117	Chloroflexia	6	2	J,F	40%	3	1	*	0%	0	24	*	0%	J,F	40%
159898356	Chloroflexia	5	1	M	38%	0	24	*	0%	0	24	*	0%	M	38%
159899014	Chloroflexia	26	1	C	51%	10	1	M	86%	1	1	*	0%	M	86%
222523563	Chloroflexia	47	1	U	51%	34	1	R	86%	0	24	*	0%	R	86%
159896601	Chloroflexia	3	1	*	0%	2	1	*	0%	0	24	*	0%	*	0%
313677233	Cytophagia	32	1	J	51%	13	1	J	86%	0	24	*	0%	J	86%
284038925	Cytophagia	16	1	J	51%	7	1	J	71%	0	24	*	0%	J	71%
294661182	Cytophagia	10	1	J	51%	7	1	J	71%	0	24	*	0%	J	71%
255036106	Cytophagia	5	1	C	38%	5	1	C	57%	0	24	*	0%	C	57%
189501811	Cytophagia	23	1	C	51%	4	2	*	0%	1	1	*	0%	C	51%
312132020	Cytophagia	11	1	M	51%	4	1	*	0%	0	24	*	0%	M	51%
338210234	Cytophagia	20	1	G	51%	3	3	*	0%	1	1	*	0%	G	51%
338212173	Cytophagia	10	1	C	51%	4	2	*	0%	0	24	*	0%	C	51%
343085206	Cytophagia	13	1	M	51%	4	1	*	0%	0	24	*	0%	M	51%
312132016	Cytophagia	9	1	G	48%	0	24	*	0%	1	1	*	0%	G	48%
343083435	Cytophagia	9	1	R	48%	3	1	*	0%	0	24	*	0%	R	48%
312132137	Cytophagia	8	3	V,R,H	45%	4	1	*	0%	0	24	*	0%	V,R,H	45%

тип предиктора		предиктор1				предиктор2				предиктор3				додељена категорија	
min_σ		5				10				5					
заснован на обрасцима из:		свих класа				свих класа				одговарајуће класе					
PID	класа	*m*	*#m*	*ЦОГ*	*p*	*m*	*#m*	*ЦОГ*	*p*	*m*	*#m*	*ЦОГ*	*p*	*ЦОГ*	*p*
343083434	Cytophagia	8	1	G	45%	4	2	*	0%	0	24	*	0%	G	45%
255039136	Cytophagia	7	1	C	42%	4	1	*	0%	0	24	*	0%	C	42%
294661313	Cytophagia	7	1	L	42%	3	2	*	0%	0	24	*	0%	L	42%
312131407	Cytophagia	7	2	J,E	42%	4	1	*	0%	0	24	*	0%	J,E	42%
313677346	Cytophagia	6	1	R	40%	2	3	*	0%	2	1	*	0%	R	40%
343087078	Cytophagia	6	1	S	40%	1	3	*	0%	0	24	*	0%	S	40%
110638895	Cytophagia	5	2	C	38%	3	1	*	0%	0	24	*	0%	C	38%
312129223	Cytophagia	5	1	G	38%	1	1	*	0%	0	24	*	0%	G	38%
338212844	Cytophagia	5	1	R	38%	1	5	*	0%	0	24	*	0%	R	38%
338213214	Cytophagia	5	1	G	38%	4	1	*	0%	0	24	*	0%	G	38%
343087529	Cytophagia	5	1	C	38%	2	1	*	0%	0	24	*	0%	C	38%
313677793	Cytophagia	38	1	M	51%	31	1	R	86%	0	24	*	0%	R	86%
338213576	Cytophagia	35	1	S	51%	32	1	G	86%	0	24	*	0%	G	86%
312131709	Cytophagia	18	2	M,R	51%	7	2	R,M	71%	0	24	*	0%	R,M	71%
313674656	Cytophagia	48	1	M	51%	7	1	G	71%	3	1	*	0%	G	71%
343083450	Cytophagia	23	1	M	51%	7	1	O	71%	0	24	*	0%	O	71%
284039112	Cytophagia	7	1	J	42%	5	1	R	57%	1	1	*	0%	R	57%
255034202	Cytophagia	3	2	*	0%	2	1	*	0%	1	1	*	0%	*	0%
312131298	Cytophagia	4	2	*	0%	2	2	*	0%	0	24	*	0%	*	0%
313675754	Cytophagia	4	3	*	0%	4	1	*	0%	0	24	*	0%	*	0%
338212338	Cytophagia	4	1	*	0%	3	1	*	0%	0	24	*	0%	*	0%
94984763	Deinococci	10	1	C	51%	6	1	C	68%	3	1	*	0%	C	68%
297567415	Deinococci	14	1	S	51%	3	2	*	0%	0	24	*	0%	S	51%
320335566	Deinococci	13	1	L	51%	4	1	*	0%	0	24	*	0%	L	51%
291297245	Deinococci	9	1	R	48%	2	1	*	0%	1	1	*	0%	R	48%
297564538	Deinococci	6	1	J	40%	1	1	*	0%	3	1	*	0%	J	40%
320335791	Deinococci	6	1	L	40%	4	2	*	0%	0	24	*	0%	L	40%
320334241	Deinococci	18	1	L	51%	5	1	M	57%	0	24	*	0%	M	57%
297567617	Deinococci	3	2	*	0%	1	3	*	0%	1	1	*	0%	*	0%
313680851	Deinococci	3	2	*	0%	2	1	*	0%	0	24	*	0%	*	0%
328951555	Deinococci	3	1	*	0%	2	1	*	0%	0	24	*	0%	*	0%
124024617	Prochlorales	64	1	P	51%	44	1	P	86%	60	1	P	99%	P	99%
72382573	Prochlorales	33	1	J	51%	17	1	J	86%	5	1	C	96%	C	96%

тип предиктора		предиктор1				предиктор2				предиктор3				додељена
min_σ		5				10				5				категорија
заснован на обрасцима из:		свих класа				свих класа				одговарајуће класе				
PID	класа	*m*	*#m*	*ЦОГ*	*p*	*m*	*#m*	*ЦОГ*	*p*	*m*	*#m*	*ЦОГ*	*p*	*ЦОГ*	*p*
33861788	Prochlorales	6	2	R,M	40%	3	1	*	0%	1	2	*	0%	R,M	40%
124024788	Prochlorales	5	1	R	38%	0	24	*	0%	1	1	*	0%	R	38%
33862529	Prochlorales	48	1	M	51%	36	1	S	86%	37	1	J	99%	J	99%
33864417	Prochlorales	33	1	T	51%	2	3	*	0%	21	1	Q	99%	Q	99%
124024522	Prochlorales	11	1	T	51%	2	5	*	0%	9	1	Q	98%	Q	98%
123965962	Prochlorales	10	1	E	51%	6	1	G	68%	2	1	*	0%	G	68%
126696810	Prochlorales	2	2	*	0%	0	24	*	0%	0	24	*	0%	*	0%
72382622	Prochlorales	1	1	*	0%	0	24	*	0%	0	24	*	0%	*	0%

m -максималан број пронађених n-грама карактеристичних за једну категорију
#m -број ЦОГ категорија за које је пронађено m карактеристичних n-грама
ЦОГ - ЦОГ категорије за које је пронађено m карактеристичних n-грама
p -прецизност предвиђања
* - није одређена категорија

Д Скраћенице коришћене у раду

Табела Д.1: Пуни називи скраћеница у раду као и скраћенице и називи њихових оригинала

Скраћеница	Назив	Оригинални назив	Оригинална скраћеница
ЦОГ	Кластери ортологих група	*Cluster of Orthhologous Groups*	*COGs*
КОГ	Еукариотске ортологе групе	*euKaryote Orthologous Groups*	*KOG*
СКОП	Структурна класификација протеина	*Structural Classification of Proteins*	*SCOP*
Пфам	Фамилије протеина	*Protein family*	*Pfam*
БЛАСТ	Претраживач основних локаних поравнања	*Basic Local Alignment Search Tool*	*BLAST*
ГенБанк	Банка гена	*Bank of genes*	*GenBank*
НЦБИ	Национални центар за биотехнолошке информације	*National Center for Biological Information*	*NCBI*
РефСик	Колекција нуклеотидних и протеинских секвенци	*Reference sequence collection*	RefSeq
СВИС-ПРОТ	Банка протеина ЕМБЛ-а и Швајцарског института за биоинформатику	*EMBL and the Swiss Institute of Bioinformatics protein sequence database*	*SWISS-PROT*
ПДБе	Протеинска банка података Европе	Protein Data Bank in Europe	PDBe
ППШ	Претрага по ширини	Breadth First Search	BFS
ППД	Претрага по дубини	Depth First Search	DFS
ОАК ниске	Основне аминокиселинске ниске	*Basic amino acid sequence*	*BAA sequence*
ГСБЈ	Генерализовани системи Булових једначина	*The generalized systems of Boolean equations*	*GSBE*
ГИ	Општи идентификатор	*GenInfo Identifier*	*GI*
ПИД	Општи идентификатор протеинске секвенце	*Pathway Interaction Database*	*PID*

Литература

[AZ07] R. Abdellatif and E. Zakaria. "Experimenting N-Grams in Text Categorization". *The International Arab Journal of Information Technology* 4(4) (2007), 377–385.

[AIS93] R. Agrawal, T. Imielinski, and A. Swami. "Mining association rules between sets of items in large databases" in *Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93*. Vol. 22(2). 1993, pp. 207–216.

[AS94] R. Agrawal and R. Srikant. "Fast algorithms for mining association rules in large databases". *Proc. of the International Conference on Very Large Databases (VLDB)* (1994), pp. 487–499.

[Alt+97] F. S. Altschul, T. L. Madden, A. A. Schäffer1, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". *Oxford University Press Nucleic Acids Research* 25(17) (1997), 3389–3402.

[Ban07] D. Banković. "Boolean inequations". *Discrete Mathematics* 307 (2007), pp. 750–755.

[Ban10] D. Banković. "Boolean equations and Boolean inequations". *Journal of Multiple-Valued Logic and Soft Computing* 16 (2010), pp. 189–196.

[BM15] D. Banković and U. Marovac. "System of two Boolean inequations". *Journal of Multiple-Valued Logic and Soft Computing* 24(5) (2015), pp. 521–528.

[Bat+00] A. Bateman, E. Birney, R. Durbin, S. R. Eddy, K. L. Howe, and E. L. L. Sonnhammer. "The Pfam protein families database". *Nucleic Acids Res.* 28 (2000), 263–266.

[BMS97] S. Brin, R. Motwani, and C. Silverstein. “Beyond market baskets: generalizing association rules to correlations” in *Proc. of ACM SIGMOD Int'l. Conference on Management of Data (SIGMOD)*. 1997, pp. 265–276.

[CT94] W. B. Cavnar and J. M. Trenkle. “n-Gram-based text categorization” in *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*. University of Nevada, Las Vegas. 1994, pp. 161–175.

[CCKS05] B. Cheng, J. Carbonell, and J. Klein-Seetharama. “Protein Classification Based on Text Document Classification Techniques”. *PROTEINS: Structure , Function, and Bioinformatics* 58 (2005), 955–997.

[Coh+00] E. Cohen, M. Datar, S. Fuijiwara, A. Cionis, P. Indyk, R. Motwani, J. Ullman, and C. Yang. “Finding interesting associations without support pruning” in 2000, pp. 489–499.

[FRM09] M. Faouzi, R. Ricco, and E. . Mourad. “A Hierarchical n-Grams Extraction Approach for Classification Problem”. 4879 (2009), pp. 211–222.

[FUM00] S. Fujiwara, J. Ullman, and R. Motwani. “Dynamic miss-counting algorithms: finding implication and similarity rules with confidence pruning” in 2000, pp. 501–511.

[Gan+02a] M. Ganapathiraju, D. Weisser, R. Rosenfeld, J. Carbonell, R. Reddy, and J. Klein-Seetharaman. “Comparative N-Gram Analysis of Whole-Genome Sequences” in *Proceedings of Human Language Technologies Conference*. California, USA. 2002.

[Gan+02b] M. Ganapathiraju, J. Klein-Seetharaman, N. Balakrishnan, R. Rosenfeld, J. Carbonell, and R. Reddy. “Rare and Frequent Amino Acid N-Grams in Whole-Genome Protein Sequences” in *IEEE Signal Processing magazine*. Washington, USA. 2002.

[Gan+04] M. Ganapathiraju, J. Klein-Seetharaman, N. Balakrishnan, and R. Reddy. “Characterization of Protein Secondary Structure Using Latent Semantic Analysis”. *IEEE Signal Processing magazine* 15 (2004), pp. 78–87.

[Gan+12] M. Ganapathiraju, A. Mitchell, M. Thahir, and K. Motwani. "Suite of tools for statistical N-gram language modeling for pattern mining in whole genome sequences". *Journal of Bioinformatics and Computational Biology* 10(6):1250016 (2012).

[HDY99] J. Han, G. Dong, and Y. Yin. "Efficient mining partial periodic patterns in time series database". *Proc. ofIEEE Int'l. Conf on Data Engineering* (1999), pp. 106–115.

[HGN00] J. Hipp, U. Güntzer, and G. Nakhaeizadeh. *Algorithms for Association Rule Mining-a General Survey and Comparison.* 2000.

[JM01] A. Jean-Marc. *Data Mining for Association Rules and Sequential Patterns.* Springer, 2001.

[KPC03] V. Kešelj, F. Peng, and N. Cercone. "N-gram-based author profiles for authorship attribution" in *Proceedings of the Conference Pacific Association for Computational Linguistics.* 2003.

[Koo] E. V. Koonin. *The Clusters of Orthologous Groups (COGs) Database: Phylogenetic Classification of Proteins from Complete Genomes.* `http://www.ncbi.nlm.nih.gov/books/NBK21090/`.

[Koo05] V. E. Koonin. "Central dogma of molecular biology". *Annual Review of Genetics* 39 (2005), pp. 309–338.

[Man+95] R. Mantegna, S. Buldyrev, A. Goldberger, S. Havlin, C. Peng, M. Simons, and H. Stanley. "Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics". *Phys. Rev.* 52 (1995), 2939–2950.

[Mar] U. Marovac. "Disjunction of Boolean equations". Publications de l'Institut Mathematique, Beograd, accepted for publication.

[Mar15] U. Marovac. "Systems of k Boolean inequations". *Journal of Multiple-Valued Logic and Soft Computing* 25 (5) (2015), D401–MVLSC.

[MM] U. Marovac and N. Mitić. "N-gram analysis of COG categorized protein sequences". MATCH: Communications in Mathematical and in Computer Chemistry, accepted for publication.

[MPLB08] N. Mitić, G. Pavlović-Lazetić, and M. Beljanski. "Could n-gram analysis contribute to genomic island determination". *Journal of biomedical informatics* 41(6) (2008), pp. 936–943.

[Mur+95] A. G. Murzin, S. Brenner, T. Hubbard, and C. Chothia. "SCOP: A structural classification of proteins database for the investigation of sequences and structures". *Journal of Molecular Biology* 247 (4) (1995), 536–540.

[Mut10] M. Muthukumarasamy. "Extraction and Prediction of System Properties Using Variable-N-Gram Modeling and Compressive Hashing". PhD thesis. University of Kentucky Doctoral Dissertations, 2010.

[OG11] H. Osmanbeyoglu and M. Ganapathiraju. "N-gram analysis of 970 microbial organisms reveals presence of biological language models". *Bioinformatics* 12(12) (2011).

[ORS98] B. Ozden, S. Ramaswamy, and A. Silberschatz. "Cyclic association rules". *Proc. 14th Int'l. Conference on Data Engineering (ICDE)* (1998), pp. 412–421.

[PLMB09] G. M. Pavlović-Lazetić, N. S. Mitić, and M. V. Beljanski. "n-Gram characterization of genomic islands in bacterial genomes". *Computer methods and programs in biomedicine* 93(3) (2009), pp. 241–256.

[PTH01] J. Pei, A. Tung, and J. Han. "Fault-tolerant frequent pattern mining: problems and challenges". *Proc. of ACM SIGMOD Int'l Workshop on Data Mining and KnowledgeDiscovery (DMKD)* (2001), pp. 194–203.

[Pei+01] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, and M. Dayal U. andHsu. "Prefixspan: mining sequential patterns by prefix-projected growth". *IEEE Int'l Conference on Data Engineernlg (ICDE)* (2001), pp. 215–224.

[PS12] T. Pirinen and M. Silfverberg. "Improving Finite-State Spell-Checker Suggestions with Part-of-Speech N-grams" in *Proceedings of International Conference on Intelligent Text Processing and Computational Linguistics CICLING*. 2012.

[Pod+07] A. Poddar, N. Chandra, M. Ganapathiraju, K. Sekar, . Klein-Seetharaman J, R. Reddy, and N. Balakrishnan. "Evolutionary insights from suffix array-based genome sequence analysis". *J Biosci.* 32(5) (2007), pp. 871–881.

[RBB07] T. Rani, S. Bhavani, and R. Bapi. "Analysis of E.coli promoter recognition problem in dinucleotide feature space". *Bioinformatics* 23(5) (2007), pp. 582–588.

[Rud74] S. Rudeanu. *Boolean functions and equations*. North-Holland, 1974.

[Rud01] S. Rudeanu. *Lattice functions and equations*. Springer, 2001.

[Sch91] J. C. Schmitt. "Trigram-based method of language identification". *US Patent* 5 (1991), pp. 62–143.

[TSV06] P. N. Tan, M. Steinbach, and K. V. *Introduction to Data Mining*. Pearson Education, 2006.

[Tat+00] R. L. Tatusov, M. Y. Galperin, D. Natale, and E. Koonin. "The COG database: a tool for genome-scale analysis of protein functions and evolution." *Nucleic Acids Res.* 28(1) (2000), pp. 33–36.

[TKL97] R. Tatusov, E. Koonin, and D. J. Lipman. "A genomic perspective on protein families". *Science* 278 (1997), 631–637.

[Tat+03] R. Tatusov et al. "The COG database: an updated version includes eukaryotes". *BMC Bioinformatics* 4(41) (2003).

[TJK06] A. Tomovic, P. Janicic, and V. Keselj. "N-gram-based classification and unsupervised hierarchical clustering of genome sequences". *Computer methods and programs in biomedicine* 81(2) (2006), pp. 137–153.

[Tro+02] O. Troyanskaya, O. Arbell, Y. Koren, G. Landau, and A. Bolshoy. "Sequence complexity profiles of prokaryotic genomic sequences: a fast algorithm for calculating linguistic complexity". *Bioinformatics* 18(5) (2002), pp. 679–688.

[Wan+05] J. T. L. Wang, M. J. Zaki, H. Toivonen, and D. Shasha. *Data Mining in Bioinformatics*. Springer, 2005.

[WJ05] W. Wei and Y. Jiong. "Mining Sequential Patterns from Large Data Sets" in *The Kluwer International Series on Advences in Database Systems*. Ed. by E. Ahmed. Vol. 28. Springer, 2005.

[Wis87] J. Wisniewski. "Effective text compression with simultaneous digram and trigram encoding". *Journal of Information Science* 13(3) (1987), 159–164.

[Wu+92] C. Wu, G. Whitson, J. Mclarty, A. Ermongkonchai, and T. Chang. "Protein classification artificial neural system". *Protein Science I* (1992), pp. 667–677.

[XJR03] Y. Xifeng, H. Jiawei, and A. Ramin. "CloSpan: Mining Closed Sequential Patterns in Large Datasets" in *Proc. of In SDM*. 2003, pp. 166–177.

[YFB01] C. Yang, U. Fayyad, and P. Bradley. “Efficient discovery of error-tolerant frequent itemsets in high dimensions”. *Proc. of ACM Int'f Cotferet ceo n Knowledge Dismver and Data Mining (KDD)* (2001), pp. 194–203.

[YWY01] J. Yang, W. Wang, and P. Yu. “InfoMiner: Mining Surprising Periodic Patterns” in *Proc. of the Seventh ACM Internatinal Conference on Knowledge Discover and Data Mining.* 2001, pp. 395–400.

[YWY02] J. Yang, W. Wang, and P. Yu. “InfoMiner+: Mining Partial Periodic Patterns with Gap Penalties” in *Proc. Second IEEE Internatinal Conference on Data Mining.* 2002.

[YWY03] J. Yang, W. Wang, and P. Yu. “STAMP: on discovery of statistically important pattern repeats in long sequential data” in *Proc. of the Third SIAM Internatinal Conference on Data Minig (SDM).* 2003.

[Zak01] M. Zaki. “SPADE-An Efficient Algorithn for Mining Frequent Sequences”. *Machine Laerning* 42 (2001), pp. 31–60.

[ZPZ81] E. Zamora, J. Pollock, and A. Zamora. “The use of Trigram Analysis for Spelling Error Detection” in *Proceedings of Information Processing and Management.* 1981, pp. 305–316.

Биографија

Основни подаци. Улфета Маровац рођена је 23. фебруара 1980. године у Чачку. Основну школу и гимназију завршила је у Новом Пазару. Школске 1998/1999. године уписала је студије на Математичком факултету у Београду (смер Рачунарство и информатика), и дипломирала је школске 2003/2004. године са просечном оценом 9,24. Школске 2007/2008. уписала је докторске студије на Математичком факултету, смер Рачунарство и информатика. Од септембара 2004. до септембра 2007. године радила је као наставник математике у Гимназији у Новом Пазару. Од септембра 2007. године запослена је на Државном универзитету у Новом Пазару као сарадник у настави. Године 2009. изабрана је у звање асистента. До сада је држала вежбе из следећих предмета: Основи информатике, Програмирање I, Програмирање II, Објектно програмирање, и Информатика. Основне области интересовања су јој истраживање података и развој и примена техника откривања знања из биолошких база података. У претходном периоду је учествовала на пројектима технолошког развоја ТР-13012 и ТР-11002. Тренутно је истраживач на пројекту Нове информационе технологије за аналитичко одлучивање базиране на организацији експеримента и њихова примена у биолошким, економским, и социолошким системима, бр.ИИИ-44007, који финансира Министарство просвете, науке и технолошког развоја Републике Србије.

Прилог 1.

Изјава о ауторству

Потписани-а Улфета Маровац

број уписа 2015/2007

Изјављујем

да је докторска дисертација под насловом

"Истраживање образаца у одређивању карактеристика протеина"

- резултат сопственог истраживачког рада,
- да предложена дисертација у целини ни у деловима није била предложена за добијање било које дипломе према студијским програмима других високошколских установа,
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио интелектуалну својину других лица.

Потпис докторанда

У Београду, 29.5.2015.

Улфета Маровац

Прилог 2.

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора: Улфета Маровац

Број уписа: 2015/2007

Студијски програм: Рачунарство и информатика

Наслов рада: "Истраживање образаца у одређивању карактеристика протеина"

Ментор: др Ненад Митић, ванредни професор

Потписани: Улфета Маровац

изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла за објављивање на порталу **Дигиталног репозиторијума Универзитета у Београду.**

Дозвољавам да се објаве моји лични подаци везани за добијање академског звања доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис докторанда

У Београду, 29.5.2015.

Улфета Маровац

Прилог 3.

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

“Истраживање образаца у одређивању карактеристика протеина”

која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигитални репозиторијум Универзитета у Београду могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

1. Ауторство

2. Ауторство - некомерцијално

3. Ауторство – некомерцијално – без прераде

4. Ауторство – некомерцијално – делити под истим условима

5. Ауторство – без прераде

6. Ауторство – делити под истим условима

(Молимо да заокружите само једну од шест понуђених лиценци, кратак опис лиценци дат је на полеђини листа).

Потпис докторанда

У Београду, 29.5.2015.

Ултрета Маровац

1. Ауторство - Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце, чак и у комерцијалне сврхе. Ово је најслободнија од свих лиценци.

2. Ауторство – некомерцијално. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела.

3. Ауторство - некомерцијално – без прераде. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела. У односу на све остале лиценце, овом лиценцом се ограничава највећи обим права коришћења дела.

4. Ауторство - некомерцијално – делити под истим условима. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца не дозвољава комерцијалну употребу дела и прерада.

5. Ауторство – без прераде. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца дозвољава комерцијалну употребу дела.

6. Ауторство - делити под истим условима. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца дозвољава комерцијалну употребу дела и прерада. Слична је софтверским лиценцама, односно лиценцама отвореног кода.