УНИВЕРЗИТЕТ У БЕОГРАДУ
МАТЕМАТИЧКИ ФАКУЛТЕТ

Јасмина Т. Јовановић

РАЗВОЈ МЕТОДА ЗА АНАЛИЗУ СЛИЧНОСТИ БИОЛОШКИХ СЕКВЕНЦИ НА ОСНОВУ КАРАКТЕРИСТИКА ПОНОВАКА

докторска дисертација

Београд, 2022

UNIVERSITY OF BELGRADE
FACULTY OF MATHEMATICS

Jasmina T. Jovanović

DESIGN AND IMPLEMENTATION OF METHODS FOR BIOLOGICAL SEQUENCE SIMILARITY ANALYSIS BASED ON REPEAT CHARACTERISTICS

Doctoral Dissertation

Belgrade, 2022

Ментор:

Проф. др Ненад Митић, редовни професор
Универзитет у Београду, Математички факултет

Чланови комисије:

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

др Јована Ковачевић, доцент
Универзитет у Београду, Математички факултет

др Зоран Огњановић, научни саветник
Српска академија наука и уметности, Математички институт

Датум одбране: __.__.2022.

Изјава захвалности:

Овај рад је резултат вишегодишње анализе, учења, конципирања и тестирања метода у циљу решавања проблема идентификовања сличности биолошких секвенци на основу различитих карактеристика поновака. Желела бих пре свега да се захвалим свом ментору др Ненаду Митићу, редовном професору Математичког факултета у Београду, на инспиративној идеји и указаном поверењу, великој помоћи, дугогодишњем усмеравању, стрпљењу и стручним саветима током мојих докторских студија.

Захваљујем се др Милошу Бељанском, на корисним саветима и подељеним идејама у смеру реализације делова ове докторске дисертације.

Хвала свим члановима комисије др Гордани Павловић Лажетић, др Јовани Ковачевић и др Зорану Огњановићу на конструктивним саветима приликом израде тезе и доприноса квалитету исте.

Захваљујем се драгим пријатељима на разумевању и подршци у току израде докторске дисертације.

На крају желим посебно да се захвалим мом супругу Ивану на мотивацији, подршци и разноврсним корисним саветима приликом припреме и израде ове тезе. Такође, велико хвала мојој породици - супругу Ивану, сестри Јели, родитељима Цици и Тодору на стрпљењу, помоћи и разумевању током студија. Хвала и мом малом Лази који је нестрпљиво чекао да се играмо након што мама заврши само још један пасус.

Посвећујем овај рад свима који искрено верују и вреднују да су посвећеност и упорност сигурни кораци који воде ка остварењу циљева.

Наслов дисертације: Развој метода за анализу сличности биолошких секвенци на основу карактеристика поновака

Сажетак: Анализа сличности биолошких секвенци омогућава утврђивање функционалних, структурних и еволуционих односа између различитих организама. Међутим, сличност биолошких секвенци и утвђивање особина нових нуклеотидних и протеинских секвенци су рачунарски захтевне методе у биоинформатици што намеће потребу за даљим развојем метода и алгоритама за њихово поређење.

У складу са брзим растом и доступношћу велике количине биолошких података, нови алгоритми се развијају са циљем што ефикасније и прецизније обраде ових података. Један од изазова код одређивања сличности биолошких секвенци јесте издвајање скупа значајних атрибута секвенци, чија кардиналност може да буде велика за примену у постојећим методама за одређивање сличности елемената. Стога је од изузетног значаја имати једноставан и ефикасан алгоритам за одређивање међусобних односа биолошких секвенци.

Циљ овог рада је формирање и имплементација нових метода за анализу сличности секвенци на основу статистички значајних поновака различитих дужина и типова. Прва метода се заснива на теорији информација узимајући у обзир позицију и учесталост статистички значајних поновака, за које се не очекује такво присуство у случајно генерисаној секвенци исте дужине. Друга метода садржи формирање потписа секвенци и профила таксономских категорија на основу парова понављајућих делова секвенци, као и растојања између елемената тих парова. Идеја ове методе је представити секвенце мањим бројем карактеристичних тачака у циљу препознавања истих као код алгоритама за препознавање лица.

Предложене методе су тестиране на различитим референтним скуповима биолошких секвенци и резултати су упоређени са резултатима добро познатих и ефикасних алгоритама који се заснивају на поравнању (BLAST, Clustal Omega) и алгоритама без поравнања који се заснивају на к-торкама. Добијени резултати показују висок ниво конзистентности са резултатима метода са којима је извршено поређење. Прецизност предложених метода није била мања од вредности добијених за постојеће методе са којима су резултати упоређивани за већи број спроведених тестирања, док је брзина добијања резултата зависила од рачунарске инфраструктуре и примера секвенци. Предложене методе представљају значајну допуну постојећим методама за одређивање сличности биолошких секвенци, јер се досадашње методе за анализу сличности биолошких секвенци нису заснивале на статистички значајним поновцима различитих карактеристика.

Кључне речи: Анализа сличности секвенци; Методе за анализу сличности секвенци без поравнања; Статистички значајни поновци; Ентропија заснована на локалној учесталости; Хијерархијско кластеровање; Вишедимензиони векторски простор; Потписи секвенци; Класификација

Научна област: Рачунарство и информатика
Ужа научна област: Биоинформатика

Dissertation title: Design and Implementation of Methods for Biological Sequence Similarity Analysis Based on Repeat Characteristics

Abstract: The analysis of biological sequence similarity between different species is significant in identifying functional, structural or evolutionary relationships among the species. Biological sequence similarity and analysis of newly discovered nucleotide and amino acid sequences are demanding tasks in bioinformatics.

As biological data is growing exponentially, new and innovative algorithms are needed to be constantly developed to get faster and more effective data processing. The challenge in sequence similarity analysis algorithms is that sequence does not always have obvious features and the dimension of sequence features may be very high for applying regular feature selection methods on sequences. It is important to have a simple and effective algorithm for determining biological sequence relationships.

This thesis proposes two new methods for sequence transformation in feature vectors that takes into consideration statistically significant repetitive parts of analyzed sequences, as well as includes different approaches for determination of nucleotide sequence similarity and sequence classification for predicting taxonomy groups of biological sequence data. The first method is based on information theory and fact that both position and frequency of repeated sequences are not expected to occur with the identical presence in a random sequence of the same length. The second method includes building signatures of biological sequences and profiles of taxonomic classes based on repetitive parts of sequences and distances between these repeats.

Proposed methods have been validated on multiple data sets and compared with results obtained using different well known and accepted methods in this field like BLAST, Clustal Omega and methods based on k-mers. Resulted precision for proposed methods is close to values provided for existing methods for the majority of tested data-sets, and time performance depends strictly to used infrastructure and sequence type. Methods provide results that are comparable with other commonly used methods focused on resolving the same problem, taking into consideration statistically significant repetitive parts of sequences with different characteristics.

Keywords: Sequence similarity analysis; Alignment-free method; Statistically significant repeat; Local frequency based entropy; Hierarchical clustering; Multi-dimensional vector space; Sequence signature; Classification

Research area: Computer Science
Research sub-area: Bioinformatics

Садржај

1. Увод

1.1 Позадина и мотивација

Биоинформатика је интердисциплинарна област која се бави развојем и коришћењем рачунарских и математичких метода и алата за прикупљање, чување, анализу, обраду, разумевање и визуелизацију биолошких информација [1] [2] [3]. Количина ових информација је изузетно велика, односи између њих су веома сложени, стога и њихова обрада у самом истраживању није једноставна. Издвајање информација из велике количине података и коришћење истих у биолошким истраживањима је важан део којим се бави биоинформатика [4].

ДНК, РНК и протеини су три основна типа макромолекула, који су есенцијални за све познате форме живота. Биолошки макромолекули полимерне природе (ДНК, РНК, протеини) се могу посматрати као ниске карактера, ознака нуклеотида и аминокиселина, јединица које се понављају. Нуклеотидна секвенца молекула ДНК се може представити као ниска карактера над азбуком A = {A, C, G, T}, док се аминокиселинска ниска може посматрати као ниска карактера над 20-ословном азбуком сачињеном од ознака амино киселина A = {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y, U, O}. Дужина нуклеотидних секвенци варира од неколико до више стотина милиона нуклеотида, док дужина аминокиселинских секвенци варира од неколико десетина до више десетина хиљада аминокиселина.

Понављајуће секвенце (репетитивне ниске) представљају делове секвенци (подсеквенце) који се јављају два или више пута у једној секвенци. У зависности од тога да ли је подсеквенца копије идентична оригиналној, или заједно са оригиналном чини палиндром, понављајуће секвенце могу бити директне или обрнуте [5] [6]. Такође понављајуће секвенце могу бити даље подељене у комплементарне и некомплементарне у зависности од тога да ли испуњавају функцију пресликавања свих карактера у њихове комплементарне карактере. Статистички значајни поновци представљају подскуп поновака секвенце за које није очекивано да ће се појавити у насумичној секвенци исте дужине [5] [6]. У скорашњим истраживањима, понављајуће ДНК секвенце су издвојене као битне у биолошким механизмима [7,5].

Анализа сличности нуклеотидних и протеинских секвенци је важна у одређивању функционалних, структурних и еволуционих односа између различитих таксономских категорија и/или других карактеристика организама. Ако су две секвенце сличне, највероватније имају заједничког претка [4]. Њена значајност се такође огледа у одређивању категорија новооткривених секвенци поређењем са секвенцама које имају познате функције. Ако се идентификује секвенца познате функције која је слична неидентификованој секвенци, са великом вероватноћом можемо да предвидимо функцију неидентификоване секвенце [8]. Сличност секвенци представља основу у истраживању података овог типа (биолошке секвенце) [4] и ова врста анализе је најчешће коришћена и примењивана у биоинформатици.

Постоје различити алгоритми за поређење секвенци, који се могу поделити на алгоритме засноване на поравнању (глобална и локална поравнања) или без поравнања секвенци. Методе засноване на поравнању као што су BLAST [9] и CLUSTAL [10] су веома прецизне и најчешће коришћене. Више алгоритама који користе технике динамичког програмирања, као и хеуристичке алгоритме су успешно развијени у циљу решавања проблема поравнања секвенци и анализе сличности секвенци. Ови алгоритми се поред биолошких секвенци могу применити и на друге секвенце (на пример у лингвистици за анализу природних језика).

Најпознатије технике истраживања података као што су класификација и кластеровање се успешно примењују и на решавање проблема анализе сличности секвенци [4].

Методе за поређење секвенци које су развијене у претходном периоду нису засноване на скуповима статистичких значајних поновака варијабилних дужина и различитих типова. Оваква врста одабира карактеристика које осликавају секвенцу дала би анализи сличности секвенци значајну предност у односу на постојеће методе које захтевају арбитрарни одабир дужине к-торки чиме се пропуштају кључне информације о поновцима другачијих дужина. С обзиром на доступност велике количине података (биолошких секвенци), јавља се мотивација за унапређењем метода за анализу сличности секвенци у циљу добијања што прецизнијег резултата. На основу поменутог, аутоматизација идентификације поновака варијабилних дужина удружена са постојећим и новим методама за поређење секвенци без поравнања би представљала напредно решење за ову врсту поређења, у складу са мањим заузећем простора неопходним за чување атрибута секвенци.

1.2 Предмет истраживања и циљ дисертације

Предмет ове дисертације биће анализа биолошких (нуклеотидних и аминокиселинских) секвенци и њихових статистички значајних поновака различитих типова и дужина у циљу развоја нових модела за одређивање сличности секвенци на основу идентификованих поновака. До сада развијени и коришћени алгоритми за поређење секвенци без поравнања, нису засновани на различитим типовима статистички значајних поновака варијабилних дужина. У формирању модела за одређивање сличности секвенци биће коришћене методе истраживања података (класификација, кластеровање и анализа текста), на основу којих ће се омогућити провера прецизности добијених резултата.

Циљ ове дисертације је развијање модела за одређивање сличности/различитости биолошких секвенци и класификацију истих. Постављени циљеви укључују:

- Преглед и разумевање постојећих метода и релевантних техника за одређивање сличности секвенци.
- Формирање базе података са особинама секвенци и њиховим поновцима у циљу одређивања и прикупљања карактеристичних тачака појединих секвенци и њихових таксономских класа, као и других особина сачуваних у облику мета података.
- Развој и имплементацију нових модела за карактеризацију и класификовање целих секвенци или делова истих, који се могу користити у биоинформатици за даља истраживања, користећи методе истраживања података. Овај развој обухвата како нумеричко, тако и визуелно представљање резултата.
- Процену квалитета предложених модела на биолошким скуповима података, као и поређење резултата са јавно доступним и коришћеним методама за одређивање сличности секвенци.

1.3 Организација тезе

Рад садржи пет поглавља и додатак. Прво поглавље садржи увод и описује циљеве дисертације. Друго поглавље садржи опис основних појмова, као и постојећих метода за анализу сличности секвенци. У трећем поглављу, које представља централни део рада, су описане новопредложене методе за анализу сличности секвенци. Изложени су модели засновани на статистички значајним поновцима и трансформацији биолошких секвенци у векторски облик погодан за даљу анализу. У четвртом поглављу су приказани резултати примена предложених метода на тестне скупове података и процена квалитета предложених модела, као и поређење развијених метода са резултатима примена познатих метода који се баве истом проблематиком. Пето поглавље садржи закључак са предлозима за даљи рад у овој области. Додатак садржи резултате за различите типове поновака који нису приказани у четвртом поглављу, као и табеле са више информација о тестним секвенцама и резултатима.

2. Основе и сродни приступи истраживања сличности биолошких секвенци

2.1 Биолошке секвенце и базе биолошких секвенци

Биолошка секвенца је непрекидан низ нуклеотида или аминокиселина, јединица које се понављају. У складу са изузетно напредним развојем метода за секвенцирање генома у претходном периоду, величине база података које садрже информације о биолошким секвенцама и њиховим особинама се повећавају експоненцијалном стопом раста [1]. Часопис *Nucleic Acids Research* годишње објављује специјално издање са најновијим пресеком актуелних база података које садрже биолошке податке. У издању за 2022. годину, објављено је 1645 актуелних база података [11], од којих 159 садржи податке о ДНК секвенцама, 107 база података које садржи информације о РНК секвенцама и 219 база података са информацијама о протеинским секвенцама [12].

Базе података које садрже информације о нуклеотидним и аминокиселинским секвенцама се називају примарним базама података. NCBI [13] је основан као национални извор микробиолошких информација [1] и садржи јавно доступне базе и алате за анализу биолошких података. GenBank је примарна база података анотираних јавно доступних биолошких секвенци. Она садржи 2.1 милијарду нуклеотидних секвенци [14].

ViPR (*Virus Pathogen Resourse*) је користан извор података коју користи заједница истраживача у области вирусологије и дизајниран је да помогне корисницима у различитим пројектима који укључују анализу секвенци и структуру вируса [15,16]. ICTV база података *(International Committee on Taxonomy of Viruses (ICTV) database - Virus Metadata repository (VMR))* [17] садржи листе примера вируса за сваку врсту као и њихове таксономије.

Често се као референтни скупови података за тестирање метода за анализу сличности биолошких секвенци користе скупови секвенци митохондријалне ДНК сисара (*mtDNK*) због њихове дужине, изолати секвенци вируса због тестирања модела над секвенцама које су изузетно сличне, као и одређени референтни скупови протеинских секвенци. Митохондријална ДНК (*mtDNK*) чини геном митохондрија, и садржи гене неопходне за функционисање свих органела кључних за енергетске процесе у ћелијама животиња. Скуп протеинских секвенци представља скуп секвенци истог протеина пореклом од различитих организама или скуп различитих протеина који деле неку заједничку особину.

2.2 Поновци

У претходном периоду се сматрало да је функционални део ДНК само онај који кодира протеине (неких 1-2% укупне секвенце ДНК човека), међутим новије студије су показале да некодирајуће секвенце имају такође важну улогу. ДНК није хомогена ниска карактера, већ комбинација (мозаик) мотива, који у синергији служе за координацију и регулацију синтезе протеина [1]. Поновљајуће секвенце представљају делове биолошких секвенци који се више пута понављају у једној секвенци. Претпоставка је да се важне еволуционе информације могу издвојити из понављајућих делова секвенци, као и њихових односа. Понављајуће секвенце се могу разликовати по дужини и броју понављања. Постоје разни типови понављајућих секвенци, као што су сателитски поновци, интермедијарни низови, тандемски поновци, итд. Тандемски поновци су понављајуће секвенце које се јављају узастопно у секвенци. Сателитски поновци су секвенце које се јављају у великом броју копија.

Статистички значајни поновци представљају подскуп понављајућих делова анализиране секвенце за које није очекивано појављивање у датом броју у случајној изабраној секвенци исте дужине [5] [6]. Постоје различити типови понављајућих секвенци у

случају нуклеотидне секвенце (Слика 1): директни комплементарни (у даљем тексту означени са DC), директни некомплементарни (у даљем тексту означени са DN), обрнути комплементарни (у даљем тексту означени са IC) и обрнути некомплементарни (у даљем тексту означени са IN) [5] [6]. У случају протеинске секвенце дефинисана су два типа поновака: директни некомплементарни (DN), и обрнути некомплементарни (IN), с обзиром да систем комплементарности важи за нуклеотидне секвенце, али није примењив на низ аминокиселина. Понављајућа секвенца је максимална ако не постоји секвенца, којој је понављајућа секвенца подсеквенца, а иста задовољава дефиницију понављајуће секвенце.

Парови подниски *x* и *y* ниске *z* који се налазе на позицијама p_x и p_y ниске *z* су:

1. директни некомплементарни поновци (DN) ако и само ако је $x = y$ и $p_x < p_y$;
2. обрнути некомплементарни поновци (IN) ако и само ако је $x = w$ и $p_x \leq p_y$ где је *w* ниска са обрнутим редоследом карактера у односу на *y* (на пример обрнута ниска ниске agt је tga) ;
3. директни комплементарни поновци (DC) ако и само ако је $x = f(y)$ и $p_x < p_y$, где је *f* функција пресликавања ниске у комплементарну ниску (на пример комплементарна ниска од agt је tca);
4. обрнути комплементарни поновци (IC) ако и само ако је $x = f(w)$ и $p_x \leq p_y$ где је *w* ниска са обрнутим редоследом карактера у односу на *y* и *f* функција пресликавања ниске у комплементарну ниску.

Слика 1: Примери различитих типова поновака на позицијама 4 и 11 код нуклеотидних секвенци (слика лево), и позицијама 4 и 15 код протеинских секвенци (слика десно)

Понављајуће секвенце могу бити различитих дужина. Максимална статистички значајна понављајућа секвенца ће у даљем тексту бити означена као поновак.

Растојање између пара поновака се израчунава као разлика позиција првих елемената парова поновака истог типа у анализираној секвенци ($d = p_y - p_x$). Позиција појављивања поновка у секвенци, поредак поновака и број појављивања истих се издвајају као додатне особине поновака које су важне у предложеним методама.

2.3 Постојеће методе за одређивање сличности секвенци

Ако степен сличности секвенци прелази 30%, сматра се да оне могу да имају хомологну везу. Хомологне секвенце имају заједничког еволуцијског претка зато и њихове структуре и функције могу да буду сличне [4].

Из перспективе програмирања, анализа биолошких секвенци се не разликује много од анализа поређења ниски и текста [16]. Постоје различити алгоритми за поређење биолошких секвенци, који се могу поделити на алгоритме засноване на поравнању (глобална и локална поравнања) или без поравнања секвенци. Код метода заснованих на поравнању, две секвенце се упоређују како би се одредиле разлике на свакој позицији, а које могу бити уметање, брисање и друге врсте неусклађености изазване еволуцијом секвенце. Процес поравнања укључује оцењивање поравнања на свакој позицији у зависности од преклапања и типа

разлике од чега зависи крајњи резултат. Најпознатији алгоритми поравнања секвенци су Needleman–Wunsch [18] и Smith–Waterman [19]. У циљу повећања брзине израчунавања и поређења великог броја секванци, развијени су нови алати као што су BLAST [9] и Clustal Omega [20] [21] [22]. Предност метода са поравнањем јесте њихова прецизност.

Поравнање две дуге ДНК секвенце дужине више милиона нуклеотида неизводљиво у пракси [23]. Стога су као алтернатива поравнања секвенци развијене такозване методе без поравнања секвенци [24], и њихова скалабилност омогућава да се примене на много веће скупове података од конвенционалних метода за поређење секвенци [23].

Методе без поравнања секвенци се могу окарактерисати као методе за одређивање сличности секвенци које не укључују поравнање секвенци у било ком кораку алгоритма [24]. Приступи методама без поравнања су математички добро утемељени у пољима линеарне алгебре, статистике, као и теорије информација и засновани су на дефинисању мере различитости између секвенци [24]. Методе без поравнања секвенци се најчешће деле у две категорије: методе засноване на броју појављивања подсеквенци одређене дужине (методе засноване на речима дужине к - к-торкама) и методе засноване на теорији информација [25]. Такође постоје и други приступи на којима су засноване методе без поравнања као што су: дужина заједничких подниски, најкраће заједничке подниске, микро поравнања секвенци, представљање секвенци на основу теорије хаоса, позиције нуклеотида, Фуријеове трансформације, и системи са поновљеним функцијама [23] [26].

Тренутно најзаступљеније методе без поравнања су методе засноване на к-торкама [27]. На основу претпоставке да сличне секвенце деле исте или сличне речи одређених дужина, већина ових метода се заснива на трансформисању секвенце у простор атрибута дефинисан фреквенцијама речи, представљајући исте нумеричким векторима који могу да се користе за рачунање растојања [23]. Начин рада метода без поравнања које су засноване на к-торкама се може поделити у три корака [24]. (1) Секвенце које се пореде се деле у колекције јединствених речи одговарајуће дужине. (2) Други корак је трансформисање секвенци у нумерички вектор пребројавањем идентификованих речи. (3) Трећи корак је израчунавање сличности секвенци на основу растојања (најчешће се користи Еуклидско растојање) између добијених нумеричких вектора. Методе засноване на речима се базирају на описаним корацима са одређеним методолошким варијацијама у њима. Први корак може да се модификује тако што ће се смањити број симбола у алфабету одређеним пресликавањем. Други корак може да се модификује на доста начина користећи различите функције за формирање нумеричких вектора уместо рачунања броја појављивања речи у секвенци. Трећи корак може да се модификује користећи различите мере сличности. На овај начин се добија велики број различитих метода без поравнања заснованих на к-торкама.

Најчешће коришћена мера заснована на информацијама, која квантификује очекивану вредност количине информације садржане у анализираном тексту, се назива Шенонова ентропија [28]. Сличан приступ (тачније мера заснована на информацијама) може да се примени у биоинформатици на биолошке секвенце и њихове подсеквенце посматрајући их као речи у анализираном тексту. Поређење секвенци се у овом случају заснива на примени ентропије користећи учесталост подсеквенци [24].

Неколико метода за поређење нуклеотидних секвенци без поравнања засноване на теорији информација су наведене у наставку. Две методе за поређење секвенци без поравнања засноване на теорији информација су mBKM [29] и CPF [30], и оне су дефинисане користећи ентропију засновану на локалној учесталости. У овим методама се примењује Шенонова ентропија на к-торке за одређену дужину речи *к* узимајући у обзир позицију, број појављивања и поредак речи у секвенци. У DMk методи без поравнања се такође користи ентропија за формирање вектора атрибута на основу позиције сваког к-мера обезбеђујући на тај начин више информација него класичне методе засноване на речима [31]. Li и Wang су у раду [32] дефинисали нормализовану релативну ентропију за формирање 12-димензионог вектора атрибута за презентовање секвенци који узима у обзир елементе секвенци и поредак између истих. Мера сличности која узима у обзир структуру

преклапајућих речи и учесталост истих је дефинисана у раду [33]. WSE мера растојања је надоградња релативне ентропије у случају велике вредности параметра *к* [34]. Такође, велики број метода за поређење секвенци без поравнања је усмерен на примену у скуповима протеинских секвнци [35] [36] [37] [38] [39] [40] [41] [42] [43].

Неке од предности метода без поравнања у односу на методе са поравнањем су следеће: брзина извршавања алгоритама и добијања резултата је већа, примењиве су за анализу читавих генома и секвенци великих дужина, не зависе од улазних параметара у великој мери као и методе са поравнањем [24]. Методе без поравнања захтевају скромније рачунарске ресурсе и користе алгоритме који су најчешће линеарне сложености [44]. У складу са тим се повећава и број метода за анализу сличности секвенци без поравнања [24] [26] [44] [45].

2.4 Анализа сличности секвенци помоћу метода истраживања података

Истраживање података је процес издвајања значајних, претходно неоткривених, потенцијално корисних и разумљивих образаца из података [46]. Истраживање података у домену биолошких секвенци се у општем случају не разликује од примене истраживања података на сличност секвенци [47]. Основни модели истраживања података укључују класификацију, кластеровање, откривање аномалија и правила придруживања [48].

2.4.1. Класификација секвенци

Главни задатак класификације јесте придруживање одређене класе (или класа) неком анализираном објекту на основу међусобне сличности или различитости. Класификација секвенци представља важан изазов у истраживању података [49]. Применом класификације на биолошке секвенце, можемо веома брзо да придружимо таксономску групу новооткривеној секвенци или делу секвенце.

Улазне податке у методама класификације представљају слогови облика $\left(x_i^1, x_i^2, \ldots, x_i^k, y_i\right)$, који се састоје од скупа атрибута x_i, као и посебног атрибута y који се назива атрибут класе. Циљ је да се утврди да ли се на основу вредности атрибута x_i може одредити вредност атрибута y. Задатак класификације је пронаћи функцију пресликавања скупа атрибута x у једну од предефинисаних вредности y. Улазни материјал се у процесу класификације дели на два дисјунктна скупа: скуп података за тренирање и скуп података за тестирање. Скуп података за тренирање се користи за формирање модела и иницијално одређивање тачности модела. Скуп података за тестирање служи за проверу исправности класификационог модела. Класификација може бити бинарна (када су дефинисане две могуће класе као излазни атрибут) и вишекласна (када је дефинисано више од две могуће класе као излазни атрибут). Такође класификациони модел можемо да означимо као једнозначан и вишезначан. Једнозначан модел представља модел у коме једној инстанци може да буде додељена тачно једна класа, док вишезначан модел омогућава доделу више од једне класе. Класификација је чврста ако се доноси бинарна одлука да ли инстанца припада класи или не, и мека када се податку придружује нумеричка вредност која означава оцену мере припадности датој класи.

Методе за класификовање секвенци се деле у три категорије, при чему неке методе могу да припадају двема категоријама [50]:

- Методе засноване на атрибутима – методе које обухватају трансформацију секвенци у простор вектора атрибута, на које се примењује познати постојећи класификациони модели. Најлакши начин за трансформисање секвенци јесте посматрати сваки елемент секвенце као атрибут. Други начин је узимати делове секвенци од *к* узастопних симбола (к-торке). На основу скупа *к*-торки, секвенца се може приказати као вектор

присуства/одсуства ових κ-торки, или као вектор учесталости појављивања κ-торки. Такође је дефинисан и метод заснован на обрасцима.

- Методе засноване на растојању – дефинисањем функције растојања која мери сличност између секвенци. Ове методе дефинишу функцију сличности/растојања парова секвенци и заснивају се на поравнању секвенци које може бити глобално или локално. Као резултат се добија вредност поравнања, која се најчешће мери бројем симбола који се поклапају или не поклапају.
- Методе засноване на моделима, као што су сакривени Марковљеви модели и други статистички модели за класификовање. За дату класу, модел М одређује расподелу вероватноће да секвенца припада тој класи.

Постојећи изазови везани за класификацију секвенци су:

- Већина постојећих класификатора (као што су дрвета одлучивања или неуронске мреже), прихватају једино вектор атрибута као улазан податак, али у подацима који садрже секвенце, обично не постоје јасно дефинисани атрибути. Формирање ефикасног класификатора може да буде тешко, јер нема јасно дефинисаних и интуитивних атрибута.
- Иако се секвенца може трансформисати у листу атрибута различитим методама за избор атрибута, ове методе нису једноставне. Број атрибута може бити велики што даље узрокује скупо израчунавање;

2.4.2. Кластеровање

Главни задатак кластеровања јесте груписање датог скупа објеката $X = \{x_1, x_2, \dots, x_n\}$ такво да су објекти x_i у истој групи (кластеру) G_R међусобно сличнији по неком критеријуму, него неком објекту x_j који припада некој другој групи G_k. Свака од група се назива кластер и има своје карактеристике. Целокупан поступак поделе улазног материјала се назива кластеровање. Сврха кластеровања је подела података на основу заједничких карактеристика у исту категорију, ради даље анализе података. Ова метода се разликује од класификације, јер број категорија (група) није унапред познат.

Различити типови кластеровања, тачније начини груписања скупа објеката у скупове кластера су [48]:

- Хијерархијско/партиционо кластеровање – Код партиционог кластеровања скуп улазних података се дели у дисјунктне подскупове (кластере) такве да сваки податак припада тачно једном кластеру. Код хијерархијског кластеровања кластери могу да садрже подкластере и улазни податак може да припада већем броју кластера на различитим нивоима хијерархије. Угнеждени кластери су организовани у дрволиком облику. Један приступ хијерархијског кластеровања је техника раздвајања. У овом приступу се полази од једног кластера, који садржи све инстанце. Дати почетни кластер се дели у складу са мером сличности, тако да се као крајњи резултат добију кластери са по једном инстанцом. Други приступ је заснован на техници спајања (сакупљајуће хијерархијско кластеровање) где се иницијално свака инстанца налази у појединачном кластеру, који се затим спајају на основу претходно изабране мере сличности. Спајање појединачних кластера се врши све док се не добије тачно један свеобухватни кластер. Крајњи резултат хијерархијског кластеровања се визуелно може представити дендрограмом, графичким приказом кластера у облику стабла повезивања, што је погодно за филогенетску анализу;
- Ексклузивно/не-ексклузивно – подела на ексклузивно и не-ексклузивно кластеровање је у зависности од тога да ли појединачни елемент који се кластерује припада само

једном (ексклузивно) или може истовремено да се налази у више кластера (не-ексклузивно кластеровање). Не-ексклузивно кластеровање се често користи када се елемент налази између два или више кластера и може да буде додељен било ком од ових кластера;
- Комплетно/делимично кластеровање – Код комплетног кластеровања се сваки елемент придружује неком кластеру, док код делимичног то није случај. Делимично кластеровање се најчешће користи када постоје елементи који не припадају јасно дефинисаним групама (као што су шум и елементи ван граница)

Два основна изазова код кластеровања ДНК секвенци јесу: издвајање карактеристика секвенци и дизајнирање ефикасне мере сличности из перспективе биолошких односа [4]. До сада су развијени алгоритми који су засновани на сличностима секвенци, теорији графова, методе засноване на векторима атрибута, фреквенцијама компоненти и дигиталним сигналима [4].

2.4.3. Мере и матрице сличности

Нека је дато m објеката $x_i \in \mathbb{R}^n, i = 1, \dots, m$. Мера одстојања између објеката је нумеричка мера која представља колико су два објекта слична или различита.

Дефиниција 1: За меру *s(x, y)* кажемо да представља меру сличности објеката *x* и *y* ако задовољава следеће услове:

1. услов нормираности: $0 \leq s(x, y) \leq 1$, за све x и y;
2. услов идентичности: $s(x, y) = 1$, само ако су *x* и *y* једнаки;
3. услов симетричности: $s(x, y) = s(y, x)$.

Дефиниција 2: Функција растојања (различитости) *d(x, y)* представља метрику између два објекта *x* и *y*, ако задовољава следеће услове за све објекте:

1. услов позитивне одређености: $d(x, y) \geq 0$, за све x и y, док је $d(x, y) = 0$, акко $x = y$;
2. услов симетричности: $d(x, y) = d(y, x)$;
3. услов неједнакости троугла: $d(x, z) \leq d(x, y) + d(y, z)$, за све x, y и z.

Већина мера сличности су универзалног значаја и користе се у многим областима. У даљем тексту је описано неколико добро познатих и широко коришћених метрика. Косинусно растојање, Жакардово растојање и коефицијент Танимота су честе мере које се користе за израчунавање сличности асиметричних вектора [48].

Дефиниција 3: Сличност два објекта који су представљени векторима *x* и *y* у *n*-димензионом векторском простору можемо израчунати на основу величине угла које та два вектора заклапају

$$\text{Cosine Similarity}(x, y) = \frac{x \cdot y}{|x||y|}$$

Ова мера се назива косинусна сличност, и користи се у векторским просторима великих димензија [51].

Дефиниција 4: Сличност два објекта *x* и *y* са бинарним атрибутима се може дефинисати помоћу Жакардовог (Jaccard) коефицијента као мере сличности

$$Jaccard(x, y) = \frac{M_{11}}{M_{01}+M_{10}+M_{11}},$$

где је:
M_{01} – број атрибута који су једнаки 0 у *x*, и 1 у *y*
M_{10} – број атрибута који су једнаки 1 у *x*, и 0 у *y*
M_{00} – број атрибута који су једнаки 0 и у *x*, и у *y*
M_{11} – број атрибута који су једнаки 1 и у *x*, и у *y*

Ова мера се често користи за податке који су представљени бинарним атрибутима, додељујући 1 или нула атрибуту у зависности да ли атрибут присутан или не. Код ове мере се подразумева да присуство атрибута у оба вектора показује сличност, док одсуство атрибута нема значаја.

Дефиниција 5: Сличност два објекта који су представљени векторима *x* и *y* у *n*-димензионом векторском простору можемо израчунати помоћу проширеног Жакардовог коефицијента (коефицијент Танимото-а) на следећи начин:

$$EJ = \frac{x \cdot y}{\|x\|^2 + \|y\|^2 - x \cdot y}$$

Дефиниција 6: Мултиваријантни (вишедимензиони) подаци се могу представити матрицом података, где елементи сваке врсте представљају објекте ($\{x_1, x_2, \dots, x_m\}$), док елементи колона представљају њихове атрибуте.

Дефиниција 7: Матрица података је матрица димензија *m x n*, где је *m* број објеката, и *n* број атрибута:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix}$$

Дефиниција 8: Матрица сличности (или различитости) је матрица димензија *m x m*, добијена из матрице података применом функције сличности или растојања на елементе из метрице података:

$$D = \begin{bmatrix} d_{11} & \cdots & d_{1m} \\ \vdots & \ddots & \vdots \\ d_{m1} & \cdots & d_{mm} \end{bmatrix}$$

2.4.4. Оцене квалитета модела

Евалуација модела представља процену његове способности исправног предвиђања и заснована је на мерама квалитета модела и на техникама еваулације модела. Постоји више различитих начина да се квантификује квалитет модела класификације и кластеровања и ове мере зависе од врсте проблема који се решава.

Процена квалитета модела класификације

Оцена квалитета модела класификације се заснива на поређењу унапред познате класе са класом коју је предвидео класификатор.

Код бинарне класификације две могуће класе се најчешће називају позитивна и негативна, и у процесу класификације се предвиђа да ли је нека инстанца заправо позитивна или негативна. Исходи код бинарне класификације су:

- Стварно позитивни (TP) – број инстанци код којих је предвиђена класа позитивна и стварна класа је позитивна
- Стварно негативни (TN) - број инстанци код којих је предвиђена класа негативна и стварна класа је негативна
- Лажно позитивни (FP) - број инстанци код којих је предвиђена класа позитивна и стварна класа је негативна
- Лажно негативни (FN) – број инстанци код којих је предвиђена класа негативна и стварна класа је позитивна

Чест приказ оцене квалитета модела је преко матрица конфузије, код које врсте и колоне представљају класе (Табела 1). Ознаке у колони дате матрице представљају број инстанци које је класификатор доделио класи (предвиђене вредности), док су ознаке у врсти означене бројевима инстанци које заиста припадају тој класи.

Табела 1. Матрица конфузије

Стварно/Предвиђено	**Позитивно**	**Негативно**
Позитивно	стварно позитивно (TP)	лажно негативно (FN)
Негативно	лажно позитивно (FP)	стварно негативно (TN)

Мере за процену квалитета бинарне класификације које се најчешће користе су приказане у табели 2 [48].

Табела 2. Мере за процену квалитета бинарне класификације

Назив мере	**Формула**
Прецизност	$\text{Прецизност} = \frac{TP}{TP + FP}$
Одзив	$\text{Одзив} = \frac{TP}{TP + FN}$
Тачност	$\text{Тачност} = \frac{TP + \text{TN}}{TP + \text{TN} + FP + FN}$
Ф1-мера	$\text{Ф1} - \text{мера} = \frac{2 * \text{Прецизност} * \text{Одзив}}{\text{Прецизност} + \text{Одзив}}$

У случају вишекласне класификације, потребно је одредити класу (C_i) сваке инстанце из скупа од n могућих класа. Матрица конфузије у овом случају има n колона и врста за сваку од могућих класа. Исходи код вишекласне класификације су:

- Стварно позитивни по класи k (TP_k) – број инстанци код којих је предвиђена класа k и стварна класа је k
- Стварно негативни по класи k (TN_k) - број инстанци код којих предвиђена класа није k и стварна класа није k
- Лажно позитивни по класи k (FP_k) - број инстанци код којих је предвиђена класа k и стварна класа није k

- Лажно негативни по класи k (FN_k) – број инстанци код којих предвиђена класа није k и стварна класа је k

Мере за процену квалитета вишекласне класификације су приказане у табели 3 [52] [53] [54] [55]. Постоје више начина да се комбинују мере по класи у циљу добијања генералних перформанси вишекласног класификатора. Макро - просечне мере представљају аритметичку средину мера по класи – израчунавају се одговарајуће мере по класи за сваку од n класа, и добијене суме се поделе са бројем класа. У овом случају дајемо једнаке тежине свим класама. Макро – тежинске мере се дефинишу када се у просек укључе и број инстанци одређене класе C_i (l_i).

Табела 3. Мере за процену квалитета вишекласне класификације

Назив мере	**Формула**
Микро прецизност (p)	$\frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}(TP_i + FP_i)}$
Микро одзив (r)	$\frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}(TP_i + FN_i)}$
Микро F1 мера (f)	$\frac{2 * p * r}{p + r}$
Микро тачност (a)	$\frac{\sum_{i=1}^{n}(TP_i + TN_i)}{\sum_{i=1}^{n}(TP_i + TN_i + FP_i + FN_i)}$
Прецизност по класи C_i (p_i)	$\frac{TP_i}{TP_i + FP_i}$
Тачност по класи C_i (a_i)	$\frac{TP_i + \mathrm{TN}_i}{TP_i + \mathrm{TN}_i + FP_i + FN_i}$
Одзив по класи C_i (r_i)	$\frac{TP_i}{TP_i + FN_i}$
Ф1-мера по класи C_i (f_i)	$\frac{2 * p_i * r_i}{p_i + r_i}$
Макро – просечна тачност	$\frac{\sum_{i=1}^{n} a_i}{n}$
Макро – просечна прецизност	$\frac{\sum_{i=1}^{n} p_i}{n}$
Макро – просечан одзив	$\frac{\sum_{i=1}^{n} r_i}{n}$
Макро – просечна Ф1-мера	$\frac{\sum_{i=1}^{n} f_i}{n}$
Макро – тежинска тачност	$\frac{\sum_{i=1}^{n} l_i * a_i}{\sum_{i=1}^{n} l_i}$
Макро – тежинска прецизност	$\frac{\sum_{i=1}^{n} l_i * p_i}{\sum_{i=1}^{n} l_i}$
Макро – тежински одзив	$\frac{\sum_{i=1}^{n} l_i * r_i}{\sum_{i=1}^{n} l_i}$
Макро – тежинска Ф1-мера	$\frac{\sum_{i=1}^{n} l_i * f_i}{\sum_{i=1}^{n} l_i}$
Губитак (Hamming loss)	$\frac{1}{n}\sum_{i=1}^{n} \frac{FN_i + FP_i}{TP_i + \mathrm{TN}_i + FP_i + FN_i}$

Процена квалитета кластеровања

Поступак процене квалитета кластеровања укључује решавање више изазова као што су: одређивање оптималног броја кластера, испитивање квалитета кластера, као и процена да ли се резултујућа подела добро слаже са основном структуром података [56]. Такође је неопходно да се утврди у поступку провере резултата кластеровања да ли постоје структуре података које нису случајне у анализираном скупу података [48]. Критеријуми исправности кластеровања су најчешће подељени у три групе:

- Мера спољашње провере коректности кластеровања се користи као процена степена слагања између поделе добијене као резултат кластер анализе и поделе формиране на основу познатих тачних информација, независно од поступка кластеровања. Оне обезбеђују објективну информацију о томе колико је добар добијени модел у односу на праве класе (златни стандард) [57]. У овом случају се тежи ка томе да кластери буду упоредиви са познатим класама тако да сваки кластер садржи елементе највише једне класе и свака класа да буде додељена тачно једно кластеру. У ову групу мера спадају мере тачност, прецизност, одзив, чистоћа, ентропија. Ова мера није употребљива у случају да тачне информације нису познате;
- Мера унутрашње провере коректности кластеровања користи информације добијене током поступка кластеровања и додатни подаци нису неопходни. Дефинисана су два типа мера унутрашње провере коректности кластеровања: мера заснована на хомогености тачака унутар кластера, тачније колико су блиско повезани елементи унутар кластера; и мера заснована на раздвојености између кластера, тачније колико су кластери међусобно различити.
- Мера релативне исправности је мера поређења кластера добијених применом истог алгоритма на различите улазне параметре кластеровања, као и различите подскупове података.

Такође је развијен велики број других техника провере исправности процеса кластеровања које обухватају слагање са постојећом класификацијом, поновљивост, слагање са експертском интуицијом, слагање са разним мултиваријантним методама, тестови значајности, Монте Карло методе, контрола интерне конзистентности и сл. [58].

Мере спољашње провере коректности кластеровања су сличне као мере које се користе у класификацији података, осим што се уместо предвиђених класа узима лабела кластера као вредност класе. Најчешће се користе мере као што су ентропија, прецизност, одзив и Ф-мера [48]. Код мера спољашње провере коректности кластеровања, матрица конфузије се формира на следећи начин:

- нека је n укупан број елемената;
- m_i број елемената који припадају кластеру i;
- c_j број елемената који припадају класи j;
- n_{ij} број елемената који су груписани у кластеру i, а припадају класи j;
- $p_{ij}=n_{ij}/m_i$ вероватноћа да елемент из кластера i припада класи j;

Табела 4. Расподела елемената по кластерима и класама

	Класа 1	**Класа 2**	**Класа 3**	**Укупно по кластеру**
Кластер 1	n_{11}	n_{12}	n_{13}	m_1
Кластер 2	n_{21}	n_{22}	n_{23}	m_2
Кластер 3	n_{31}	n_{32}	n_{33}	m_3
Укупно по класи	c_1	c_2	c_3	

Табела 5. Расподела вероватноће припадности елемента кластера *i* у класи *j*

	Класа 1	**Класа 2**	**Класа 3**
Кластер 1	p_{11}	p_{12}	p_{13}
Кластер 2	p_{21}	p_{22}	p_{23}
Кластер 3	p_{31}	p_{32}	p_{33}

Мере спољашње провере коректности кластеровања су приказане у табели 6 [48] [59].

Табела 6. Мере спољашње провере коректности кластеровања

Назив мере	**Формула**
Ентропија - кластер *i*	$e_i = -\sum_{j=1}^{L} p_{ij} \log p_{ij}$
Ентропија - кластеровање	$e = \sum_{i=1}^{K} \frac{m_i}{n} e_i$
Чистоћа - кластер *i*	$p_i = \max_j p_{ij}$
Чистоћа - кластеровање	$p = \sum_{i=1}^{K} \frac{m_i}{n} p_i$
Прецизност кластера *i* у односу на класу *j*	$Prec(i,j) = p_{ij}$
Одзив кластера *i* у односу на класу *j*	$Rec(i,j) = \frac{n_{ij}}{c_i}$
Ф-мера кластера *i* у односу на класу *j*	$F(i,j) = \frac{2 * Prec(i,j) * Rec(i,j)}{Prec(i,j) + Rec(i,j)}$
Прецизност - кластер *i*	Кластеру *i* је додељена класа k_i , тако да важи $k_i = arg\ \max_j n_{ij}$, $Prec(i) = \frac{n_{ik_i}}{m_i}$
Прецизност - кластеровање	$Prec = \sum_i \frac{m_i}{n} Prec(i)$
Одзив – кластер *i*	$Prec(i) = \frac{n_{ik_i}}{c_{k_i}}$
Одзив – кластеровање	$Rec = \sum_i \frac{m_i}{n} Rec(i)$
Ф-мера - кластеровање	$F = \frac{2 * Prec * Rec}{Prec + Rec}$

Методе за поређење резултата кластеровања

Методе за поређење резултата кластеровања се могу заснивати на нумеричким подацима добијеним током процеса кластеровања, као и поређењем добијених дендрограма као резултата кластер анализе.

Кофенетски кеофицијент корелације и Хопкинсова статистика су начини евалуације хијерархијског кластеровања, и припадају мерама унутрашње провере коректности кластеровања [48]. Циљ ових метода јесте показати да су кластери који су добијени коректни и да структуре у подацима нису случајне.

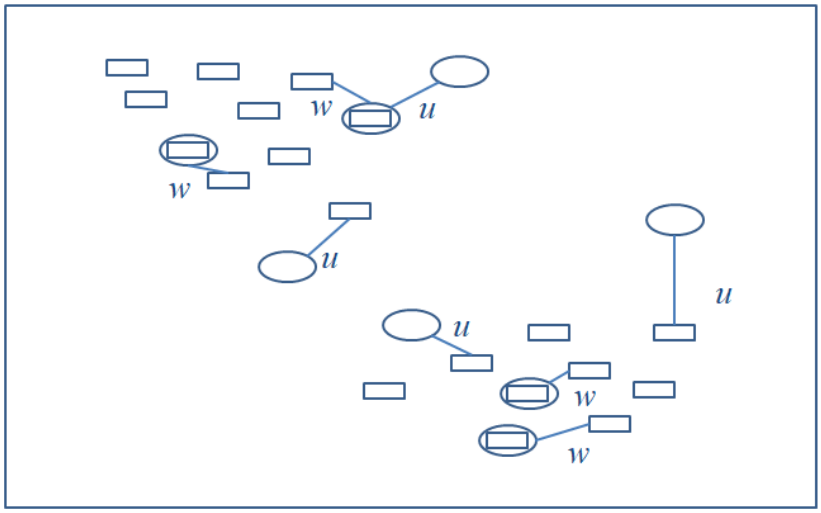
Кофенетски коефицијент корелације кластеровања се добија на основу корелације између матрице сличности и кофенетске матрице растојања. Кофенетска матрица растојања се генерише тако што све тачке у истом кластеру добијају једнаку вредност кофенетског растојања кластера. Вредност кофенетског растојања кластера је најмање растојање између

два кластера која су спојена у тренутку када се формира тај кластер први пут. Најчешће се користи ради утврђивања који тип хијерархијског кластеровања је најбољи.

Хопкинскова статистика се користи за процену тендеције груписања скупа података тестирањем случајне расподеле података у простору. Формира се скуп U од p случајно распоређених тачака у простору, као и скуп W од p тачака из стварног скупа података. За оба скупа података израчунавамо растојање са најближим суседом у изворном скупу података. Са w_i означимо растојања тачака из скупа W са најближим тачкама из оригиналног скупа. Са u_i означимо растојања тачака из скупа U са најближим тачкама из оригиналног скупа (Слика 2). Хопкинсова статистика је дефинисана следећом једначином:

$$H = \frac{\sum_{i=1}^{p} u_i}{\sum_{i=1}^{p} w_i + \sum_{i=1}^{p} u_i}$$

У случају да не постоји структура у скупу података, онда ће растојања између оригиналних тачака w_i и случајно распоређених тачака u_i у просеку бити слична. Када случајно изабране тачке и тачке из изворног скупа података имају сличне вредности растојања, тада је вредност H близу 0.5. У случају да су групе присутне у скупу података, онда ће вредности растојања случајних тачака u_i бити знатно већи од растојања оригиналних тачака w_i, чиме ће вредност статистике H бити близу 1. Вредности H које су близу 1 показују да су подаци веома кластерабилни. Вредности H које су близу нуле показују да су подаци равномерно дистрибуирани у датом простору.



Слика 2. Правоугаоницима су представљене тачке из изворног скупа података. Тачке скупа случајно распоређених тачака у простору су представене круговима. *w* представља растојање тачака скупа W и њима најближих тачака. *u* означава растојање случајно распоређене тачке у простору са најближом тачком из изворног скупа података. Слика је преузета од Lawson и сарадника [60] и адаптирана.

Развијено је више метода за поређење дендрограма. Једна метода која је међу првима развијена је кофенетски коефицијент корелације између два дендрограма, који представља изузетно једноставан и ефикасан метод за поређење различитих врста дендрограма [61]. Он се заснива на исцртавању једнако удаљених хоризонталних линија, зване фенонске линије на дендрограмима, и подели дендрограма на зоне које су нумерички означене. Затим се формира матрица кофенетских вредности између сваког објекта дендрограма, чије вредности представљају ознаку зоне у којој се налази чвор који спаја та два објекта. Израчунавањем кеофицијента корелације између елемената две матрице кофенетских вредности два различита дендрограма, добија се кофенетски коефицијент корелације два дендрограма.

Виша вредност кофенетског коефицијента корелације указује и на већу сличност између два дендрограма која се упоређују.

Друга метода се назива Robinson Folds растојање (у даљем тексту је означено са Р-Ф растојање). Р-Ф метрика је најчешће коришћена у анализи поређења филогенетских стабала, и заснива се на трансформацији стабла из једног у друго [62] [63]. Вредност Р-Ф растојања која је једнака нули означава једнаке топологије два дрвета (дендрограма) који се упоређују. Вредност Р-Ф растојања се повећава у случају када се сличност два дендрограма смањује.

2.4.5. Нормализација података

Нормализација података представља трансформацију података у циљу добијања резултата у одређеном интервалу, ради лакшег поређења резултата. Најчешће коришћени интервал је од -1 до 1 и од 0 до 1. Постоје разне методе нормализације, и једна од њих је мин-макс нормализација описана у наставку за интервал од 0 до 1.

Дефиниција 9: Мин-макс нормализација је линеарна трансформација која мапира оригиналну вредност N у нормализовану вредност $\widehat{N}$ у интервалу $[0,1]$ користећи следећу формулу:

$$\widehat{N} = \frac{N - \min(N)}{\max(N) - \min(N)}$$

2.4.6. Визуелизација података

Визуелизацијом података се пресликавају подаци (укључујући и резултате неке анализе) у графички или табеларни формат ради лакшег и бржег разумевања информација. Начин и врста приказа зависе од података. Неке од метода визуелизације које су коришћене у раду су набројане у наставку.

Шеме са распршеним елементима (тачкасти приказ) се често користе за приказ резултата кластеровања као и односа између два или више објеката. Сваки објекат је приказан као тачка у равни. Припадност различитим кластерима је означено различитим маркерима или бојама (Слика 3a).

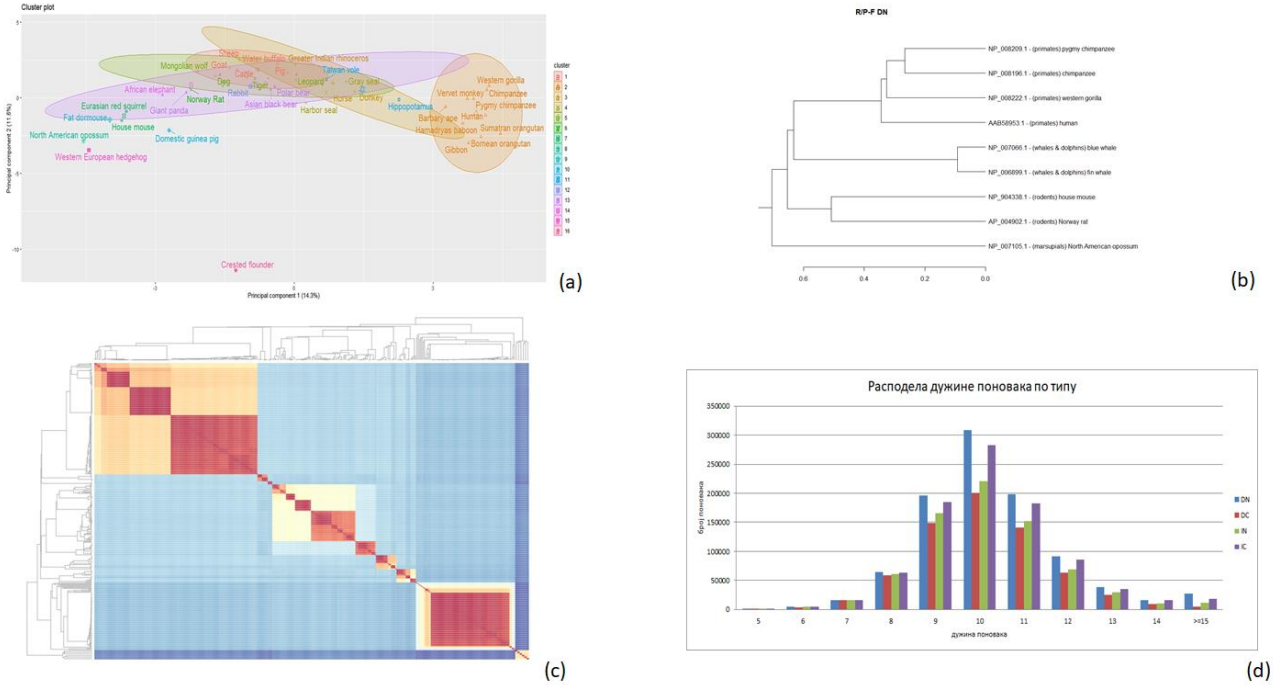
Дендрограм је дијаграм у облику стабла који приказује узајамне односе различитих објеката. Често се користи за приказ резултата хијерархијског кластеровања илуструјући распоред кластера као и сличност између објеката, као и у филогенетици за приказ еволуцијских односа између врста (Слика 3b).

Шема са матрицама (топлотна мапа) представља графичку репрезентацију података где су величине променљивих представљене бојама у дводимензионој матрици. Матрице садржи информације о сличностима и различитостима и корисне су за представљање односа између објеката. Варијације у боји могу бити према нијанси или интензитету, које у комбинацији са сортирањем редова или колона даје јаснији приказ о могућим кластерима (Слика 3c).

Бар графикон (*bar plot*) представља графички приказ који омогућава поређење описаних ставки које су представљене графиконом. Варијације у боји омогућава додатно поређење по групама. Ова врста графикона је примењена на визуелизацију вредности мера за процену квалитета модела класификације формираних у анализи и примењених на различите параметре (Слика 3d).

Анализа главних компоненти представља статистичку анализу редукције димензионалности скупа података, на начин да буде обухваћена што већа количина варијансе података. Примењује се када постоје редундантне променљиве у анализираном скупу података у циљу траснформације оригиналних променљивих у нове (главне компоненте) које су мањих димензија. Примена методе анализе главних компоненти је

широка приликом решавања различитих проблема и добијене главне компоненте најчешће представљају улазни податак у даљу анализу. Додатан значај ове методе јесте што омогућава визуализацију вишедимензионих података. Применом ове методе, се једноставно визуализују вишедимензиони подаци приказом истих у две или три димензије са минималним губитком. На тај начин се добија приказ података који је једноставнији за разумевање (Слика 3а).



Слика 3. (а) пример графичког приказа резултата кластеровања; (b) пример приказа разултата хијерархијског кластеровања у облику дендрограма; (c) пример приказа резултата користећи топлотну мапу; (d) пример представљања резултата користећи бар графикон.

3. Нове методе за анализу сличности биолошких секвенци

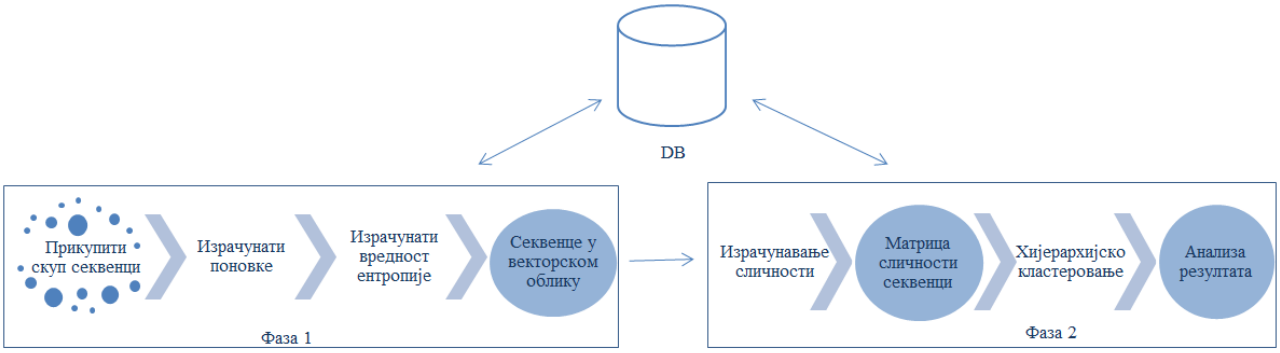
Главни допринос ове дисертације огледа се у развијеним новим методама за анализу сличности биолошких секвенци. Методе су засноване на статистички значајним поновцима различитих типова и дужина, као и примени метода из области истраживања података: трансформације података, мера сличности објеката, класификације и кластеровања. Развијене методе припадају категорији метода за анализу сличности секвенци без поравнања. Предложене методе обухватају трансформацију биолошких секвенци и представљање истих у облику који је погодан за даље истраживање и поједностављену обраду података. Прва метода се заснива на концепту теорије информација и кластеровању података. Друга метода је утемељена на формирању потписа секвенци и профила таксономских категорија, као и класификацији секвенци представљених атрибутима.

3.1. Метода заснована на позицији и локалној учесталости поновака

Метода заснована на позицији и локалној учесталости поновака, у даљем тексту означено са R-P/F метод (енгл. *Repeats-Position/Frequency method*), за одређивање сличности биолошких секвенци се заснива на теорији информација, узимајући у обзир број појављивања, позицију понављајуће секвенце, као и чињеницу да није очекивано истоветно појављивање поновака у случајно изабраним секвенцама исте дужине [64]. Секвенце су представљене нумеричким векторима у вишедимензионом векторском простору и односи између секвенци су идентификовани користећи мере за одређивање растојања вектора у векторском простору. На основу добијених резултата формира се матрица сличности, која се даље користи у алгоритмима хијерархијског кластеровања.

3.1.1 Опис *R-P/F* методе

Поступак анализе сличности секвенци R-P/F методом реализује се у две фазе (Слика 4). Прва фаза обухвата припрему података и упис у базу података, док друга фаза обухвата израчунавање сличности секвенци и упис резултата у базу података.



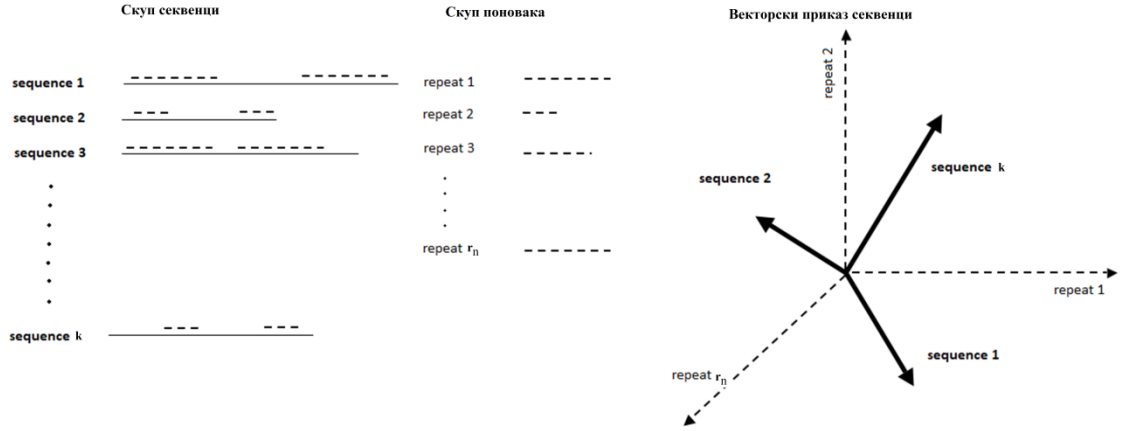
Слика 4. Фазе R-P/F методе. Круговима су представљени улазни и излазни подаци модула. Стрелице означавају функционалне модуле. Слика је преузета из [64] и адаптирана

База података се састоји од три дела:

- скупа јединствених поновака садржаних у секвенцама које се анализирају
- карактеристика секвенци које се анализирају
- матрице сличности/растојања секвенци које су представљене векторима

Секвенца је трансформисана у нумерички вектор (Слика 5) користећи статистички значајне поновке различитих дужина. Компоненте векторске репрезентације једне секвенце

су израчунате на основу ентропије засноване на локалној учесталости поновка, узимајући у обзир како учесталост поновка, тако и позицију појављивања поновка који се налази у скупу јединствених поновака анализираних секвенци (прва компонента описане базе података). Димензија датог векторског простора је једнака броју поновака који се користе за формирање вектора. Свака компонента вектора наглашава значајност одређеног поновка у секвенци. У овом кораку садржај базе података, као и димензија векторског простора се динамички повећава, у случају да улазна секвенца садржи поновке који се не налазе у јединственом скупу поновака. У другом кораку се израчунава/проширује матрица сличности улазних секвенци, применом мере сличности на векторске приказе секвенци. Добијена матрица сличности се користи као улазни податак у алгоритам хијерархијског кластеровања, којим се одређује сличност између улазне секвенце и секвенци сачуваних у бази података. Детаљи о подацима и фазама методе су описани у даљем тексту.



Слика 5. Векторски приказ секвенци. Секвенце су представљене векторима истих димензија. Димензија векторског простора је једнака броју различитих поновака истог типа. Свака секвенца има 4 различита векторска приказа у односу на изабрани тип поновка. Слика је преузета из [64] и адаптирана

3.1.2 Фаза припреме података и пуњење базе података

Идентификација поновака

У R-P/F методи се користе максимални статистички значајни поновци, ради издвајања изузетно важних биолошких сигнала у секвенци. На тај начин се понављајуће секвенце, које се вероватно јављају и у случајно генерисаној секвенци исте дужине, искључују из анализе. Не мењајући форму, предложена метода прихвата сва 4 типа поновака (DN, IN, DC и IC) описаних у претходном тексту за нуклеотидну секвенцу и два типа поновака (DN и IN) за протеинску секвенцу. Нови идентификовани поновци се уписују у базу података, динамички проширујући димензије векторског простора.

Израчунавање ентропије засноване на локалној учесталости поновка

Свака компонента вектора одражава важност одређеног поновка у датој секвенци и једнака је вредности ентропије засноване на локалној учесталости датог поновка. Ова вредност се израчунава за сваки поновак из скупа поновака R_c и сваку секвенцу из скупа секвенци S_c у бази података. Улазна секвенца се на овај начин се трансформише у нумерички вектор (Слика 5).

Израчунавање вредности ентропије засноване на локалној учесталости (h_{sr}) за секвенцу $s \in S_c$ дужине m и поновка $r \in R_c$ је описано у корацима испод:

(1) Нека је n број појављивања поновка r у секвенци s;
(2) Дефинисати l_p^{rs} као позицију p-тог појављивања поновка r у секвенци s, где је l_0^{rs} једнако 0 и важи да $p\epsilon\{1,2,\dots,n\}$;
(3) Израчунати LF_i^{rs} −вредност локалне учесталости поновка r у секвенци s пребројавајући растојања између два суседна појављивања поновка r у секвенци s узимајући реципрочну вредност у обзир на следећи начин:

$$LF_i^{rs} = \begin{cases} \frac{1}{l_p^{rs} - l_{p-1}^{rs}}, & \text{ако се поновак } r \text{ налази на позицији } i = l_p^{rs} \\ 0, & \text{иначе} \end{cases}, i = 1, \dots, m \quad (1)$$

(4) Формирати вектор вредности локалне учесталости поновка r у секвенци s дужине m:

$$LF_{rs}=(LF_1^{rs}, LF_2^{rs}, \dots, LF_m^{rs}),\ m \text{ је дужина секвенце } s. \quad (2)$$

(5) Израчунати парцијалне суме LF_{rs} на следећи начин:

$$u_{ri} = \sum_{l=1}^{i} LF_l^{rs}, i \in \{1, .., m\} \quad (3)$$

(6) Формирати вектор парцијалних сума U_r:

$$U_r = \{u_{r1}, u_{r2}, \dots, u_{rm}\}$$
$$= \{LF_1^{rs}, LF_1^{rs} + LF_2^{rs}, LF_1^{rs} + LF_2^{rs} + LF_3^{rs}, \dots, LF_1^{rs} + LF_2^{rs} + \dots + LF_m^{rs}\} \quad (4)$$

(7) Израчунати Z које је једнако збиру елемената вектора U_r:

$$Z = \sum_{i=1}^{m} u_{ri} = \sum_{i=1}^{m} \sum_{l=1}^{i} LF_l^{rs} \quad (5)$$

(8) Израчунати вредност q_i на следећи начин:

$$q_i = \frac{u_{ri}}{Z}, (i = 1,2, \dots, m) \quad (6)$$

(9) На крају, вредност ентропије засноване на локалној учесталости h_{sr} секвенце s и поновка r се израчунава на основу следеће формуле

$$h_{sr} = -\sum_{i=1}^{m} q_i log_2 q_i \quad (7)$$

Формирање векторског простора

Векторски простор се формира на основу скупа од k секвенци S_c и скупа од r_n јединствених поновака R_c. Вредност ентропије засноване на локалној учесталости поновка се израчунава за сваку секвенцу у скупу S_c и поновак из скупа R_c. На овај начин, свака секвенца је представљена r_n -димензионим нумеричким вектором, где су елементи вектора једнаки вредности ентропије засноване на локалној учесталости поновка

$$\mathrm{S_i} = \left(\mathrm{h_{s_i r_1}}, \mathrm{h_{s_i r_2}}, \dots, \mathrm{h_{s_i r_n}}\right), \mathrm{i} = 1,2, \dots, \mathrm{k}.$$

Величина векторског простора $\mathrm{r_n}$ је једнака кардиналности скупа поновака укључених у анализу. Уколико секвенца не садржи дати поновак, вредност $\mathrm{h_{sr}}$ је једнака нули.

Пример израчунавања ентропије засноване на локалној учесталости и примене R-P/F методе

Пример израчунавања ентропије засноване на локалној учесталости поновака DN типа и примене R-P/F методе за секвенцу *s=TAAGTCTGACCGTCAGACT* и њене поновке је описан у следећим корацима:

1. Идентификовати директне некомплементарне поновке (DN) дуже од два нуклеотида у секвенци *s=TAAGTCTGACCGTCAGACT*:
   1.1. Директни некомплементарни поновак GTC се налази на позицијама 4 и 12 у секвенци TAAGTCTGACCGTCAGACT
   1.2. Директни некомплементарни поновак GAC се налази на позицијама 8 и 16 у секвенци TAAGTCTGACCGTCAGACT,

   напомена: за издвајање поновака у примеру коришћен је програм описан у делу имплементација (поглавље 3.1.5).
2. Формирати векторски простор на основу идентификованих различитих статистички значајних поновака истог типа V = {GTC, GAC}, r_n =2, где је r_n димензија векторског простора. Секвенца ће бити представљена у векторском простору димензије 2,
3. Вредност ентропије засноване на локалној учесталости за поновак r=GTC у секвенци s=TAAGTCTGACCGTCAGACT, где је m = 19 дужина секвенце се израчунава пратећи следеће кораке:
   3.1. Израчунати елементе вектора који садржи вредности локалне учесталости поновка r=GTC у секвенци s на позицијама 4 и 12 на следећи начин:

$$LF_4^{rs} = \frac{1}{4-0} = \frac{1}{4},$$
$$LF_{12}^{rs} = \frac{1}{12-4} = \frac{1}{8},$$
$$LF_p^{rs} = 0, p \neq 4 \text{ и } p \neq 12, p > 0 \text{ и } p \leq m$$
$$LF_{rs}=\{0, 0, 0, \frac{1}{4}, 0, 0, 0, 0, 0, 0, 0, \frac{1}{8}, 0, 0, 0, 0, 0, 0, 0\}$$

   3.2. Израчунати вектор парцијалних сума на следећи начин:

$$u_{r1} = LF_1^{rs} = 0$$
$$\dots$$
$$u_{r4} = LF_1^{rs} + LF_2^{rs} + LF_3^{rs} + LF_4^{rs} = 0 + 0 + 0 + \frac{1}{4} = \frac{1}{4}$$
$$u_{r5} = LF_1^{rs} + LF_2^{rs} + LF_3^{rs} + LF_4^{rs} + LF_5^{rs} = 0 + 0 + 0 + \frac{1}{4} + 0 = \frac{1}{4}$$
$$\dots$$
$$U_r = \left\{0, 0, 0, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{3}{8}, \frac{3}{8}, \frac{3}{8}, \frac{3}{8}, \frac{3}{8}, \frac{3}{8}, \frac{3}{8}\right\}$$

   3.3. Израчунати вредност Z:

$$Z = 0 + 0 + 0 + \frac{1}{4} + \frac{1}{4} + \cdots + \frac{3}{8} = 5$$

   3.4. Израчунати q_i вредности:

$$q_1 = \frac{u_{r1}}{Z} = \frac{0}{5} = 0$$
$$\dots$$
$$q_4 = \frac{u_{r4}}{Z} = \frac{\frac{1}{4}}{5} = \frac{1}{20}$$
$$\dots$$
$$q_{12} = \frac{u_{r12}}{Z} = \frac{\frac{3}{8}}{5} = \frac{3}{40}$$

...

3.5. Израчунати вредност ентропије засноване на локалној учесталости поновка p=GTC за секвенцу s: TAAGTCTGACCGTCAGACT

$$h_{sr_1} = -\sum_{i=1}^{m} q_i \log_2 q_i = 16.66$$

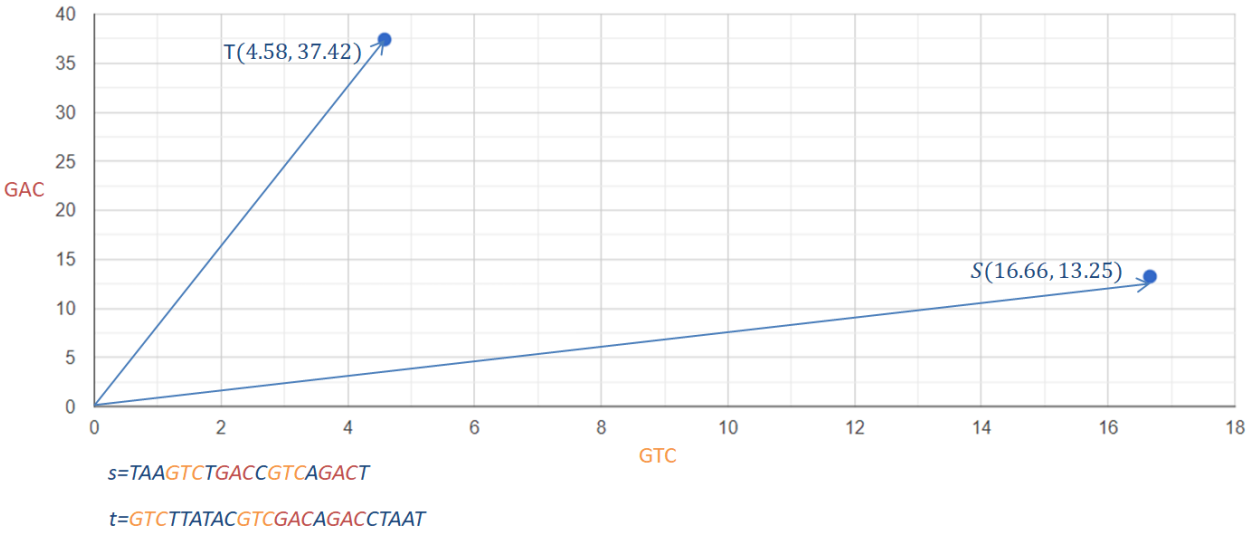
4. На исти начин израчунати вредност ентропије засноване на локалној учесталости поновка GAC у секвенци TAAGTCTGACCGTCAGACT:

$$h_{sr_2} = -\sum_{i=1}^{m} q_i \log_2 q_i = 13.25$$

5. Секвенца s је представљена у дводимензионом векторском простору на основу горе израчунатих вредности ентропија заснованих на локалној учесталости поновака {GTC, GAC} (Слика 5):

$$\vec{s} = (16.66, 13.25)$$

На сличан начин се за секвенцу t= *GTCTTATACGTCGACAGACCTAAT* и њене поновке {GTC, GAC} применом R-P/F методе добија дводимензиони вектор $\vec{t} = (4.58, 37.42)$ (Слика 6).



Слика 6. Илустрација векторског приказа секвенци s=TAAGTCTGACCGTCAGACT и t=GTCTTATACGTCGACAGACCTAAT добијена на основу примене R-P/F методе за DN тип поновака. Секвенце су представљене у дводимензионом векторском простору, где *x* координата представља поновак GTC и *y* координата поновак GAC.

Пуњење базе података

Нека дата база података садржи скуп секвенци $S_c = \{s_1, s_2, \dots, s_k\}$ и њему припадајући скуп јединствених поновака $R_c = \{r_1, r_2, \dots, r_{nc}\}$. Скуп вектора са карактеристикама секвенци је означен са $V_c = \{v_1, v_2, \dots, v_k\}$. Дужина вектора v_i је једнака броју јединствених поновака садржаних у бази података $|v_i| = r_{nc}$. J-ти елемент вектора v_i једнак је вредности ентропије засноване на локалној учесталости поновка r_j у секвенци s_i.

Додавање нове секвенце Q у базу података је описано следећим корацима:

1. Израчунати скуп јединствених поновака R_Q у улазној секвенци Q и применити унију истог на постојећи скуп R_c. Нов добијени скуп јединствених поновака је означен са $R = \{r_1, r_2, \dots, r_n\}$, $r_n \geq r_{nc}$
2. Било која секвенца s_i из скупа $S = \{s_1, s_2, \dots, s_{k,} s_{k+1,}\}$, где је $Q = s_{k+1}$ је представљена вектором дужине r_n и трансформише се на следећи начин:

2.1. Ако је $i<=k$ (секвенца већ постоји у бази података), нове координате вектора се додају на постојећу репрезентацију секвенце v_i за сваки ново идентификовани поновак у секвенци Q. Вредност нових координата је једнака нули. Нова димензија вектора карактеристика једне секвенце је једнака r_n. Вредности додатих координата се уписују у базу података.

2.2. Ако је $i = k+1$, израчунати вредности координата вектора v_{k+1}, дужине r_n на основу описаног поступка за израчунавање вредности ентропије локалне учесталости поновка за секвенцу Q и скуп поновака R. Вредности координата вектора v_{k+1} се уписују у базу података.

3.1.3 Фаза израчунавања сличности секвенци и пуњење базе података

Формирање матрице сличности

Сличност нуклеотидних секвенци S_i и S_j се добија израчунавањем косинусне сличности векторских репрезентација датих секвенци. Израчунавањем сличности свих парова секвенци у бази података, формира се матрица сличности секвенци M_c. На описан начин такође могу да се примене и друге мере сличности вектора.

Нека је $M_c = \{\{m_{11}, m_{12}, \dots, m_{1k}\}, \{m_{21}, m_{22}, \dots, m_{2k}\}, \dots, \{m_{k1}, m_{k2}, \dots, m_{kk}\}\}$ постојећа матрица сличности за k секвенци из скупа S_c. Приликом додавања нове секвенце Q у базу података, матрица сличности M се израчунава тако што се додају вредности $m_{(k+1)i}$ за нову секвенцу Q у постојећу матрицу M_c, израчунавањем сличности векторских репрезентација нове секвенце Q и векторских репрезентација већ постојећих секвенци у бази података.

Фаза обраде података

У овој фази се примењује алгоритам сакупљајућег хијерархијског кластеровања, ради груписања сличних секвенци. Нуклеотидна секвенца се придружује групи секвенци којима је највише слична на основу мере косинусне сличности примењене на векторске репрезентације секвенци.

3.1.4 Временска сложеност израчунавања

Временска сложеност израчунавања вредности ентропије засноване на локалној учесталости поновка за секвенцу дужине m и један поновак је $O(m)$. Временска сложеност R-P/F методе за трансформисања секвенце у вектор је $O(m*r_n)$, где је са m означена дужина секвенце и r_n је једнако димензији векторског простора, тачније броју јединствених поновака укључених у анализу. Временска сложеност за формирање свих вектора у бази података је $O(m*r_n*k)$ за k секвенци.

3.1.5 Имплементација и прикупљање података

За израчунавање различитих типова поновака коришћен је већ имплементиран програм StatRepeats (прво издање, верзија 2) [5] [6]. Подразумевана p=0.05 вредност је коришћена као један од улазних параметара. Подаци о улазним секвенцама, идентификовани поновци различитих типова, векторске репрезентације секвенци и матрица сличности су чувани у IBM Db2 релационој бази података [65], која се користи као репозиторијум података током фаза припреме, обраде и анализе података.

R-P/F метода је имплементирана у R програмском језику [66]. Имплементирана верзија укључује преузимање неопходних филтрираних података неопходних за анализу из релационе базе података, израчунавање вредности локалне учесталости поновака, формирање векторске репрезентације секвенци, израчунавање матрице сличности секвенци, кластеровање и проверу. Коришћени R пакети су приказани у наставку:

- пакет *RODBC* [67] за повезивање на IBM Db2 релациону базу података и обраду података;
- пакет *parallel* [66] за паралелно извођење већег броја израчунавања;
- пакет *Lsa* [68] за израчунавање матрице сличности на основу метрике косинусног растојања, као и за исцртавање дендрограма;
- пакет *stats* [66] за примену алгоритма кластеровања на добијену матрицу сличности;
- пакети *d3heatmap* [69], *purr* [70] и *factoextra* [71] за визуелизацију резултата кластеровања;
- пакети *corrplot* [72], *dendextend* [73] и *phangorn* [74] [75] за поређење денрограма

Провера имплементиране методе је извршена на различитим скуповима биолошких секвенци. Подаци о секвенцама су преузети из NCBI [13] и ViPR [15] база података Такође, подаци о таксономијама секвенци су преузети из NCBI базе података користећи FTP сервер (линк ftp://ftp.ncbi.nih.gov/pub/taxonomy/accession2taxid/). У R програмском језику написан је скрипт за издвајање само оних података о таксономији који су неопходни за даљу анализу у изабраном скупу података. Пакет *taxonomizr* [76] је коришћен за упаривање и издвајање таксономских података за појединачне секвенце од интереса.

3.2. Методе засноване на потписима секвенци и профилима категорија

У циљу доприноса описаним изазовима класификације секвенци, с обзиром на велику примену истих, као и проналажењу другачијих метода за решавање наведених проблема, у овом одељку су предложене нове методе за карактеризацију биолошких секвенци, одређивање сличности биолошких секвенци као и класификовање истих. Проблему идентификовања сродних секвенци, као и њиховој класификацији, се у овом одељку приступа тако што се граде потписи биолошких секвенци, као и профили таксономских група. Ове методе се темеље како на статистички значајним понављајућим деловима секвенци, тако и на растојањима између поновака. Идеја коришћена у методама се заснива на коришћењу (мањег) скупа карактеристичних тачака секвенци, уместо читавих секвенци.

Ове методе су засноване на трансформацији биолошких секвенци у потписе секвенци и профиле категорија на основу њихових статистички значајних понављајућих делова секвенци различитих типова. Верује се да секвенце, које припадају истим групама, деле исту структуру репрезентативних секвенци. Анализа је фокусирана на понављајуће делове секвенци, јер се претпоставља да се важне еволуционе информације могу издвојити из понављајућих делова секвенци и њихових односа, као и растојања између њих. Потписи секвенци и профили таксономских група репрезентативних биолошких секвенци се користе у формирању базе података ради даље анализе биоинформатичким методама.

Развијене су две нове методе:

- одређивање сличних секвенци међусобним поређењем потписа секвенци (у даљем тексту *1:1 поређење потписа секвенци*)
- класификација секвенци заснована на профилима категорија

Прва метода омогућава одређивање сличности секвенци, укључујући и секвенце које припадају различитим таксономским категоријама. Друга метода омогућава одређивање ужег скупа података, тачније категорију којој та секвенца припада.

Растојања између парова поновака (d_r) су коришћена ради смањења броја могућих промена које су се десиле током еволуције. Растојање између леве и десне компоненте пара поновака (d_r) у датој секвенци (s) се израчунава као разлика између почетних позиција леве (p_l) и десне (p_r) компоненте пара поновка ($d_r = p_r - p_l$). Лева и десна компонента пара поновака, као и растојање између ових компоненти је означено уређеним паром $<r, d_r>$ у даљем тексту. Идентични поновци који су идентификовани на више различитих позиција у секвенци се пресликавају у више уређених парова $<r, d_r>$.

На основу идентификованих парова поновака и њихових растојања $<r, d_r>$ се генерише скуп потписа секвенци и профила категорија. Потпис једне секвенце дефинишемо као скуп свих уређених парова $<r, d_r>$ који су садржани у датој секвенци. Потпис секвенце је вектор уређених парова, који може да садржи већи број $<r, d_r>$ са идентичним вредностима за поновак r и растојање d_r. Приликом формирања модела, коришћене су секвенце са познатим таксономским категоријама. Профил таксономске категорије се формира на основу потписа секвенци које припадају датој категорији у скупу података за тренирање. Профил таксономске категорије се састоји од различитих уређених парова $<r, d_r>$. Број секвенци неке категорије које садрже уређени пар $<r, d_r>$ представља карактеристику категорије. У случају да је уређен пар $<r, d_r>$ идентификован у две или већем броју категорија, исти се искључује из профила тих категорија.

Сличност улазних секвенци се израчунава одређивањем сличности између потписа секвенци. Сврха методе *1:1 поређење потписа секвенци* јесте идентификовање сличних секвенци као и њихових односа. Такође, дата метода може да се користи у идентификовању таксономске категорије секвенце (потпуна или делимична) која нема идентификовану таксономску категорију поређењем потписа те секвенце са базом потписа познатих секвенци. С обзиром да је метода независна од дужине секвенце, иста може да се користи за поређење дела секвенце са скупом секвенци у бази података.

Прва иницијална фаза методе се извршава само једном и садржи формирање базе потписа секвенци скупа података за тренирање. У другом кораку се идентификују сличне секвенце са улазном секвенцом поређењем потписа улазне секвенце са потписима секвенци који су већ садржани у бази података. Додатно се база података проширује потписом улазне секвенце. Поред секвенци и њихових поновака различитих карактеристика и дужина у бази података се чувају и остали мета подаци као што су: информације о различитим нивоима таксономије за одређену секвенцу, карактеристике категорије, тип поновака, дужина поновака.

Формирање потписа секвенци представља трансформацију биолошких секвенци у векторе атрибута на које се могу применити добро познате методе истраживања података као што су кластеровање и класификација. Додатно, на овај начин можемо да идентификујемо сличне секвенце само на основу нуклеотидне секвенце, без додатних информација о особинама истих. Различите мере сличности се могу применити на добијене потписе секвенци ради утврђивања степена сличности.

Класификација секвенци заснована на профилима категорија представља идентификовање категорије улазне секвенце на основу метода истраживања података. Класификација улазне секвенце се заснива на одређивању потписа дате секвенце и поређењем истог са базом профила категорија. Предложени метод може да се примени на различите таксономске категорије, или друге групе за које секвенце деле заједничке особине.

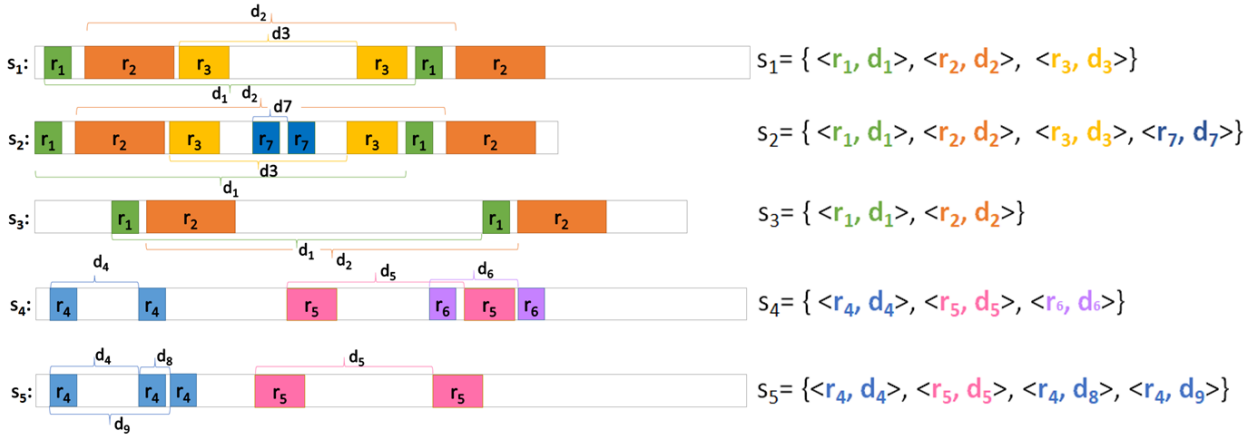
3.2.1. Опис методе Одређивање сличних секвенци поређењем потписа секвенци

У првом кораку методе се формира база података која садржи потписе улазних секвенци. За сваку секвенцу s_i из скупа података за тренирање S = {s_1, s_2,..., s_n} формира се вектор атрибута на основу функције *CreateSequenceSignature* (Слика 7). Улазни параметри су секвенца *s*, тип поновка (могуће вредности су *DN*, *DC*, *IN* или *IC*) и минимална дужина поновка. Функција се извршава више пута за различите комбинације улазних параметара.

```
function CreateSequenceSignature (ns, rt, min_repeat_length, initial)
/* ns  улазна нуклеотидна секвенца */
/* rt тип поновка: DN, IN, DC или IC, min repeat length - integer constant ≥ 0   */
/* initial = 1 у случају позива функције у првом кораку, 0 у осталим случајевима */
begin
израчунати растојања између леве и десне компоненте поновка типа rt у секвенци ns;
креирати sequence_signature листу уређених парова <r, dr> тако да важи dr > 0;
    if (initial=1)
        then begin
            if sequence_signature за улазне параметре већ постоји у бази података;
                then do_nothing;
            else сачувај израчунату sequence_signature и метаподатке у бази података;
             end
     return  sequence_signature;
end
```

Слика 7. Опис корака функције CreateSequenceSignature формирања потписа секвенци

Израчунат потпис секвенце се чува у бази података, заједно са вредностима мета-података (тип поновка и минимална дужина поновка). Добијени потпис секвенце може да садржи више идентичних уређених парова <*r*, d_r> у случају да се исти пар поновака са идентичним растојањем налази на различитим позицијама у секвенци. Пример вектора атрибута који одговара трансформисаној секвенци је приказан на слици 8.

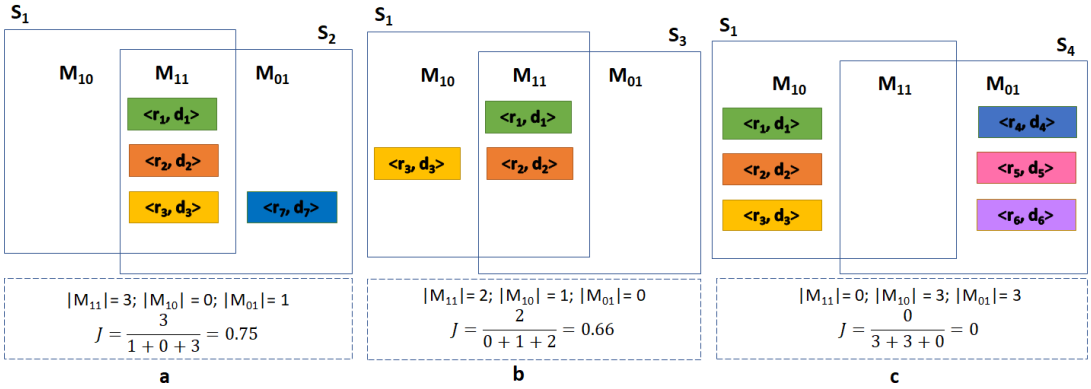


Слика 8. Пример формирања вектора атрибута (потписа секвенци) S_1, S_2, S_3, S_4 и S_5 за 5 различитих секвенци које су означене са s_1, s_2, s_3, s_4 и s_5. Различити уређени парови <*r*, d_r> су означени различитим бојама. Идентични уређени парови <*r*, d_r> су означени истом бојом у различитим секвенцама. Позиције поновака који су део уређеног пара <*r*, d_r> се могу разликовати унутар различитих секвенци.

Други корак методе садржи одређивање потписа секвенце S_u за улазну секвенцу s_u применом фукнције *CreateSequenceSignature*. Када је формиран потпис улазне секвенце, израчунава се сличност између потписа улазне секвенце и потписа сачуваних секвенци у бази за исте улазне параметре (тип поновака и минимална дужина поновка) применом мере сличности између два потписа секвеци представљених вектором атрибута. Мере које се користе су Жакардово растојање, косинусно растојање и коефицијент Танимотоа (описано у поглављу 2.4.3). Приликом примене косинусног растојања и коефицијента Танимотоа, број појављивања уређеног пара <r, d_r> учествује у одређивању сличности секвенци. Приликом примене Жакардовог растојања, присуство атрибута, тачније уређеног пара <r, d_r> је означено са 1 у случају да постоји један или више уређених парова са истим вредностима <r,

$d_r>$, и 0 у случају одсуства тог атрибута. Коришћење различитих мера сличности може директно да утиче на прецизност добијеног модела.

Илустративни пример одређивања сличности парова потписа секвенци, користећи Жакардову меру сличности је приказан на слици 9. Што је већа вредност Жакардовог коефицијента између потписа две секвенце, то су секвенце сличније.



Слика 9. Нека су дате 4 секвенце илустроване на слици 6: s_1, s_2, s_3 и s_4 и њихови потписи S_1 = {<r_1, d_1>, <r_2, d_2>, <r_3, d_3>}, S_2 = {<r_1, d_1>, <r_2, d_2>, <r_3, d_3>, <r_7,d_7>}, S_3= {<r_1, d_1>, <r_2, d_2>} и S_4 = {<r_4, d_4>, <r_5, d_5>, <r_6, d_6>}. На сликама a, b и c је приказано израчунавање Жакардове мере сличности између парова секвенци представљених векторима атрибута за пар S_1 и S_2 (на делу слике који је означен са a), за пар S_1 и S_3 (на делу слике који је означен са b), и за пар секвенци S_1 и S_4 (на делу слике који је означен са c). M_{11} представља подскуп скупа елемената потписа секвенци (уређених парова <r,d>) који припадају потписима обе секвенце. M_{10} и M_{01} представљају подскупове скупова елемената потписа секвенци који припадају једној, али не припадају другој секвенци. Број елемената сваког подскупа M_{11}, M_{10}, M_{01} се користи за израчунавање Жакардове сличности између две секвенце. На основу добијених резултата може да се утврди да је секвенца s_1 најсличнија секвенци s_2 ($J(s_1, s_2)=0.75$), и најмање слична секвенци s_4 ($J(s_1, s_4)=0$) у приказаном примеру.

3.2.2. Пример примене методе одређивање сличних секвенци поређењем потписа секвенци са Жакардовом мером сличности

Нека су дате 3 ДНК секвенце s_1, s_2 и s_3:

s_1:ACATAAGGTCCAAGCTAAGCTCCGCGCGCTGAAATTGCGATGCTGATTCCGCGCATGATCTCGTCTCGGTCATTCGCGCCG

s_2:AACATAAGGTCCGCTCCGCGCGGCGATGCTGATTCCGCGCATGATCTCTCGGTCATTCGCGCCGAAATATCTTAAATGTC

s_3:TCAAGGTCCGCTCCGCGCGGCGCATGCAGACCCCCCGTTCAGCTATTGGTCATTCGCGCTGTTGCATGAAGGCCTCACGCACGG

и улазни подаци који садрже DN тип поновака, као и 0 за вредност улазног параметра минимална дужина поновака (тачније сви статистички значајни поновци су укључени у анализу).

Корак 1: Као што је описано у тексту, поновци представљају максималне директне некомплементарне статистички значајне понављајуће делове секвенци. Подразумевана вредност p=0.05 је изабрана за одређивање статистичке значајности поновака. Када је нула наведена као вредност улазног параметра минимална дужина поновака, поновци са најнижом могућом вредношћу дужине која задовољава постављене услове статистичке значајности се издвајају. Растојање између поновака представља број нуклеотида између почетних позиција леве и десне компоненте пара идентификованог поновка. Израчунавањем поновака и растојања су издвојени следећи потписи секвенци. Додатно су наглашени елементи који се јављају у бар два потписа секвенци.

S_1 (укупно 37 уређених парова): {<ATT, 38>; <ATT, 12>; <CCG, 30>; <CCG, 57>; <CGCGC, 25>; <CGCGC, 2>; <CGCGC, 25>; <CGCGC, 52>; <CGCGC, 50>; **<CGC, 2>**; <CGC, 29>; <CGC, 27>; <CGC, 54>; **<CGC, 23>**; <CGC, 4>; **<CGC, 23>**; **<CGC, 2>**; **<CGC, 48>**; <CTC, 45>; <CTC, 40>; <GAT, 18>; <GAT, 6>; <GCG, 14>; <GCG, 39>; <GCG, 13>; <GCG, 11>; <GCT, 14>; <GCT, 9>; <GCT, 28>; <GCT, 23>; <GCT, 6>; <GCT, 55>; <TCCGCGC, 27>; <TCC, 12>; <TCC, 39>; <TCG, 8>; <TCG, 13>;}

S_2 (укупно 27 уређених парова): {<ATG, 35>; <ATG, 50>; <ATG, 15>; <CAT, 37>; <CAT, 51>; <CAT, 14>; <CCG, 27>; <CCG, 46>; <CCG, 51>; <CGCGC, 22>; **<CGCGC, 41>**; <CGCG, 17>; **<CGCG, 2>**; **<CGCG, 39>**; <CGC, 26>; **<CGC, 48>**; **<CGC, 7>**; **<CGC, 46>**; **<CGC, 2>**; **<CGC, 21>**; <CGC, 24>; **<CGC, 43>**; <CGC, 20>; **<CGC, 2>**; <TCCGCGC, 19>; **<TCCGC, 5>**; <TCCGC, 24>;}

S_3 (укупно 26 уређених парова): {<CCCCC, 1>; **<CGCGC, 41>**; **<CGCG, 2>**; **<CGCG, 39>**; <CGC, 64>; **<CGC, 7>**; <CGC, 62>; **<CGC, 23>**; <CGC, 34>; <CGC, 69>; **<CGC, 7>**; <CGC, 12>; **<CGC, 46>**; **<CGC, 21>**; **<CGC, 43>**; <CGC, 2>; <GCATG, 42>; <GCGC, 5>; <GCGC, 36>; <TCA, 74>; <TCA, 49>; <TCA, 38>; <TCA, 25>; <TCA, 36>; <TCA, 11>; **<TCCGC, 5>**;}

На примеру потписа секвенце S_1 се примећује да елемент <ATT, 26> не припада истом, јер StatRepeats прогам не идентификује поновак ATT на позицијама 46 и 72 као максималан директан некомплементаран статистички значајан понављајући део секвенце, зато што је део дужег поновка ATTC који је максималан, али није и статистички значајан.

Корак 2: У овом кораку се израчунава сличност између дате три секвенце (за парове секвенци S_1 и S_2, S_1 и S_3 , S_2 и S_3) на основу њихових потписа, тачније идентификованих парова $<r, d_r>$ (Табела 7 и Табела 8). На основу добијених резултата секвенце S_2 и S_3 су међусобно сличније (J=0.21429) него било која од њих упарена са секвенцом S_1.

Табела 7. Скуп парова <r, d_r> неопходних за израчунавање Жакардовог растојања код парова секвенци S_1 и S_2, S_1 и S_3 , S_2 и S_3

Скуп парова	**S_1 - S_2**	**S_1 - S_3**	**S_2 - S_3**
M_{11}	**1. <CGC, 2>; 2. <CGC, 48>;**	**1. <CGC, 2>; 2. <CGC, 23>;**	**1. <CGC, 2>; 2. <CGC, 46>; 3. <CGC, 43>; 4. <CGC, 21>; 5. <CGC, 7>; 6. <CGCG, 2>; 7. <CGCG, 39>; 8. <CGCGC, 41>; 9. <TCCGC, 5>;**
M_{10}	1. <ATT, 12>; 2. <ATT, 38>; 3. <CCG, 30>; 4. <CCG, 57>; 5. <CGC, 23>; 6. <CGC, 27>; 7. <CGC, 29>; 8. <CGC, 4>; 9. <CGC, 54>; 10. <CGCGC, 2>; 11. <CGCGC, 25>; 12. <CGCGC, 50>; 13. <CGCGC, 52>; 14. <CTC, 40>; 15. <CTC, 45>; 16. <GAT, 18>; 17. <GAT, 6>; 18. <GCG, 11>; 19. <GCG, 13>; 20. <GCG, 14>; 21. <GCG, 39>; 22. <GCT, 14>; 23. <GCT, 23>; 24. <GCT, 28>; 25. <GCT, 55>; 26. <GCT, 6>; 27. <GCT, 9>; 28. <TCC, 12>; 29. <TCC, 39>; 30. <TCCGCGC, 27>; 31. <TCG, 13>; 32. <TCG, 8>;	1. <ATT, 12>; 2. <ATT, 38>; 3. <CCG, 30>; 4. <CCG, 57>; 5. <CGC, 27>; 6. <CGC, 29>; 7. <CGC, 4>; 8. <CGC, 48>; 9. <CGC, 54>; 10. <CGCGC, 2>; 11. <CGCGC, 25>; 12. <CGCGC, 50>; 13. <CGCGC, 52>; 14. <CTC, 40>; 15. <CTC, 45>; 16. <GAT, 18>; 17. <GAT, 6>; 18. <GCG, 11>; 19. <GCG, 13>; 20. <GCG, 14>; 21. <GCG, 39>; 22. <GCT, 14>; 23. <GCT, 23>; 24. <GCT, 28>; 25. <GCT, 55>; 26. <GCT, 6>; 27. <GCT, 9>; 28. <TCC, 12>; 29. <TCC, 39>; 30. <TCCGCGC, 27>; 31. <TCG, 13>; 32. <TCG, 8>;	1. <ATG, 15>; 2. <ATG, 35>; 3. <ATG, 50>; 4. <CAT, 14>; 5. <CAT, 37>; 6. <CAT, 51>; 7. <CCG, 27>; 8. <CCG, 46>; 9. <CCG, 51>; 10. <CGC, 20>; 11. <CGC, 24>; 12. <CGC, 26>; 13. <CGC, 48>; 14. <CGCG, 17>; 15. <CGCGC, 22>; 16. <TCCGC, 24>; 17. <TCCGCGC, 19>;
M_{01}	1. <ATG, 15>; 2. <ATG, 35>; 3. <ATG, 50>; 4. <CAT, 14>; 5. <CAT, 37>; 6. <CAT, 51>; 7. <CCG, 27>; 8. <CCG, 46>; 9. <CCG, 51>; 10. <CGC, 20>; 11. <CGC, 21>; 12. <CGC, 24>; 13. <CGC, 26>; 14. <CGC, 43>; 15. <CGC, 46>; 16. <CGC, 7>; 17. <CGCG, 17>; 18. <CGCG, 2>; 19. <CGCG, 39>; 20. <CGCGC, 22>; 21. <CGCGC, 41>; 22. <TCCGC, 24>; 23. <TCCGC, 5>; 24. <TCCGCGC, 19>;	1. <CCCCC, 1>; 2. <CGC, 12>; 3. <CGC, 21>; 4. <CGC, 34>; 5. <CGC, 43>; 6. <CGC, 46>; 7. <CGC, 62>; 8. <CGC, 64>; 9. <CGC, 69>; 10. <CGC, 7>; 11. <CGCG, 2>; 12. <CGCG, 39>; 13. <CGCGC, 41>; 14. <GCATG, 42>; 15. <GCGC, 36>; 16. <GCGC, 5>; 17. <TCA, 11>; 18. <TCA, 25>; 19. <TCA, 36>; 20. <TCA, 38>; 21. <TCA, 49>; 22. <TCA, 74>; 23. <TCCGC, 5>;	1. <CCCCC, 1>; 2. <CGC, 12>; 3. <CGC, 23>; 4. <CGC, 34>; 5. <CGC, 62>; 6. <CGC, 64>; 7. <CGC, 69>; 8. <GCATG, 42>; 9. <GCGC, 36>; 10. <GCGC, 5>; 11. <TCA, 11>; 12. <TCA, 25>; 13. <TCA, 36>; 14. <TCA, 38>; 15. <TCA, 49>; 16. <TCA, 74>;

M_{11}- скуп парова <r, d_r> који припадају обема секвенцама; M_{10}- скуп парова <r, d_r> који припадају првој секвенци и не припадају другој секвенци; M_{01}- скуп парова <r, d_r> који припадају другој секвенци и не припадају првој секвенци;

Табела 8. Број парова <r, d_r> неопходних за израчунавање мере сличности (Жакардово растојање), као и израчуната вредност сличности за секвенце из примера

	S_1 - S_2	**S_1 - S_3**	**S_2 - S_3**
$\|M_{11}\|$	2	2	9
$\|M_{10}\|$	32	32	17
$\|M_{01}\|$	24	23	16
J	**0.03448**	**0.03509**	**0.21429**

$|M_{11}|$- број елемената скупа парова <*r*, d_r> који припадају обема секвенцама; $|M_{10}|$- број елемената скупа парова <*r*, d_r> који припадају првој секвенци и не припадају другој секвенци; $|M_{01}|$- број елемената скупа парова <*r*, d_r> који припадају другој секвенци и не припадају првој секвенци;

3.2.3. Опис методе *Класификација секвенци заснована на профилима категорија*

Профил категорије *C* која садржи секвенце $S_c = \{s_1, s_2, \ldots, s_k\}$ се формира на основу скупа потписа секвенци $SS_c = \{S_1, S_2, \ldots, S_k\}$ дате категорије *C*. Скуп потписа секвенци SS_c се формира пратећи кораке описане функције *CreateSequenceSignature*.

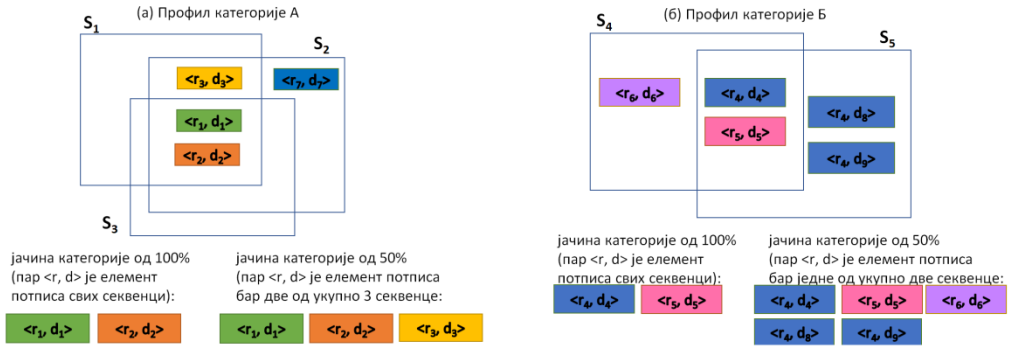
Први корак у израчунавању профила категорије *C* је идентификовање свих уређених парова <*r*, d_r> који су присутни у одређеном броју потписа секвенци које припадају категорији *C*. Означимо са *X* однос броја различитих потписа секвенци у којима се појављује уређени пар <*r*, d_r> и укупног броја секвенци које припадају категорији *C*. Профил категорије се конструише узимајући у обзир све уређене парове <*r*, d_r> за које је вредност *X* већа од одређене границе. Означимо са $C_{X\%}$ јачину категорије *C*, за границу *X*.

За сваку категорију из скупа $C=\{c_1, c_2, \ldots, c_m\}$ се примењује функција *CreateCategoryProfile* описана доле ради израчунавања профила категорије (Слика 10).

```
function CreateCategoryProfile (taxc, SS, CX)
/*taxc - таксономска категорија, SS - скуп потписа секвенци, CX - јачина профила категорије */
begin
израчунати distinct_ordered_pairs скуп различитих уређених парова <r, dr> из скупа SS;
for each <r, dr> in distinct_ordered_pairs
    begin
    if percent of <r, dr> ≥ CX
    then if (ако постоји неки други профил категорије у бази података такав да садржи исте ordered_pair и metadata параметре)
    then <r, dr> и metadata се искључују из профила ове категорије;
        else do_nothing;
    else додати ordered_pair <r, dr> и metadata у профил категорије taxc;
    end;
унети профил категорије taxc у базу података;
return taxc;
end
```

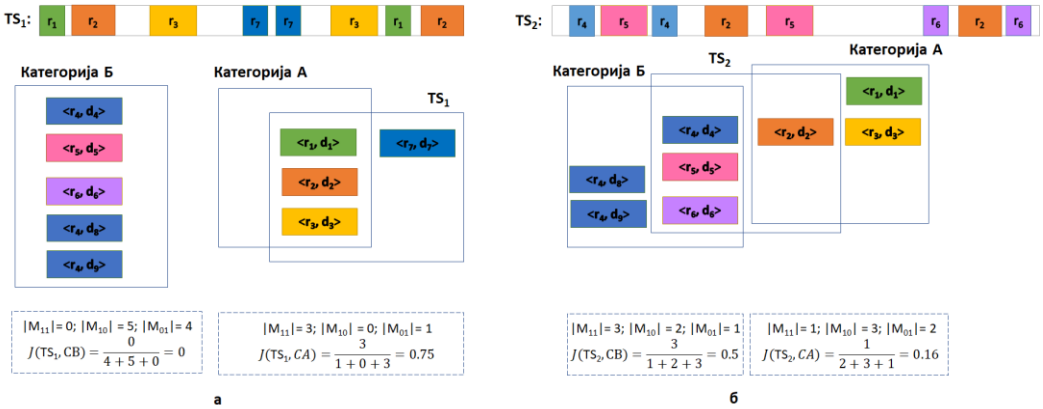
Слика 10. Опис корака функције CreateSequenceSignature за формирање профила категорије

Сви профили категорија се чувају у бази података. Профили категорија садрже различите уређене парове <*r*, d_r>, при чему број појављивања уређеног пара није укључен у даљу анализу. Ако две различите категорије садрже исти пар <*r*, d_r>, дати пар <*r*, d_r> се искључује из профила обе категорије. Пример формирања профила категорија је приказан на слици 11.



Слика 11. Пример формирања профила категорије за одређене вредности јачине категорије ($C_{100\%}$ и $C_{50\%}$ у примеру)
(а) Профил категорије А која садржи 3 секвенце: S_1, S_2 и S_3. Први пример (а) лево за вредност јачине категорије $C_{100\%}$ значи да уређени пар $<r, d_r>$ припада профилу категорије ако припада свим секвенцама у датој категорији. Два различита уређена пара испуњавају овај услов (на слици су означени наранџастом и зеленом бојом: {$<r_1, d_1>$ и $<r_2,d_2>$}). У другом примеру (а) десно је презентација јачине категорије $C_{50\%}$, за коју уређени пар $<r, d_r>$ припада профилу категорије ако је елемент потписа бар две од 3 секвенце у датој категорији. Три различита уређена пара испуњавају овај услов (означени наранџастом, жутом и зеленом бојом: {$<r_1, d_1>$; $<r_2,d_2>$ и $<r_3, d_3>$}).
(б) Профил категорије Б која садржи 2 секвенце: S_4 и S_5. Први пример (б) лево за вредност јачине категорије $C_{100\%}$ значи да уређени пар $<r, d_r>$ припада профилу категорије ако припада свим секвенцама у датој категорији. Два различита уређена пара испуњавају овај услов (на слици су означени са плавом и розе бојом: {$<r_4, d_4>$ и $<r_5,d_5>$}). У другом примеру (б) десно је презентација јачине категорије $C_{50\%}$, за коју уређени пар $<r, d_r>$ припада профилу категорије ако је елемент потписа бар једне од две секвенце у датој категорији. Пет различитих уређених парова испуњавају овај услов (на слици су означени са плавом, љубичастом и розе бојом: {$<r_4, d_4>$; $<r_5,d_5>$; $<r_6,d_6>$; $<r_4,d_8>$; $<r_4,d_9>$}).

Категорија непознате секвенце се одређује поређењем потписа секвенце са профилима категорија у бази података. Сличност између потписа секвенце и профила категорије се израчунава користећи једну од мера сличности описаних у поглављу 2.4.3 (Жакардово растојање, косинусно растојање, или коефицијент Танимотоа). Поједностављен пример је приказан на слици 12.



Слика 12. Пример поређења потписа тестних секвенци TS_1 (а) и TS_2 (б) са профилима категорије А и В, за изабрану вредност јачине категорије $C_{50\%}$ и Жакардове мере као мере сличности. Три уређена пара $<r, d_r>$ припадају потпису секвенце TS_1 и профилу категорије А, али нема заједничких елемената између елемената потписа секвенце TS_1 и профила категорије Б (а). Применом Жакардове мере сличности ($J = \frac{|M_{11}|}{|M_{01}|+|M_{10}|+|M_{11}|}$) на потписе секвенци и профиле категорија, закључује се да је секвенца TS_1 више слична категорији А ($J(TS_1,CA)=0.75$) него категорији Б ($J(TS_1,CB)=0$). Три уређена пара $<r, d_r>$ припадају потпису секвенце TS_2 и профилу категорије Б, такође један уређени пар $<r, d_r>$ припада потпису секвенце TS_2 и категорије А (б). Применом Жакардове мере сличности, закључује се да је секвенца TS_2 више слична категорији Б ($J(TS_2,CB)=0.5$) него категорији А ($J(TS_2,CA)=0.16$).

3.2.4. Пример формирања профила категорија

Формирање профила категорије биће објашњено на примеру три улазне секвенце које припадају категорији C, и коришћене су у претходном примеру (израчунавање потписа секвенци - поглавље 3.2.2). Директан некомплементарни тип поновка је изабран у примеру као улазни податак методе. Вредност параметра минимална дужина поновка је једнака нули, што означава да се узимају у обзир сви поновци са најнижом могућом вредношћу дужине која задовољава постављене услове статистичке значајности.

Корак 1: Начин израчунавања потписа секвенци је идентичан као у првом примеру (поглавље 3.2.2).

S_1 (укупно 37 уређених парова): {<ATT, 38>; <ATT, 12>; <CCG, 30>; <CCG, 57>; <CGCGC, 25>; <CGCGC, 2>; <CGCGC, 25>; <CGCGC, 52>; <CGCGC, 50>; **<CGC, 2>**; <CGC, 29>; <CGC, 27>; <CGC, 54>; <CGC, 23>; <CGC, 4>; <CGC, 23>; **<CGC, 2>**; <CGC, 48>; <CTC, 45>; <CTC, 40>; <GAT, 18>; <GAT, 6>; <GCG, 14>; <GCG, 39>; <GCG, 13>; <GCG, 11>; <GCT, 14>; <GCT, 9>; <GCT, 28>; <GCT, 23>; <GCT, 6>; <GCT, 55>; <TCCGCGC, 27>; <TCC, 12>; <TCC, 39>; <TCG, 8>; <TCG, 13>;}

S_2 (укупно 27 уређених парова): {<ATG, 35>; <ATG, 50>; <ATG, 15>; <CAT, 37>; <CAT, 51>; <CAT, 14>; <CCG, 27>; <CCG, 46>; <CCG, 51>; <CGCGC, 22>; <CGCGC, 41>; <CGCG, 17>; <CGCG, 2>; <CGCG, 39>; <CGC, 26>; <CGC, 48>; <CGC, 7>; <CGC, 46>; **<CGC, 2>**; <CGC, 21>; <CGC, 24>; <CGC, 43>; <CGC, 20>; **<CGC, 2>**; <TCCGCGC, 19>; <TCCGC, 5>; <TCCGC, 24>;}

S_3 (укупно 26 уређених парова): {<CCCCC, 1>; <CGCGC, 41>; <CGCG, 2>; <CGCG, 39>; <CGC, 64>; <CGC, 7>; <CGC, 62>; <CGC, 23>; <CGC, 34>; <CGC, 69>; <CGC, 7>; <CGC, 12>; <CGC, 46>; <CGC, 21>; <CGC, 43>; **<CGC, 2>**; <GCATG, 42>; <GCGC, 5>; <GCGC, 36>; <TCA, 74>; <TCA, 49>; <TCA, 38>; <TCA, 25>; <TCA, 36>; <TCA, 11>; <TCCGC, 5>;}

Корак 2: У овом кораку се израчунава профил категорије C на основу потписа секвенци које припадају овој категорији као и вредности јачине категорије.

Избор вредности јачине категорије $C_{100\%}$, означава да профил категорије треба да садржи парове $<r, d_r>$ који припадају свим потписима секвенци s_1, s_2 и s_3 из примера који дефинишу категорију C. У датом примеру постоји само један овакав пар <CGC, 2> (парови који испуњавају услов су подебљани у тексту).
На основу претходног, профил категорије $C_{100\%}$ је скуп {<CGC, 2>}.

Избор вредности јачине категорије $C_{50\%}$, означава да профил категорије треба да садржи парове $<r, d_r>$ који припадају потписима бар две секвенце из примера, тачније скупа секвенци s_1, s_2 и s_3 које дефинишу категорију C. У датом примеру следећи парови испуњавају дати услов (парови су подвучени у примеру): {<CGC, 2>; <CGC, 23>; <CGC, 48>; <CGC, 7>; <CGC, 21>; <CGC, 43>; <CGC, 46>; <CGCG, 2>;<CGCG, 39>; <CGCGC; 41>; <TCCGC; 5>}. Дати скуп парова представља профил категорије C за вредност параметра јачина категорије $C_{50\%}$.

3.2.5. Имплементација и прикупљање података

За израчунавање различитих типова поновака коришћен је већ помињани програм StatRepeats (прво издање, верзија 2) [5] [6]. Као улазни параметар, коришћена је подразумевана p=0.05 вредност. Сви описани метаподаци су чувани у IBM Db2 релационој бази података [65], која се користи као репозиторијум података који се користи током фаза припреме, обраде и анализе података коришћењем SQL упита.

Метода је тестирана на основу чланова таксономских категорија различитог нивоа, као што су класа, породица, род или врста. Секвенце су преузете из ICTV базе података [17]. Недостајуће информације о таксономијама су преузете из NCBI базе са таксономским подацима [77].

3.3. Упоредни преглед и анализа метода

Добро је познат изазов издвајања атрибута биолошких секвенци због њихове дужине, комплексности и различитости. Предмет ове дисертације је анализа биолошких (нуклеотидних и аминокиселинских) секвенци и њихових статистички значајних поновака различитих типова и дужина у циљу развоја нових модела за одређивање сличности секвенци на основу идентификованих поновака. Опште правило код нуклеотидних секвенци је то да што је подсеквенца краћа (у нашем случају поновак), то је већа вероватноћа да се појављује насумично у низу. Избором само статистички значајних понављајућих секвенци, овај проблем се успешно превазилази.

У складу са тим су развијена два нова и корисна приступа за представљање биолошких секвенци на основу особина поновака, у облику који је флексибилан и погодан за примену у даљим анализама. Први приступ трансформисања биолошке секвенце у векторски облик, који представља прву фазу R-P/F методе, се заснива на концепту теорије информација укључујући позицију, поредак и учесталост појављивања поновака. Други приступ се заснива на скуповима уређених парова формираних на основу пара поновака и растојања између истих. На овај начин се граде потписи секвенци.

Прва разлика између ова два приступа је у начину формирања вектора и подацима коришћеним за формирање вектора. У првом приступу, улазни подаци неопходни за формирање вектора су поновак, позиције на којима се поновак појављује укључујући и његов поредак, укупан број појављивања тог поновка (који је најчешће већи од два), растојања између најближих суседа и дужина секвенце. Иако је први развијени приступ дао задовољавајуће резултате, било је неопходно одредити додатно да ли са мањом количином информација (карактеристичних тачака) могу да се добију довољно прецизни резултати. Стога су уместо поновака узети у обзир парови поновака, као елементи карактеристичних тачака, и додатно је укључено растојање између парова поновака у анализу, како би смањили зависност анализе од могућих промена које су се десиле током еволуције. Такође, сама процедура трансформације секвенци у векторе је једноставнија у другом приступу.

Друга фаза анализе је усмерена на одређивање сличности између секвенци користећи презентације секвенци у векторском простору и примену метода истрживања података ради даље анализе и провере. У случају R-P/F методе је примењено косинусно растојање у анализи сличности, као и хијерархијско кластеровање. У случају методе формирања потписа секвенци и профила таксономских група је примењено косинусно растојање, Жакардова коефицијент и коефицијент Танимото-а за одређивање сличности секвенци и методе класификације за одређивање таксономских група.

Показано је да оба модела прецизно функционишу на статистички значајним поновцима различитих особина (типа и дужина). Представљене методе се категоришу као методе без поравнања од којих је прва заснована на теорији информација, а друга се категорише као класификација секвенци заснована на атрибутима.

4. Резултати и дискусија

У оквиру ове главе биће приказани резултати новоразвијених метода. Прво су приказани резултати тестирања R-P/F методе (поглавље 4.1). Затим су приказани резултати методе одређивање сличних секвенци, поређењем потписа секвенци (поглавље 4.2.1) и на крају су приказани резултати класификације секвенци засноване на профилима категорија (поглавље 4.2.2). Методе су проверене на различитим скуповима референтних нуклеотидних и протеинских секвенци.

У циљу провере изводљивости и предности нових модела, извршено је више тестова на скуповима биолошких секвенци преузетих из актуелних биолошких репозиторијума. Сумирани преглед скупова података над којима су тестиране методе је дат у табели 9. Резултати предложених метода су упоређени са резултатима најчешће коришћених метода за поређење секвенци са поравнањем и без поравнања, и оцењени мерама за евалуацију резултата кластеровања и класификације. Установљено је да се са предложеним новим методама, заснованим на статистички значајним поновцима, постижу упоредни резултати са најпопуларнијим алгоритмима за анализу сличности секвенци на скуповима тестираних примера. Сумирана методологија (садржај/упутство за читање резултата) је дата у табели 10.

Табела 9. Сумирани преглед скупова података за тестирање метода

Скуп података	Тип секвенце	Бр. секвенци	Дужина секвенци	Број различитих поновака/парова поновака				Метода	Извор скупа података
				DN	DC	IN	IC		
Митохондријална ДНК различитих врста сисара	нуклеотидна	46	16295 - 18642	53.504	12.848	33.419	15.258	R-P/F метода	NCBI [13]
РНК секвенце вируса еболе, марбург вируса, и бетакоронавируса	нуклеотидна	381	18794 - 30309	85.305	32.894	53.831	38.385	R-P/F метода	ViPR база података [15]
NADH дехидрогеназа субјединица 5 (MT-ND5)	протеинска	9	602-604	311	N/A	275	N/A	R-P/F метода	Секвенце преузете из NCBI базе података [13], идентификатори секвенци из [78]
NADH дехидрогеназа субјединица 6 (MT-ND6)	протеинска	8	167-172	69	N/A	94	N/A	R-P/F метода	
Ксиланаза	протеинска	20	210-291	267	N/A	244	N/A	R-P/F метода	
Трансферин	протеинска	24	677-697	630	N/A	244	N/A	R-P/F метода	
Spike протеин корона вируса	протеинска	50	1162-1283	1.191	N/A	791	N/A	R-P/F метода	
Бета-глобин	протеинска	50	141-146	106	N/A	92	N/A	R-P/F метода	
Нуклеотидне целе секвенце вируса	нуклеотидна	4231	229-196858	962.880	674.214	744.198	890.201	Потписи секвенци и профили категорија	ICTV - Virus Metadata repository (VMR) [17]

Табела 10. Сумирани преглед коришћених метода за различите скупове података

Скуп података	Примењена метода	Методологија Примењена мера сличности Методе истраживања података
Митохондријална ДНК различитих врста сисара	R-P/F метода	**Методологија** • Косинусно растојање • Хопкинсова статистика • Хијерархијско кластеровање **Поређење резултата за различите параметре модела** • Кофенетски коефицијент корелације за различите типове поновака **Поређење са другим методама** • Clustal Omega – кофенетски коефицијент • BLAST – Спирманов коефицијент • Методе без поравнања – Спирманов коефицијент

РНК секвенце вируса еболе, марбург вируса, као и бетакоронавируса	R-P/F метода	**Методологија** • Косинусно растојање • ПЦА димензиона редукција • Хопкинсова статистика • Хијерархијско кластеровање **Поређење резултата за различите параметре модела** • Мера спољашње провере коректности кластеровања **Спољашња провера коректности кластеровања** Мере: прецизност, одзив, Ф-мера, ентропија, чистоћа
Секвенце протеина (6 различитих скупова)	R-P/F метода	**Методологија** • Косинусно растојање • Хопкинсова статистика • Хијерархијско кластеровање **Поређење са другим методама** • Clustal Omega – кофенетски коефицијент корелације • Clustal Omega – РФ растојање
Нуклеотидне секвенце вируса	1:1 поређење потписа секвенци	**Методологија** • Косинусно растојање • Жакардов коефицијент • Коефицијент Танимотоа **Поређење резултата за различите параметре модела** • Коефицијент корелације за различите типове поновака и мере сличности **Поређење са другим методама** • BLAST – коефицијент корелације **Примена на целим и парцијалним секвенцама**
Нуклеотидне секвенце вируса	Класификација секвенци заснована на профилима категорија	**Методологија** • Косинусно растојање • Коефицијент Танимотоа • Класификација

4.1 Метода заснована на позицији и локалној учесталости поновака

R-P/F метода заснована на позицији и локалној учесталости поновака је тестирана на два различита скупа нуклеотидних секвенци: 1) митохондријална ДНК различитих врста сисара и 2) секвенце РНК вируса еболе, марбург вируса и бетакоронавируса. Такође је ова метода примењена на шест скупова референтних протеинских секвенци: 1) секвенце протеина MT-ND5, 2) секвенце протеина MT-ND6, 3) секвенце протеина ксиланазе, 4) секвенце протеина трансферина, 5) секвенце *Spike* протеина корона вируса и 6) секвенце протеина бета-глобина. Добијени резултати су упоређени са резултатима BLAST и Clustal Omega поравнања над истим скупом секвенци, као и различитим добро познатим алгоритмима поређења секвенци без поравнања заснованим на к-торкама користећи растојања: Еуклидско растојање, растојање Менхетн (енг. *Manhattan*), д2 мера, д2* мера и косинусна сличност. Метода је оцењена мерама за процену квалитета резултата кластеровања: Хопкинсова статистика, тачност, одзив, РФ растојање и кофенетски коефицијент корелације.

4.1.1. Тестирање *R-P/F* методе на скупу секвенци митохондријалне ДНК различитих врста сисара

Први скуп садржи 45 секвенци митохондријалне ДНК сисара (Табела додатак 6.1.1). Скуп је допуњен митохондријалном секвенцом рибе листа (Crested flounder), који представља елемент ван граница додат са намером додатне провере. Ове секвенце су преузете са NCBI платформе [13] у FASTA формату. Дужина анализираних нуклеотидних секвенци је између 16295 и 18642 нуклеотида.

Приликом идентификовања поновака, коришћена је уобичајна *p* вредност (p=0.05) за одређивање статистичке значајности поновка. Минимална дужина статистички значајних поновака у овом скупу података је 5 нуклеотида. Максимална дужина поновака, као и број парова поновака у датим секвенцама по типу поновка су приказани у табели 11. Највећи број статистички значајних поновака у овом скупу података је дужине 8 и 9 нуклеотида за сва 4 типа поновака (Слика 13).

Табела 11. Особине поновака за скуп података митохондријалне ДНК сисара

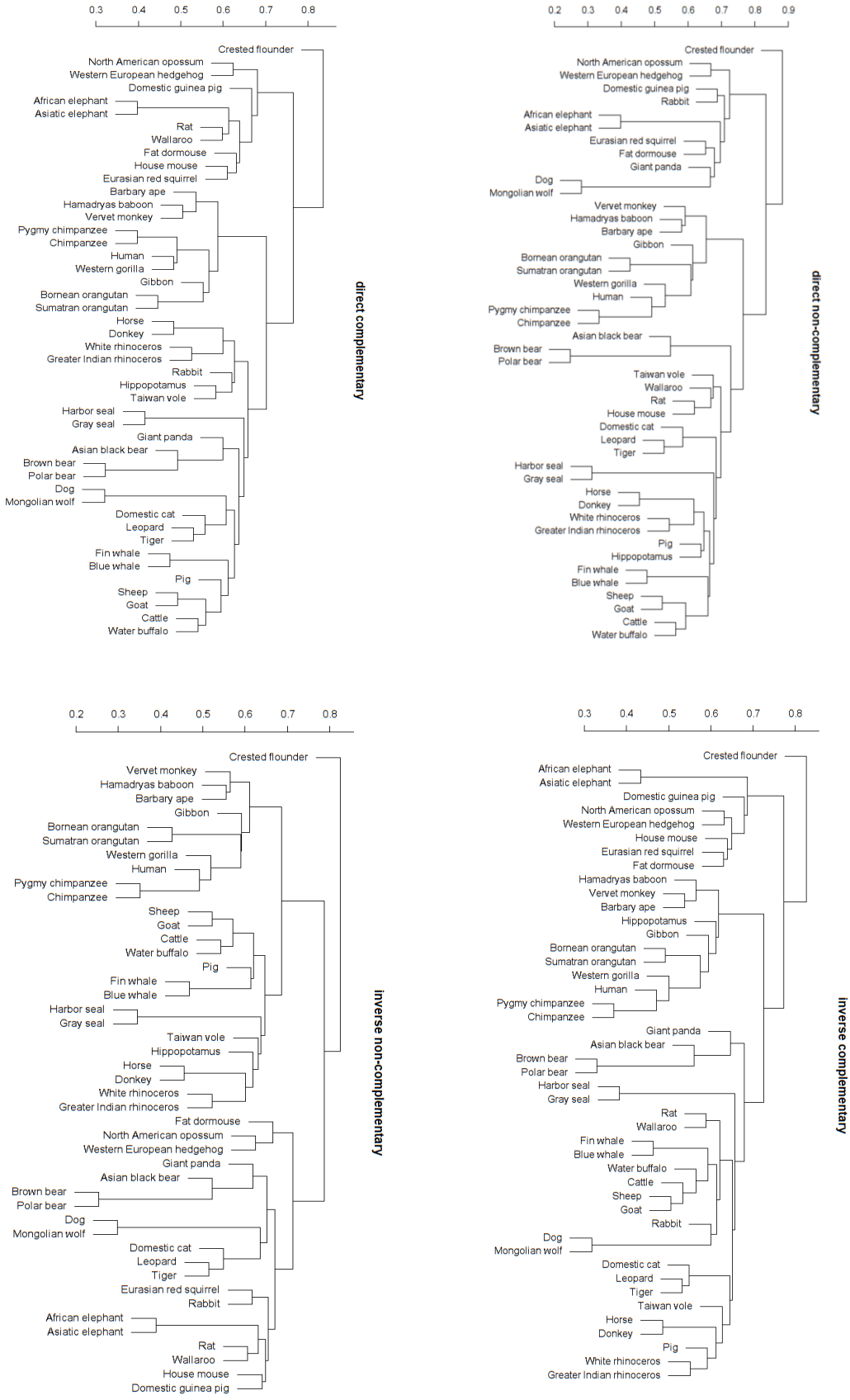
Тип поновка	**Број секвенци**	**Минимална дужина поновака**	**Максимална дужина поновака**	**Број поновака (димензија векторског простора)**
DN	46	5	784	53.504
DC	46	5	22	12.846
IN	46	5	291	33.413
IC	46	5	24	15.258

Слика 13. Расподела дужине поновака за скуп података митохондријалне ДНК сисара

Анализа је поновљена за све четири врсте поновака. За сваку секвенцу и поновак у бази, израчуната је вредност ентропије локалне учесталости поновка на основу описане методе. Секвенце су представљене векторима у вишедимензионом векторском простору. Димензије векторских простора су представљене у табели 11. Свака компонента вектора истиче удео одређеног поновка у секвенци.

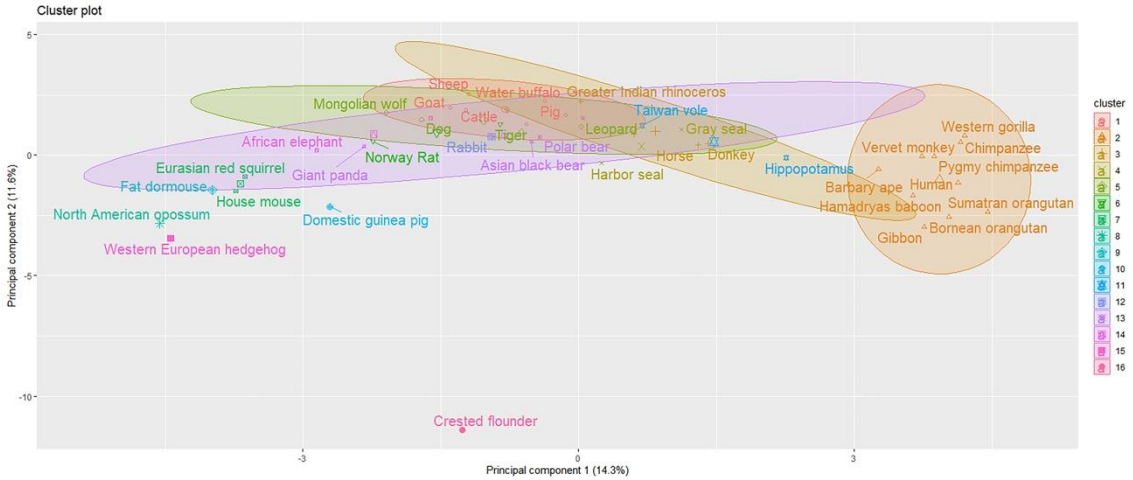
Израчуната је вредност Хопкинсове статистике за сва четири типа поновака и она износи 0.6634177 за DN тип поновака, 0.6747377 за DC тип поновака, за 0.6537529 IN тип поновака и 0.6600711 за IC тип поновака.

Матрица сличности између свих парова секвенци је израчуната користећи косинусну сличност. Примењено је хијерархијско кластеровање и добијени резултати су представљени дендрограмима (Слика 14).



Слика 14. Резултат примене R-P/F методе на скупу секвенци митохондријалне ДНК сисара приказан у облику дендрограма за 4 различита типа поновака

Добијени резултати су у складу са еволуционим односима датих врста. Груписање врста по таксономији је уочено за све четири врсте поновака. Намерно додавање елемента ван граница (риба лист *Crested flounder*) је омогућило додатну потврду исправности методе због јасног одвајања истог у резултатима. Риба лист (*Crested flounder*), једина митохондријална ДНК која не припада сисарима у анализираном скупу података је јасно издвојена на слици 13. Групе се такође могу уочити на графичком приказу резултата кластеровања (Слика 15) и на топлотној мапи (Слика додатак 6.1.1) за сва 4 типа поновака.



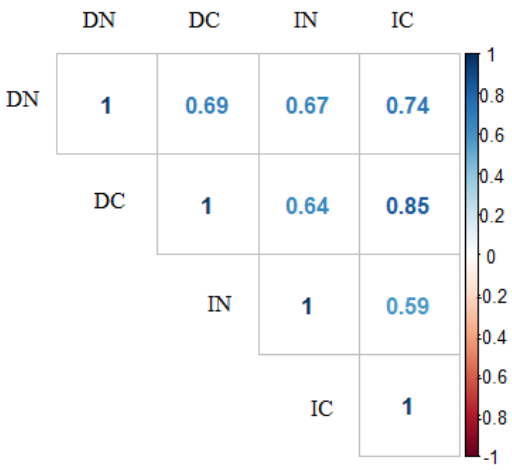
Слика 15. Тачкасти приказ резултата кластеровања 46 секвенци митохондријалне ДНК. Резултати су добијени на основу матрице сличности добијене као резултат R-P/F методе примењене на директне комплементарне поновке. Секвенце су приказане тачкама користећи прве две димензије анализе главних компоненти. Различитим бојама је означено којим кластерима припадају. Кластери су ограничени елипсама.

Пет парова врста са најсличнијим секвенцама митохондријалне ДНК је приказано у табели 12.

Табела 12. Вредности матрице сличности за пет парова насличнијих врста добијених као резултат Р-Ф/П методе примењене на секвенце митохондријалних ДНК сисара

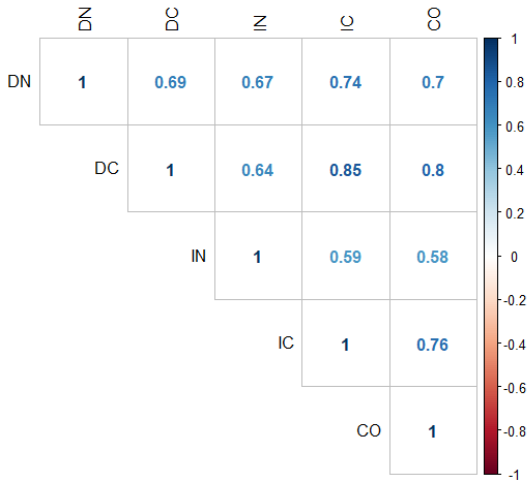
Секвенца/тип поновака		**DN**		**DC**		**IN**		**IC**	
Домаћин бр. 1	*Домаћин бр. 2*	*р. б.*	*сличност*	*р. б.*	*сличност*	*р. б.*	*сличност*	*р. б.*	*сличност*
Polar bear	**Brown bear**	1	0,753	2	0,677	1	0,746	2	0,671
Mongolian wolf	**Dog**	2	0,719	1	0,680	2	0,702	1	0,683
Gray seal	**Harbor seal**	3	0,688	5	0,585	3	0,655	4	0,616
Chimpanzee	**Pygmy chimpanzee**	4	0,666	3	0,602	4	0,648	3	0,630
Asiatic elephant	**African elephant**	5	0,603	4	0,602	5	0,611	5	0,567

Добијени дендрограми за различите типове поновака су упоређени међусобно, израчунавањем кофенетског коефицијента корелације. Показано је да су дендрограми развијени R-P/F методом користећи DC и IC, као и DN и IC типове поновака више корелисани, док су IN и IC најмање корелисани међусобно (Слика 16).



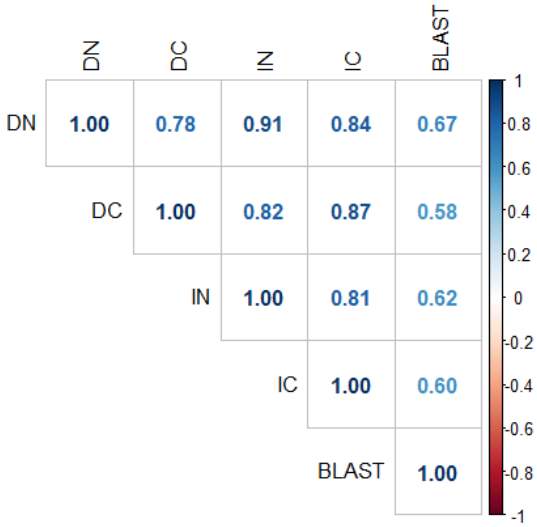
Слика 16. Матрица корелације између 4 дендрограма добијених применом R-P/F методе над истим скупом секвенци за различите типове поновака, користећи кофенетску методу за поређење дендрограма.

Ради поређења резултата добијених R-P/F методом са резултатима добијеним методом Clustal Omega, генерисано је филогенетско стабло (Слика додатак 6.1.2) користећи UPGMA алгоритам примењен на резултате вишеструког поравнања секвенци Clustal Omega методом. Израчуната је матрица корелације између филогенетског стабла добијеног Clustal Omega методом и 4 дендрограма добијених применом R-P/F методе над истим скупом секвенци за различите типове поновака (Слика 13), користећи кофенетску методу за поређење дендрограма (Слика 17). Резултирајући дендрограм добијен Clustal Omega методом је највише корелисан са дедрограмом добијеним као резултат R-P/F методе примењене на директне комплементарне поновке.

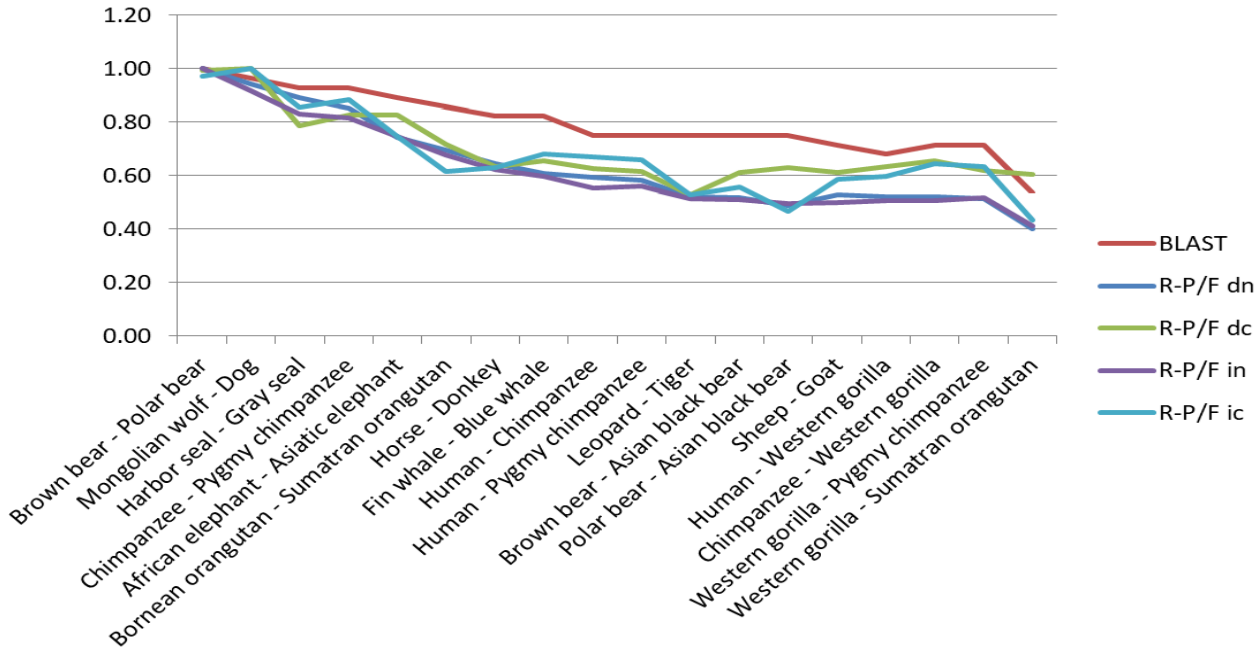


Слика 17. Матрица корелације између добијеног филогенетског стабла на основу резултата Clustal Omega методе и 4 дендрограма за различите типове поновака добијених применом R-P/F методе над истим скупом секвенци, користећи кофенетску методу за поређење дендрограма.

Одређивање сличности парова секвенци је извршено применом BLAST методе. Добијени резултати, тачније вредност идентичности секвенци (енг. *BLAST ident value*) су нормализовани и упоређени са резултатима R-P/F методе користећи Спирманов коефицијент корелације. Показано је да су резултати добијени BLAST методом најсличнији резултатима добијених применом R-P/F методе примењене на директне некомплементарне поновке (Слика 18 и 19).

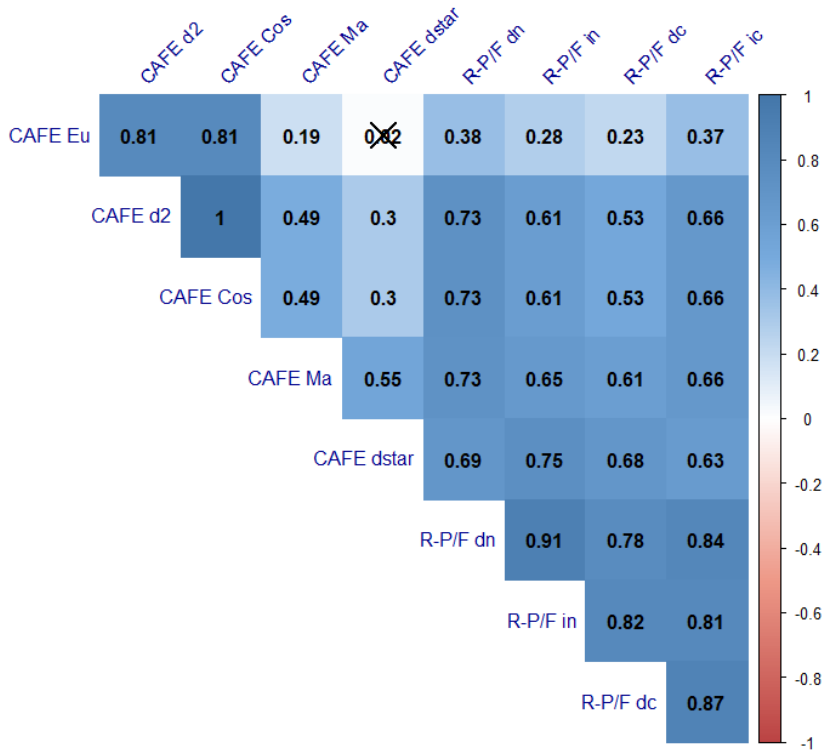


Слика 18. Матрица корелације (Спирманов коефицијент корелације) између нормализованих вредности идентичности секвенци добијених BLAST методом и елемената матрице сличности добијених применом R-P/F методе над истим скупом секвенци (митохондријалне ДНК сисара) за различите типове поновака.



Слика 19. Приказ поређења нормализованих резултата BLAST методе (вредност ident) и R-P/F методе над истим скупом секвенци (митохондријалне ДНК сисара) за различите типове поновака. Приказани су резултати који су имали високе вредности сличности за било коју примењену методу.

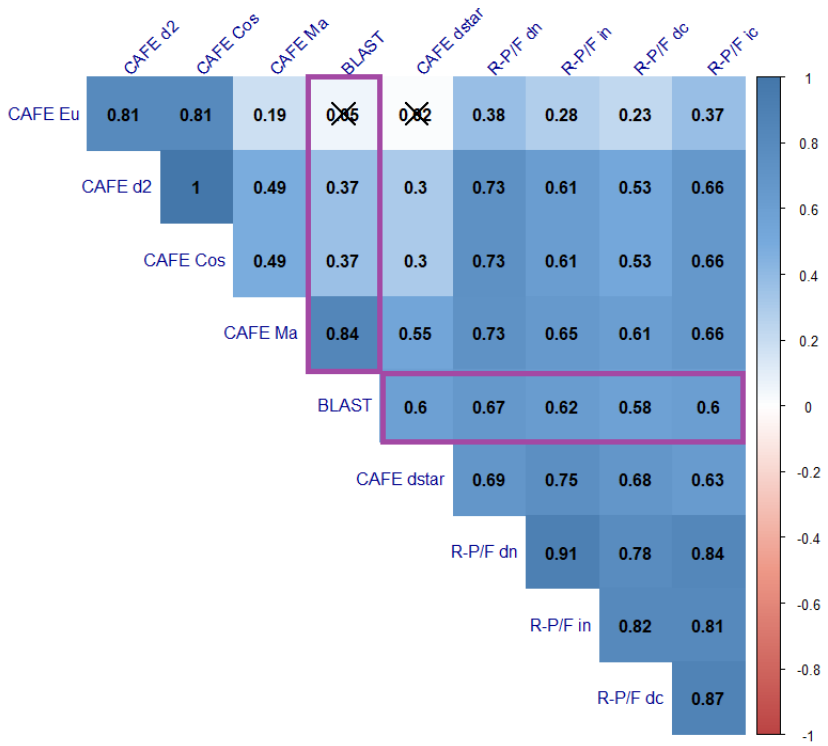
На истом скупу секвенци су примењени добро познати и најчешће коришћени алгоритми за поређење секвенци без поравнања засновани на к-торкама. Коришћен је програм aCcelerated Alignment-FrEe sequence analysis (CAFE) [79]. Мере сличности које су коришћене за поређење су: Еуклидско растојање (вредност параметра к=14), растојање Менхетн (вредност параметра к=14), д2 мера (вредност параметра к=8), д2* мера (вредност параметра к=8) и косинусна сличност (вредност параметра к=8). Добијене сличности између парова секвенци су упоређене са резултатима R-P/F методе за сва четири типа поновка на основу Спирмановог коефицијента корелације (Слика 20). Показано је да су резултати добијени R-P/F методом за директне некомплементарне поновке у већој корелацији са резултатима метода заснованим на к-торкама, у односу на остале типове поновака.



Слика 20. Спирманов коефицијент корелације добијен упоређивањем резултата методе без поравнања за различита растојања (Еуклидско растојање (вредност параметра к=14), растојање Менхетн (вредност параметра к=14), д2 мера (вредност параметра к=8), д2* мера (вредност параметра к=8) и косинусна сличност (вредност параметра к=8)) са резултатима R-P/F методе над секвенцама митохондријане ДНК сисара. Резултати (елементи матрице) који нису значајни на основу коришћене p вредности (p>0.01) су прецртани.

Показано је да су резултати R-P/F методе примењене на директне некомплементарне и инверзне некомплементарне поновке тачнији од резултата добијених методама без поравнања за Еуклидску, д2, д2* меру и косинусну сличност узимајући као стандард резултате BLAST методе. Такође су мање тачни у односу на Менхент растојање на тестираном скупу података (Слика 20).

Резултати R-P/F методе примењене на директне комплементарне и инверзне комплементарне поновке су тачнији од резултата добијених применом метода без поравнања за Еуклидску, д2 и косинусну сличност узимајући као стандард резултате BLAST методе. Такође су мање тачни у односу на Менхетн растојање и д2* меру на тестираном скупу података (Слика 21).



Слика 21. Спирманов коефицијент корелације добијен при поређењу резултата методе без поравнања за различита растојања (Еуклидско растојање (вредност параметра к=14), растојање Менхетн (вредност параметра к=14), д2 мера (вредност параметра к=8), д2* мера (вредност параметра к=8) и косинусна сличност (вредност параметра к=8)), нормализованих резултата BLAST методе и резултата R-P/F методе примењене на 46 секвенци митохондријане ДНК сисара. Резултати (елементи матрице) који нису значајни на основу коришћене p вредности (p>0.01) су прецртани.

4.1.2. Тестирање *R-P/F* методе на скупу нуклеотидних секвенци РНК вируса еболе, марбург вируса, и бетакоронавируса

Други скуп података садржи 381 нуклеотидну секвенцу изолата РНК вируса еболе, марбург, и бетакоронавируса. Скуп комплетних секвенци генома је преузет из ViPR базе података [15]. Информације о таксономији су преузете из NCBI базе података [13]. Секвенце вируса еболе и марбург вируса припадају истој таксономској фамилији *Filoviridae*, док секвенце изолата SARS вируса (из рода бетакоронавируса) припадају фамилији Coronavirinae (Табела 13). Дужина анализираних РНК секвенци вируса је између 18794 и 30309 нуклеотида.

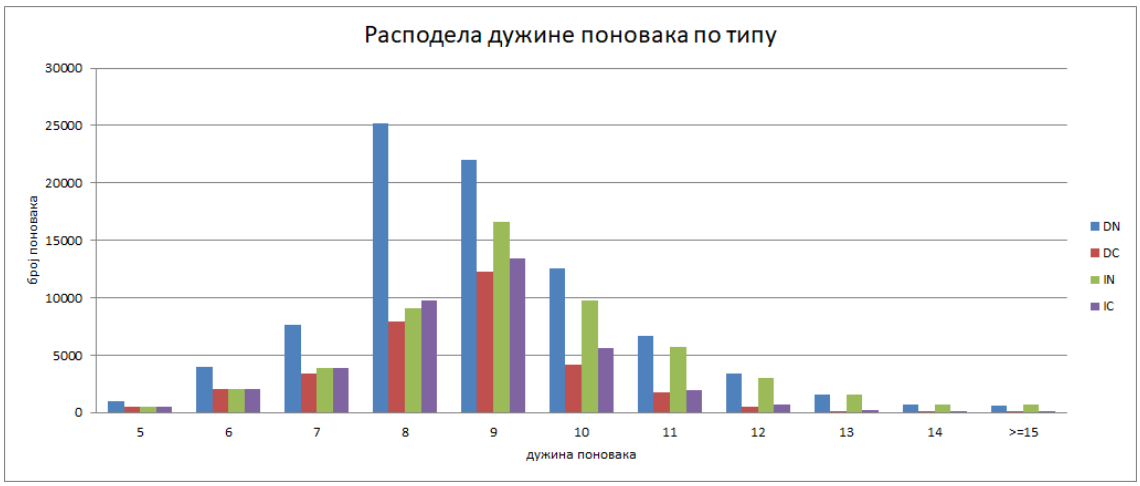
Приликом идентификовања поновака, коришћена је уобичајна *p* вредност (p=0.05) за одређивање статистичке значајности поновка. Минимална дужина статистички значајних поновака у овом скупу података је 5 нуклеотида. Максимална дужина поновака, као и број парова поновака у датим секвенцама по типу поновка су приказани у табели 14. Највећи број статистички значајних поновака у овом скупу података је дужине 8 и 9 нуклеотида за сва 4 типа поновака (Слика 22).

Табела 13. Таксономска подела скупа нуклеотидних секвенци изолата РНК вируса еболе, марбург, и бетакоронавируса

Таксономија	Број секвенци
Phylum: Negarnaviricota	**294**
Class: Monjiviricetes	**294**
Order: Mononegavirales	294
Family: Filoviridae	**294**
Genus: Ebolavirus	204
Species: Bundibugyo ebolavirus	9
Species: Ebola virus sp.	13
Species: Reston ebolavirus	18
Species: Sudan ebolavirus	17
Species: Tai Forest ebolavirus	4
Species: Zaire ebolavirus	143
Genus: Marburgvirus	90
Species: Marburg marburgvirus	90
Phylum: Pisuviricota	**87**
Class: Pisoniviricetes	**87**
Order: Nidovirales	87
Family: Coronaviridae	**87**
Genus: Betacoronavirus	87
Species: SARS-like coronavirus WIV16	1
Species: Severe acute respiratory syndrome-related coronavirus	86
Укупан број секвенци	**381**

Табела 14. Особине поновака за скуп нуклеотидних секвенци изолата РНК вируса еболе, марбург, и бетакоронавируса

Тип поновка	Број секвенци	Минимална дужина поновака	Максимална дужина поновака	Број поновака (димензија векторског простора)
DN	381	5	87	85.305
DC	381	5	78	32.894
IN	381	5	79	53.831
IC	381	5	81	38.385

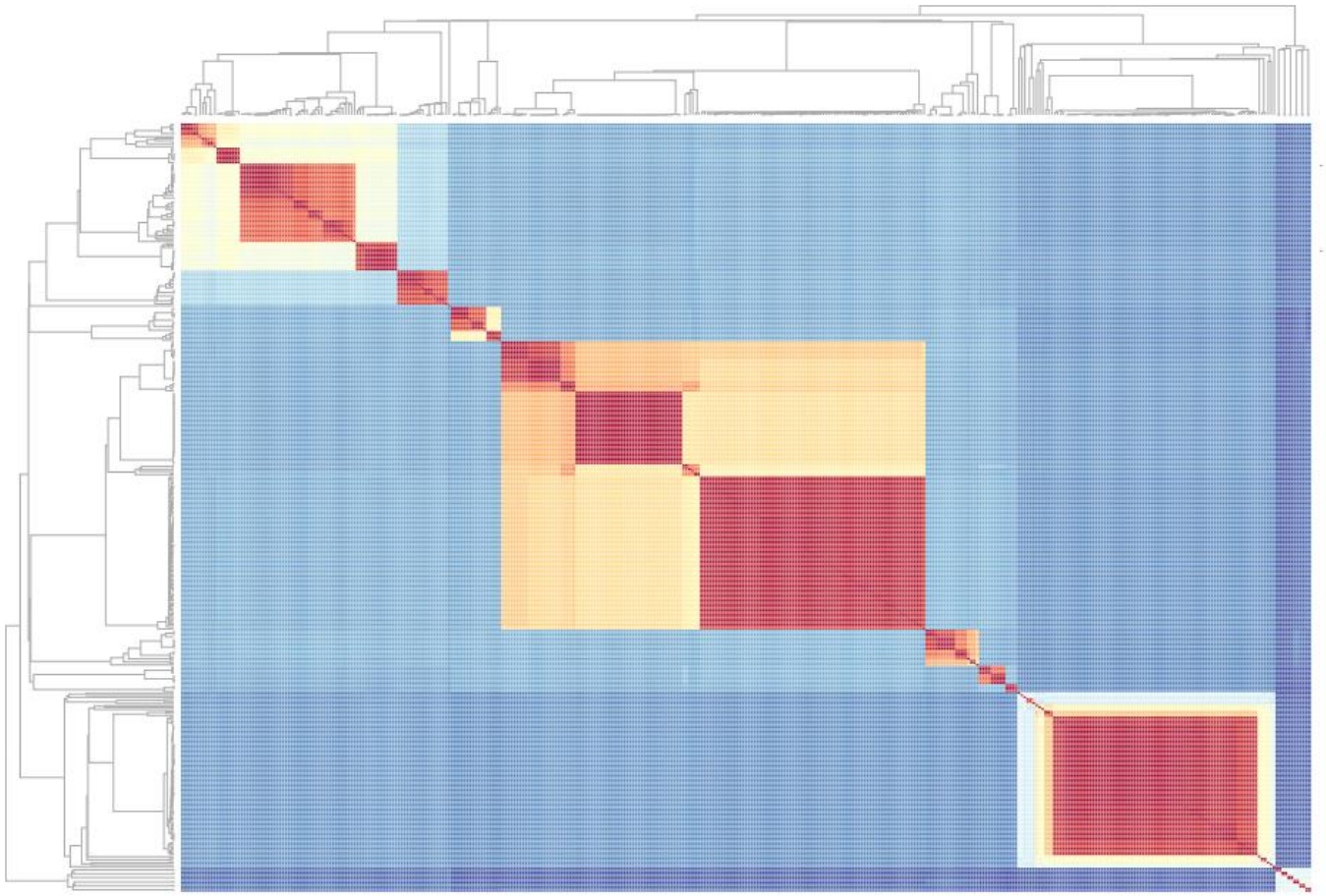


Слика 22. Расподела дужине поновака за скуп нуклеотидних секвенци изолата РНК вируса еболе, марбург, и бетакоронавируса

Анализа је поновљена за све четири врсте поновака. За сваку секвенцу и поновак у бази, израчуната је вредност ентропије локалне учесталости поновка на основу описане методе. Секвенце су представљене векторима у вишедимензионом векторском простору. Димензије векторских простора су представљене у табели 14. Свака компонента вектора истиче удео одређеног поновка у секвенци.

Израчуната је вредност Хопкинсове статистике за сва четири типа поновака на скупу података на коме је примењена метода анализе главних компоненти. За DN тип поновака узето је у обзир првих 54 димензија које покривају 95% материјала и добијена вредност Хопкинсове статистике износи 0,9781896. За DC тип поновака узето је у обзир првих 66 димензија које покривају 95% материјала и добијена вредност Хопкинсове статистике износи 0,9749552. Добијена вредност Хопкинсове статистике износи 0,9682564 за IN тип поновака, при чему је узето у обзир првих 53 димензија које покривају 95% материјала. Вредност Хопкинсове статистике износи 0,9721908 за IC тип поновака при чему је узето у обзир 63 димензије које покривају 95% материјала.

Матрица сличности између свих парова секвенци је израчуната користећи косинусну сличност, након чега је примењено хијерархијско кластеровање. Добијени резултати су графички представљени топлотном мапом (Слика 23 за DN тип поновака, додатак слике 6.1.3, 6.1.4 и 6.1.5 за остале типове поновака). Уочена је јасна подела која је у складу са познатим таксономским класификацијама.



Слика 23. Матрица сличности добијена као резултат R-P/F методе примењене на скуп нуклеотидних секвенци изолата РНК вируса еболе, марбург, и бетакоронавируса за DN тип поновака приказана топлотном мапом. Сличност секвнеци је наглашена интензитетом боје. Секвенце су поређане на основу дендрограма који је добијен као резултат методе хијерархијског кластеровања.

Израчунате су мере спољашње провере коректности кластеровања на основу познатих информација о таксономским групама за род и врсту. Такође, број кластера је изабран на основу познатог броја таксономских група (3 за род и 9 за врсту). Када се као познат елемент приликом израчунавања мера спољашње провере коректности кластеровања користи вредност рода из таксономске класификације, добијају се резултати који су идентични за све типове поновака (Табела 15). У случају коришћења вредности врста из таксономске класификације, приликом израчунавања мера спољашње провере коректности кластеровања добијени су идентични резултати за DN, IN и DC тип поновака (Табела 16), док се резултати за IC тип података веома занемарљиво разликују.

Табела 15. Мере спољашње провере коректности кластеровања изолата РНК вируса еболе, марбург, и бетакоронавируса за род (позната таксономија) на основу R-P/F методе. За све типове поновака су добијене исте вредности.

Кластер /врста	**Betacoronavirus**	**Ebolavirus**	**Marburgvirus**	**Прецизност**	**Одзив**	**Ф-мера**	**Ентропија**	**Чистоћа**
1	0	192	90	0,68085	0,94118	0,79012	0,90345	192
2	87	0	0	1	1	1	0	87
3	0	12	0	1	0,05882	0,11111	0	12
Укупно	**87**	**204**	**90**	**0,76378**	**0,92682**	**0,83744**	**0,66870**	**162,35433**

Табела 16. Мере спољашње провере коректности кластеровања изолата РНК вируса еболе, марбург, и бетакоронавируса за род (позната таксономија) на основу R-P/F методе у случају DN, IN и DC тип поновака

Кластер/ врста	**Bundibugyo ebolavirus**	**Ebola virus sp.**	**Marburg marburgvirus**	**Reston ebolavirus**	**SARS-like coronavirus WIV16**	**Severe acute respiratory syndrome-related coronavirus**	**Sudan ebolavirus**	**Tai Forest ebolavirus**	**Zaire ebolavirus**	**Прецизност**	**Одзив**	**Ф-мера**	**Ентропија**	**Чистоћа**
1	0	0	0	**18**	0	0	0	0	0	1	1	1	0	18
2	0	0	0	0	0	0	0	0	**143**	1	1	1	0	143
3	0	0	0	0	1	**86**	0	0	0	0,98851	1	0,99422	0,09054	86
4	0	1	**89**	0	0	0	0	0	0	0,98889	0,98889	0,98889	0,08807	89
5	0	0	0	0	0	0	17	0	0	1	1	1	0	17
6	9	0	0	0	0	0	0	0	0	1	1	1	0	9
7	0	0	0	0	0	0	0	4	0	1	1	1	0	4
8	0	12	0	0	0	0	0	0	0	1	0,92308	0,96000	0	12
9	0	0	1	0	0	0	0	0	0	1	0,01111	0,02198	0	1
Укупно	**9**	**13**	**90**	**18**	**1**	**86**	**17**	**4**	**143**	**0,99475**	**0,99236**	**0,99355**	**0,04148**	**96,57743**

4.1.3. Тестирање *R-P/F* методе на референтним скуповима протеинских секвенци

Метода је примењена на 6 скупова референтних протеинских секвенци различитих организама: 1) NADH дехидрогеназу субјединица 5 (MT-ND5), 2) NADH дехидрогеназу субјединица 6 (MT-ND6), 3) ксиланазу (чије су секвенце категорисане групама: F10 i G11), 4) трансферин, 5) бета-глобин и 6) *spike* протеин корона вируса. Протеинске секвенце су преузете из NCBI базе података [13] у FASTA формату. Такође, из NCBI базе података, су преузети детаљи о организмима чије су протеинске секвенце анализиране и њиховој таксономији. Детаљи о секвенцама наведених скупова су приложени у додатку (додатне табеле 6.1.2 – 6.1.7). Предлог скупова за тестирање методе (одабир организама и протеинских секвенци њима заједничких протеина) је преузет из студије Saw и сарадници 2019. [78]. Скупови су минимално ажурирани у складу са најновијим стањем у NCBI бази података.

Број протеинских секвенци у скуповима је од 8 до 50 у зависности од анализираног протеина (Табела 17). Дужина секвенци је у опсегу од 141 аминокиселине (минимална дужина секвенце бета-глобина) до 1.447 (максимална дужина секвенце *spike* протеина корона вируса). Карактеристике наведених скупова података су приказане у табели 17. R-P/F метода је примењена на 6 скупова протеинских секвенци, за оба типа поновака примењивих на протеинске секвенце (DN и IN).

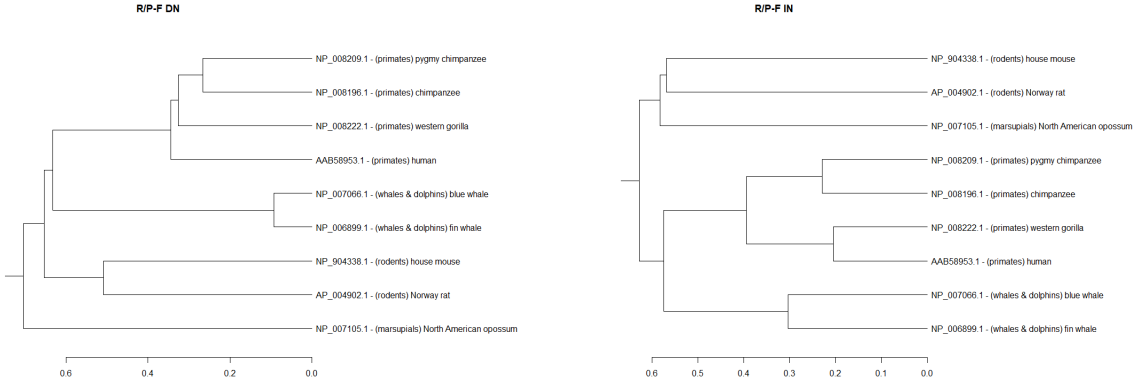
Табела 17. Особине поновака за 6 скупова референтних протеинских секвенци

Скуп протеинских секвенци	**Број секвенци**	**Дужина секвенци**			**Величина векторског простора – број јединствених поновака**	
		минимална вредност	**максимална вредност**	**средња вредност**	**DN тип поновака**	**IN тип поновака**
MT-ND5	9	602	610	604	311	275
MT-ND6	8	167	175	172	69	94
Ксиланаза	20	210	484	291	267	244
Трансферин	24	677	717	697	630	244
Корона вирус	50	1.162	1.447	1.283	1.191	791
Бета-глобин	50	141	148	146	106	92

Уочено је да је највећи број поновака дужине 3 (аминокиселине) у већини тестних скупова протеинских секвенци (Слика 24). Краћи поновци у протеинским секвенцама у односу на нуклеотидне секвенце су резултат комбинације већег броја карактера (4 карактера код нуклеотидних секвенци, и 20 карактера код протеинских секвенци). Међутим, будући да су аминокиселинске секвенце одговорне за функционалност самог протеина (прецизно увијање протеина, формирање активних места итд.) поновци представљају кључне карактеристике за анализу сличности између секвенци јер су потенцијално фаворизовани кроз процес еволуционе конзервације.

Слика 24. Расподела дужине поновака за 6 скупова референтних протеинских секвенци

Матрица сличности између свих векторских репрезентација парова секвенци појединачних протеина је израчуната користећи косинусну сличност. Примењено је хијерархијско кластеровање и добијени резултати су представљени дендрограмима (Слика 25 за MT-ND5 скуп података, додатне слике 6.1.9, 6.1.10, 6.1.14, 6.1.15, 6.1.19, 6.1.20, 6.1.24, 6.1.25, 6.1.29, 6.1.30 за остале скупове података). За различите скупове протеинских секвенци израчунате су вредности Хопкинсове статистике и добијене вредности су приказане у Табели 18.



(а) (б)

Слика 25. Дендрограмски приказ секвенци протеина NADH дехидрогеназе субјединица 5 (MT-ND5) конструисано R-P/F методом за DN (а) и IN (б) тип поновака. У листовима су приказани идентификатори секвенци, иза којих следи информација о врсти којој та секвенца припада.

У случају скупа протеина *MT-ND5* за DN тип поновака и скупа протеина *MT-ND6* за IN скуп поновака**,** на основу генерисаних дендрограма се уочава успешно груписање секвенци које је у складу са таксономском класификацијом. Такође је и јасно визуелно издвајање врсте North American opossum, у односу на друге таксономске категорије. Примена R-P/F методе над скупом протеина *ксиланазе*, није дала очекиване резултате груписања протеинских секвенци по групама (F10 и G11). У случају скупа протеина трансферина за DN тип поновака, R-P/F методе је дала успешну поделу на трансферин и лактоферин/лактотрансферин протеине. *Корона вируси* се деле у 4 групе на основу домаћина: коронавируси сисара су класификовани групом I и II, група III садржи корона вирусе птица, док група IV садржи SARS-CoVs вирусе [78]. R-P/F метода је дала исправне резултате за *spike протеине корона вируса* у случају DN типа поновака. Како је *spike* протеин корона

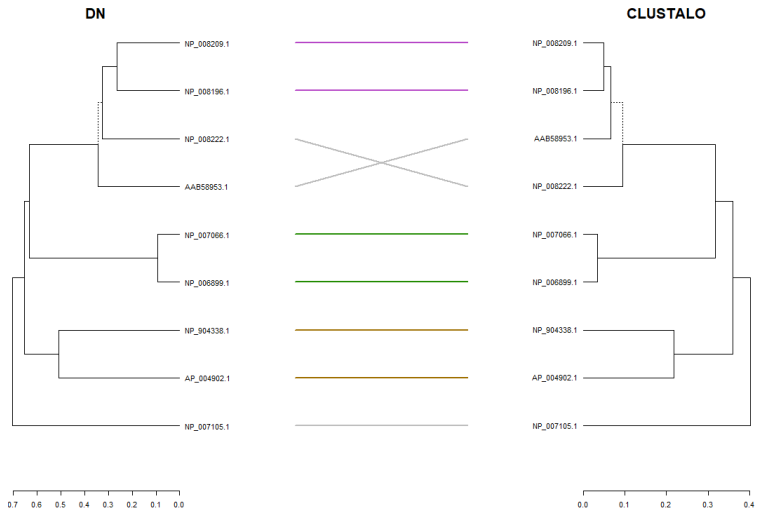
вируса Covid-19 (YP_009724390.1) накнадно додат у тестни скуп података, за ову секвенцу није била дефинисана група у изворном скупу података. На основу извршене анализе R-P/F методом, *spike* протеин корона вируса Covid-19 (YP_009724390.1) је најближи групи IV како за DN, тако и за IN тип поновака што је у складу са класификацијом описаном у литератури [80] [81]. На основу додатних слика 6.1.24 и 6.1.25 се такође уочава исправно раздвајање секвенци по групама II, III и IV, док је група I подељена на два дела за IN тип поновака. Подела за скуп протеина *бета-глобина* није у складу са таксономијом.

У циљу процене квалитета методе у односу на друге јавно расположиве методе, упоређивани су резултати добијени овом методом и резултати поређења секвенци засноване на поравнању добијени методом *Clustal Omega*. Методе су примењене на истим скуповима секвенци. На добијене матрице растојања из *Clustal Omega* методе је примењен *UPGMA* алгоритам за хијерархијско кластеровање и резултати су приказани дендрограмима. Дендрограми добијени као резултат R-P/F методе, Clustal Omega методе и упоредни прикази су представљени на слици 26 и додатним сликама 6.1.6, 6.1.11, 6.1.16, 6.1.21, 6.1.26, 6.1.31, 6.1.7, 6.1.12, 6.1.13, 6.1.17, 6.1.18, 6.1.22, 6.1.23, 6.1.27, 6.1.28, 6.1.32, 6.1.33.

Парови дендрограма добијени применом R-P/F методе и *Clustal Omega* методе над истим скуповима протеинских секвенци су упоређени израчунавањем кофенетског коефицијента корелације и РФ растојања (Табела 18). Највише вредности кофенетског коефицијента корелације (висока вредност сличности између дендрограма добијених из различитих метода) су добијене за скупове протеинских секвценци *MT-ND5* и *Spike протеине корона вируса*. Најниже вредности РФ растојања (највише слични дендрограми) су добијени за *MT-ND5, MT-ND6* скупове података, затим *трансферин* и *Spike протеине корона вируса*.

Табела 18. Вредности Хопкинсове статистике за различите скупове протеинских секвенци, као и кофенетског коефицијента корелације и РФ растојања између дендрограма добијених из *Clustal Omega* и Р/П-Ф метода. Подебљане су највише вредности кофенетског коефицијента корелације (висока вредност сличности између дендрограма добијених из различитих метода) и најниже вредности РФ растојања (највише слични дендрограми).

Скуп протеинских секвенци	**Тип поновака**	**Хопкинсова статистика**	**Кофенетски коефицијент корелације**	**РФ растојање**
MT-ND5	DN	0,5439108	**0,9773738**	**2**
	IN	0,5911131	**0,9319781**	**2**
MT-ND6	DN	0,5725814	0,8539581	**2**
	IN	0,5765676	0,9188384	**2**
Ксиланаза	DN	0,6897214	0,720081	24
	IN	0,6657664	0,8338732	22
Трансферин	DN	0,6679793	0,9271463	**16**
	IN	0,6408383	0,9259313	20
Корона вирус	DN	0,8456782	**0,9794029**	**14**
	IN	0,860263	**0,9800219**	18
Бета-глобин	DN	0,8058485	0,7115278	76
	IN	0,7787247	0,7271148	64



Слика 26. Компаративни приказ дендрограма секвенци протеина дехидрогеназе субјединица 5 (MT-ND5) конструисано R-P/F методом за DN тип поновака и UPGMA методом примењеном на матрицу растојања добијену из *Clustal Omega* методе.

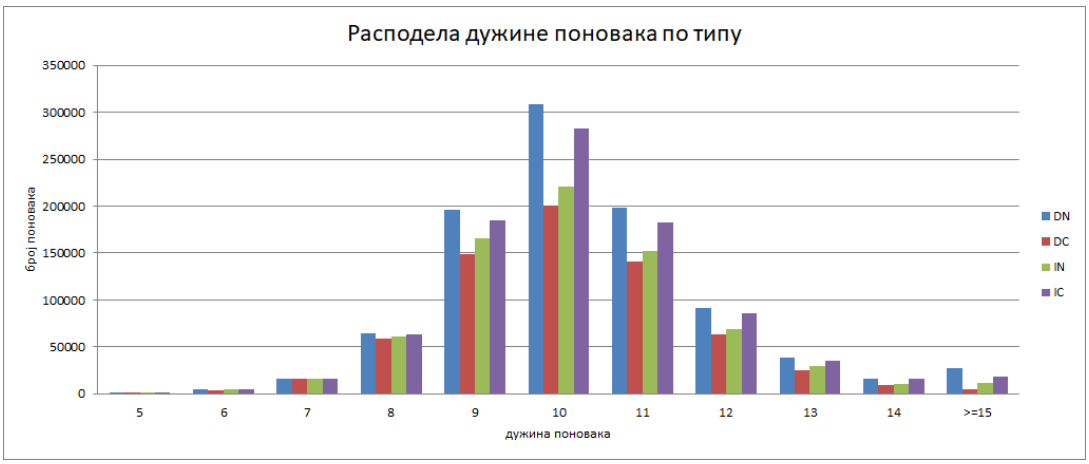
4.2 Метода заснована на профилима секвенци и потписима категорија

У овом одељку је описана примена новог начина трансформације секвенци у векторе атрибута у анализи сличности и класификацији биолошких секвенци. Скалабилност ових приступа је проверена на скупу података целих вирусних секвенци чија дужина варира од 229 до 196.858 нуклеотида, као и делова секвенци. Различите таксономске групе су коришћене у анализи као што су: род, породица, ред, иако се предложене методе могу применити на различите категорије или заједничке особине секвенци. Анализирано је 4.231 примера секвенци вируса преузетих из *ICTV* базе података *(International Committee on Taxonomy of Viruses (ICTV) database - Virus Metadata repository (VMR))* [17]. Ова база садржи листу примера вируса за сваку врсту познату у ICTV бази секвенци, као и линк ка нуклеотидној секвенци исте. Таксономске категорије за анализиране секвенце су преузете из *NCBI* базе таксономија [77]. Дати скуп податка је небалансиран, што представља додатан изазов у методама истраживања података. Пре извршавања анализе, скуп је методом случајног избора подељен на скуп података за тренирање и тест. Резултати су упоређени са резултатима *BLAST* алгоритма. Такође су проучаване перформансе класификације у односу на различите вредности улазних параметара.

Приликом идентификовања поновака, коришћена је p=0.05 вредност за одређивање статистичке значајности поновка. Минимална дужина статистички значајних поновака у овом скупу података је 5 нуклеотида. Максимална дужина поновака, као и број парова поновака у датим секвенцама по типу поновка су приказани у табели 19. Највећи број статистички значајних поновака је дужине 9, 10 и 11 за сва 4 типа поновака (Слика 27) у овом скупу података.

Табела 19. Особине поновака за скуп примера секвенци вируса

Тип поновка	**Број секвенци**	**Минимална дужина поновака**	**Максимална дужина поновака**	**Број парова поновака**
DN	4.231	5	5.168	962.880
DC	4.231	5	237	674.214
IN	4.231	5	295	744.198
IC	4.231	5	29.139	890.201



Слика 27. Расподела дужине поновака за скуп вирусних секвенци

4.2.1. Резултати методе *Одређивање сличних секвенци поређењем потписа секвенци*

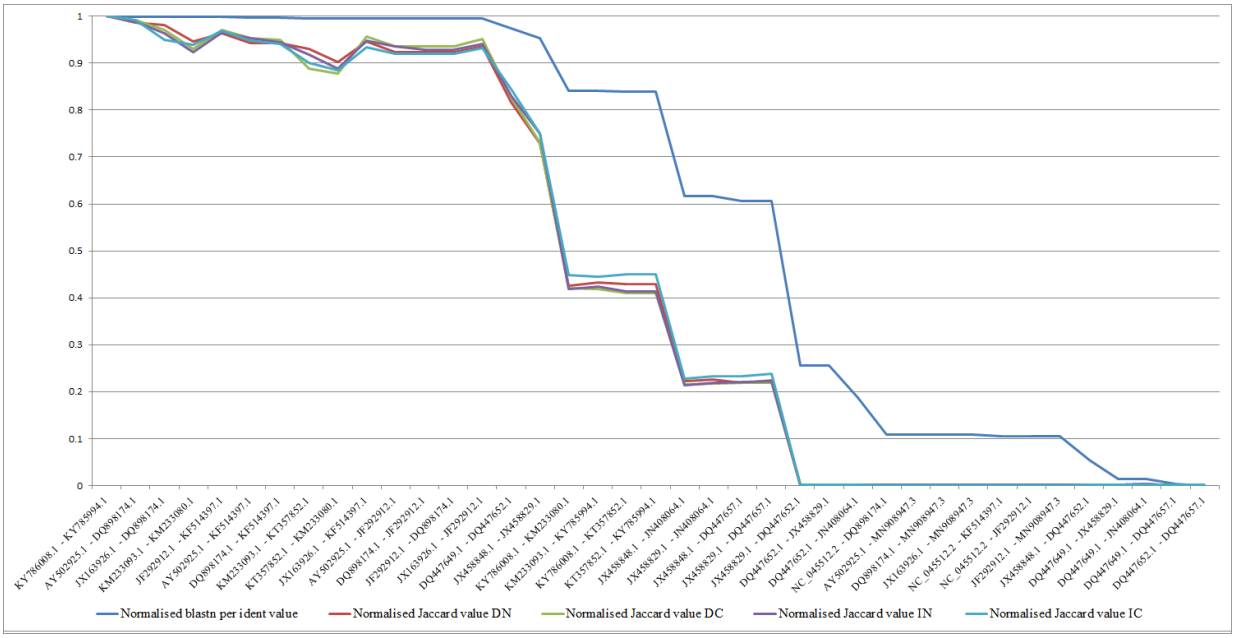
Оригинални скуп података је подељен на два скупа методом случајног избора. Први скуп секвенци садржи 16 секвенци целог генома, други 742 секвенце. Тестирано је укупно око 1200 парова секвенци (сви парови секвенци од којих прва секвенца припада првом скупу и друга секвенца припада другом скупу), користећи различите параметре модела. Добијени резултати су упоређени са резултатима *BLAST* алгоритма који је примењен на истим паровима секвенци [82]. Такође, метода је тестирана на деловима секвенци. Део секвенце (одређен почетком и дужином секвенце) је изабран насумично и посматран је као потпуно нова секвенца у бази података. Неколико елемената ван граница је додато са намером додатне провере исправности методе.

Извршавање методе је поновљено неколико пута за различите мере сличности: Жакардово растојање, косинусно растојање и коефицијент Танимотоа. Такође је извршено за сва четири типа поновака DN, DC, IN и IC. Коефицијент корелације између резултата методе извршене за различите мере сличности је представљен у табели 20. Добијени резултати су високо корелисани у случају коришћења различитих мера сличности. У случају анализе користећи IC тип поновака добијени су идентични резултати приликом коришћења Жакардовог коефицијента и коефицијента Танимотоа као мере сличности за тестирани скуп података.

Табела 20. Вредности коефицијента корелације над резултатима методе одређивање сличних секвенци поређењем потписа секвенци примењено на целе секвенце генома користећи различите мере сличности: Жакардов коефицијент, косинусно растојање и коефицијент Танимотоа за различите типове поновака.

Тип поновка	**Жакардов коефицијент - косинусно растојање**	**Жакардов коефицијент - коефицијент Танимотоа**	**Косинусно растојање - коефицијент Танимотоа**
DN	0,98966	0,99995	0,99056
DC	0,99089	0,99999	0,99138
IN	0,99120	0,99987	0,99153
IC	0,99159	1,00000	0,99191

Као резултат *BLAST* алгоритма, издвојено је 39 парова секвенци са значајним поравнањем које подразумева да је проценат делова улазне секвенце садржан у секвенци са којом се улазна секвенца пореди (енг. *query cover value*) већи од 70, као и сличност секвенци (енг. *per ident value*) већи од 80. Добијени парови који су имали значајно поравнање на основу *BLAST* алгоритма су такође показали високе вредности сличности за различите параметре методе засноване на потписима секвенци и све три мере сличности. Подаци су нормализовани користећи мин-макс методу за нормализацију података. Сирови и нормализовани резултати поређења издвојених парова секвенци (39 укупно) су приказани на слици 28 и табели 21 на примеру Жакардовог растојања за DN тип поновака. Остали парови, са израчунатом нижом вредношћу сличности (J) (ниже од 0.023507 у тестираном скупу података), нису били значајни ни као резултат добијен *BLAST* методом.



Слика 28. Компаративни приказ резултата метода *BLAST* и *поређење потписа секвенци* за сва 4 типа поновака (DN, DC, IN и IC). Приказани су резултати методе поређења потписа секвенци у случају Жакардове мере сличности. Поређење је приказано на 39 парова секвенци који су резултирали значајним поравнањем *BLAST* методе. На *x* оси су приказани парови секвенци, док су на *y* оси приказане вредности сличности за одређену методу.

Табела 21. Поређење резултата метода *BLAST* и *поређење потписа секвенци* за DN тип поновка у случају Жакардове мере сличности. Поређење је приказано на 39 парова секвенци који су резултирали значајним поравнањем у случају методе *BLAST*. Подаци су поређани по нормализованој вредноси резултата сличности користећи Жакардову меру.

Идентификатор секвенце 1	**Идентификатор секвенце 2**	**Проценат покривености упитне секвенце** (blast query cover value)	**Проценат идентичних секвенци** (blast per ident value)	**Нормализована вредност процента идентичних секвенци** (Normalised blast per ident value)	**Вредност Жакардове мере** (DN)	**Нормализована вредност Жакардове мере** (DN)↓
KY786008.1	KY785994.1	100,00	100,00	1,00000	1,00000	1,00000
AY502925.1	DQ898174.1	100,00	99,99	0,99950	0,98781	0,98734
JX163926.1	DQ898174.1	100,00	99,99	0,99950	0,98178	0,98109
JF292912.1	KF514397.1	100,00	99,97	0,99849	0,96543	0,96412
JX163926.1	KF514397.1	100,00	99,92	0,99597	0,94878	0,94684
KM233093.1	KM233080.1	100,00	99,98	0,99899	0,94866	0,94672
AY502925.1	KF514397.1	99,00	99,93	0,99647	0,94471	0,94261
DQ898174.1	KF514397.1	99,00	99,93	0,99647	0,94463	0,94253
JX163926.1	JF292912.1	100,00	99,9	0,99496	0,93808	0,93574
KM233093.1	KT357852.1	98,00	99,92	0,99597	0,93290	0,93036
DQ898174.1	JF292912.1	99,00	99,9	0,99496	0,92579	0,92298
JF292912.1	DQ898174.1	100,00	99,9	0,99496	0,92579	0,92298
AY502925.1	JF292912.1	99,00	99,9	0,99496	0,92572	0,92291
KT357852.1	KM233080.1	99,00	99,92	0,99597	0,90633	0,90279
DQ447649.1	DQ447652.1	100,00	99,48	0,97379	0,82267	0,81597
JX458848.1	JX458829.1	100,00	99,06	0,95262	0,73854	0,72866
KM233093.1	KY785994.1	99,00	96,86	0,84173	0,45288	0,43221
KY786008.1	KT357852.1	98,00	96,8	0,83871	0,44986	0,42908
KT357852.1	KY785994.1	99,00	96,8	0,83871	0,44986	0,42908
KY786008.1	KM233080.1	99,00	96,86	0,84173	0,44745	0,42658
JX458829.1	JN408064.1	99,00	92,38	0,61593	0,25488	0,22673
JX458829.1	DQ447657.1	100,00	92,18	0,60585	0,25169	0,22342
JX458848.1	JN408064.1	99,00	92,39	0,61643	0,25105	0,22275
JX458848.1	DQ447657.1	100,00	92,2	0,60685	0,24761	0,21918
DQ447649.1	JN408064.1	88,00	80,42	0,01310	0,03905	0,00275
DQ447652.1	JN408064.1	88,00	83,87	0,18700	0,03879	0,00248
DQ447649.1	JX458829.1	80,00	80,42	0,01310	0,03751	0,00116
JX458848.1	DQ447652.1	77,00	81,24	0,05444	0,03738	0,00101
NC_045512.2	DQ898174.1	88,00	82,3	0,10786	0,03691	0,00053
DQ898174.1	NC_045512.2	89,00	82,3	0,10786	0,03691	0,00053
DQ447649.1	DQ447657.1	90,00	80,23	0,00353	0,03690	0,00052
JX458829.1	DQ447652.1	74,00	85,22	0,25504	0,03688	0,00050
DQ447652.1	JX458829.1	74,00	85,22	0,25504	0,03688	0,00050
DQ447652.1	DQ447657.1	89,00	80,16	0,00000	0,03680	0,00042
AY502925.1	NC_045512.2	89,00	82,3	0,10786	0,03668	0,00029
JX163926.1	NC_045512.2	89,00	82,3	0,10786	0,03666	0,00027
NC_045512.2	KF514397.1	88,00	82,26	0,10585	0,03643	0,00003
NC_045512.2	JF292912.1	88,00	82,24	0,10484	0,03639	0,00000
JF292912.1	NC_045512.2	89,00	82,24	0,10484	0,03639	0,00000

Израчунат је коефицијент корелације значајних резултата добијених *BLAST* методом и резултата метода 1:1 поређење потписа секвенци за различите мере сличности (Табела 22). Добијени резултати су упоредиви са резултатима *BLAST* методе и прихватљиви за поређење целих секвенци. Вредности коефицијента корелације указују на значајну везу између резултата добијених на основу *BLAST* методе и 1:1 поређења потписа секвенци (Табела 22). Највећа вредност коефицијента корелације је добијена у случају избора косинусног растојања као мере сличности између две секвенце.

Табела 22. Вредности коефицијента корелације над добијеним значајним резултатима *BLAST* методе и методе одређивање сличних секвенци поређењем потписа секвенци за различите мере сличности: Жакардов коефицијент, косинусно растојање и коефицијент Танимотоа. Подебљане су вредности типа поновка (IC тип поновка) и мере сличности (косинусна сличност) за које су добијени најбољи резултати.

Тип поновка	**Жакардов коефицијент**	**Косинусно растојање**	**Коефицијент Танимотоа**
DN	0,73361	**0,79258**	0,73521
DC	0,73051	**0,79025**	0,73179
IN	0,73144	**0,79072**	0,73272
IC	**0,74119**	**0,79853**	**0,74248**

Додатно је израчуната вредност сличности 6 парцијалних секвенци са секвенцама из базе података (Табела 23). Парцијалне секвенце су формиране исецањем дела оригиналне секвенце, случајним избором почетне позиције и дужине. Израчунате су сличности са свим секвенцама у бази, укључујући и оригиналне секвенце (4.464 парова укупно). Пет (од укупно шест парцијалних секвенци) је резултирало високом вредношћу сличности са оригиналном секвенцом (Табела 23). Такође је примећено да парови секвенци са високом вредношћу сличности имају исту категорију. Најтачнији резултати су добијени применом косинусне сличности као мере коришћене за поређење потписа секвенци. У случају анализе делова секвенци, методе су показале боље резултате на дужим секвенцама.

Метода је независна од избора мере сличности, и различите/другачије мере могу да се примене. Коефицијент корелације израчунат над резултатима добијеним за различите мере сличности показује високу вредност за све типове поновака.

Изазов методе израчунавања се односи на трајање извршавања за дуге секвенце. Исти може да се превазиђе променом параметра дужине поновка, у случају да је бржи одговор неопходан. Добијени резултати се могу користити у алгоритмима кластеровања.

Табела 23. Резултати методе одређивање сличних секвенци поређењем потписа секвенци примењене на деловима секвенци у поређењу са оригиналном секвенцом. Подебљане су вредности типа поновка који даје најбоље резултате – вредност сличности секвенци. Најтачнији резултати су добијени променом косинусне сличности као мере коришћене за поређење потписа секвенци и ови резултати су такође подебљани.

Парцијална секвенца ID	Парцијална секв. дужина	Оригинална секвенца ID	Оригинална секв. дужина	Покривеност (заступљеност) дужине	Тип поновка	Жакардов коефицијент	Косинусно растојање	Коефицијент Танимотоа
XX447652	7140	DQ447652	19114	37.35%	DN	0,10416	**0,31186**	0,10528
					DC	**0,11512**	**0,32792**	**0,11552**
					IN	0,11399	**0,33102**	0,11446
					IC	0,10987	**0,32328**	0,11051
XX292912	10746	JF292912	29646	36.25%	DN	0,09854	**0,31105**	0,09953
					DC	0,09936	**0,31285**	0,10000
					IN	0,09692	**0,30819**	0,09797
					IC	**0,10149**	**0,31341**	**0,10242**
XX357852	4830	KT357852	18875	25.59%	DN	**0,04577**	**0,19854**	**0,04636**
					DC	0,02460	**0,14703**	0,02491
					IN	0,03512	**0,17436**	0,03533
					IC	0,03981	**0,18903**	0,04001
XX028835	1470	MK028835	18936	7.76%	DN	**0,00293**	**0,04877**	**0,00320**
					DC	0,00082	**0,02884**	0,00082
					IN	0,00219	**0,04239**	0,00219
					IC	0,00213	**0,03926**	0,00210
XX898174	2100	DQ898174	29751	7.06%	DN	**0,00259**	**0,04866**	**0,00265**
					DC	0,00196	**0,04142**	0,00194
					IN	0,00219	**0,04594**	0,00218
					IC	0,00235	**0,04468**	0,00233
XX458829	1120	JX458829	19114	5.86%	DN	0,00112	**0,03314**	0,00131
					DC	0,00120	**0,03237**	0,00121
					IN	0,00123	**0,03048**	0,00122
					IC	**0,00144**	**0,03774**	**0,00146**

4.2.2. Резултати методе *Класификација секвенци заснована на профилима категорија*

Анализирани скуп података садржи 4.231 нуклеотидних секвенци вируса. Број категорија анализираних секвенци варира од 11 до 419 (Табела 24) за различите таксономске групе (род, породица, ред). Тестирани скуп података је небалансиран и велики број категорија садржи мали број секвенци. Само мали број група је имао довољно значајан број секвенци који се препоручује у методама истраживања података. Пре извршавања анализе,

скуп је методом случајног избора подељен на скуп података за тренирање и тест. Приближно 2.5% улазног материјала је издвојено за скуп података за тестирање.

Табела 24. Особине скупа података (нуклеотидне секвенце вируса) по таксономској категорији

Таксономска категорија	Број секвенци у скупу података за тренинг	Број категорија	Број секвенци у скупу података за тест	Број парцијалних секвенци	Број секвенци елемената ван граница
род	3.971	419	109	6	2
породица	4.060	93	109	6	2
ред	2.590	11	67	6	1

За изабрани тренинг скуп секвенци, подељен по различитим таксономским категоријама (укупно 3 категорије) тренирана су по 4 класификациона модела за сваки тип поновка (DN, DC, IN и IC) и 4 различите вредности параметра јачина категорије $C_{X\%}$ (50%, 75%, 90%, 95%). Косинусно растојање и коефицијент Танимотоа су коришћени као различите мере сличности у одређивању сличности потписа секвенце и профила категорије. На тај начин је развијено и процењено 48 класификационих модела за сваку меру сличности (96 модела укупно). Приликом класификације улазне секвенце, идентификована класа је добијена за највишу вредност сличности потписа улазне секвенце и профила категорије.

Добијене вредности сличности за изабрана три примера секвенци KF514369.1, MF599507.1, JX458829.1 (од 100) су приказане у додатној табели 6.2.1. Процена квалитета модела је тестирана користећи око 100 тестних секвенци за које је позната категорија, које нису биле укључене у скуп података за тренирање (Табела 24) и извршена на основу микро/макро прецизности, одзива, Ф-мере (просечне и тежинске), метрика по класи и Хамингово растојање, у односу на различите вредности улазних параметара (додатне табеле 6.2.2, 6.2.3).

Такође су тестиране парцијалне секвенце са циљем одређивања категорије истих. Шест делимичних секвенци је коришћено у тесту. Све тестиране парцијалне секвенце су класификоване исправно бар једним класификационим моделом. Већи број модела је класификовао исправно три (од укупно шест) парцијалних секвенци (Табела 25).

Табела 25. Број/проценат различитих класификационих модела (од 48) који су исправно класификовали таксономску категорију за тестиране парцијалне секвенце

Id	Косинусно растојање	Коефицијент Танимотоа
XX898174.1	37 (77.08%)	37 (77.08%)
XX292912.1	34 (70.83%)	30 (62.5%)
XX357852.1	30 (62.5%)	25 (52.08%)
XX447652.1	19 (39.58%)	16 (33.33%)
XX458829.1	15 (31.25%)	13 (27.08%)
XX028835.1	3 (6.25%)	3 (6.25%)

Секвенца/елемент ван граница NC_001367 (дужина секвенце 6.395 нуклеотида) која припада Tobamovirus, резултирао је ниском вредношћу добијене сличности (<=0.005 приликом коришћења мере косинусног растојања, и <=0.002 приликом коришћењакоефицијента Танимотоа), што је и очекивано јер у скупу података за тренирање није било података везаних за ову категорију. Секвенца/елемент ван граница NC_045512.2 (Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, дужина секвенце 29.903 нуклеотида) је класификован коректно у складу са познатом таксономијом преузетом из ICTV и NCBI база података са већином класификационих модела (34 модела (од 48) у

случају косинусног растојања и 28 модела (од 48) у случају Танимото коефицијента) (додатне табеле 6.2.4, 6.2.5).

Као и већина метода за класификацију, и овај метод зависи од скупа података за тренирање – референтних генома који се користе у анализи. Самим тим квалитет модела директно зависи од броја секвенци у одређеној таксономској категорији. Главна предност методе је што је независан од таксономије и може да се примени на било које друге заједничке карактеристике секвенци. Предложени алгоритам даје задовољавајуће резултате и на скуповима података за тренирање који су небалансирани, што је показано у примеру.

Анализирајући резултате за различите параметре модела, могу да се изведу следећа запажања везана за квалитет модела. На основу тежинске прецизности може да се закључи да модели засновани на DN типу поновака дају прецизније резултате од модела који су изграђени на осталим типовима поновака. На основу овога може да се изведе препорука за давање предности у коришћењу DN типа поновака у односу на преостала три. Друго, квалитет модела се разликује за различите вредности параметра јачина категорије. Модели са мањом вредношћу параметра јачина категорије дају резултате са већом вредношћу прецизности. Такође, постоје разлике у резултатима у односу на избор таксономске категорије. Овај закључак је везан искључиво за конкретан избор скупа података за тестирање, а не за генералне особине модела. Применом методе на делове секвенци су добијени просечни резултати и модели захтевају надоградњу како би обезбедили прецизније резултате за класификовање делова секвенци. Елементи ван граница су јасно одвојени. Бољи резултати су добијени за класе са већим бројем секвенци у скупу података за тренирање.

Узети заједно, ови резултати показују да је могуће извести вишекласну класификацију биолошких секвенци на основу поновака различитих карактеристика и предлажу стратегију за будући развој предложених приступа.

5. Закључак и даљи рад

У оквиру ове тезе разматран је проблем анализе сличности биолошких секвенци, који је један од базичних и најактуелнијих проблема у области биоинформатике. Наведене су постојеће методе и описане нове за одређивање сличности секвенци и класификацију истих. Представљени модели су евалуирани и упоређени са тренутно најактуелнијим моделима за одређивање сличности секвенци.

Развојем и провером ових метода, заснованих на статистичким и рачунарским приступима који су примењени на понављајућим деловима секвенци, у оквиру ове тезе се доноси нов поглед на запис секвенце, проблеме одређивања сличности секвенци као и класификацију биолошких секвенци. Изазов у дизајнирању модела за одређивање сличности секвенци, је одабир начина представљања секвенци, такав да је довољно информативан, да може да укључи сложене информације, и да је истовремено тачан и да има добре перформансе. Предложено истраживање је засновано на новом, и до сада непримењиваном, приступу за поређење секвенци без поравнања на основу статистички значајних скупова поновака варијабилних дужина и различитих типова.

Када је нова секвенца откривена и потребно је да се идентификује њена функција, једна релативно поуздана метода је коришћење претпоставке да сличне секвенце имају сличну функцију. Стога предложени модели могу бити значајни за добијање информација о функцији из свих врста нуклеотидних и аминокиселинских секвенци. Нови модели би могли да буду веома корисна допуна постојећим системима. Ово је важан корак у моделовању и анализи биолошких података који је довео до интересантних и потпуно оригиналних резултата који представљају новину у овој области.

Показано је да се биолошке секвенце разврставају са великом прецизношћу по таксономским групама род, ред и породица користећи само поновке секвенци. Методе су примењиве на секвенце различитих дужина, као и на делове секвенци. Горе наведени експериментални резултати и анализа показују да предложене нове методе засноване на статистички значајним поновцима могу постићи упоредне резултате са најпопуларнијим алгоритмима за анализу сличности секвенци.

Даља обрада ове теме обухвата следеће планове:

- Проширење базе података која садржи скупове секвенци и њихове поновке ради омогућавања већег броја анализа за различите типове секвенци. Количина података у бази има директног утицаја на прецизност модела и коректност класификације непознатих секвенци
- Унапређење R-P/F методе у циљу побољшања брзине извршавања, без губитка прецизности методе на основу филтрирања поновака и примена метода димензионе редукције
- Мерење и евалуација перформанси модела за различите врсте мера сличности. Тренутно су проверене косинусна сличност, Жакардово растојање и коефицијент Танимото-а, и у плану су Хамингово растојање, Еуклидско растојање и растојање Минковског са тежинама
- Укључивање у алгоритме међусобног растојања свака два пара поновака у секвенци за различите типове поновака. Растојање не може да буде наведено статички због броја поновака, већ мора да се израчунава динамички, што знатно утиче на перформансе израчунавања, али ће знатно повећати прецизност
- Унапређење методе класификације секвенци засноване на профилима категорија у смеру укључивања поновака са одређеним бројем мутација
- Провера метода заснованих на потписима секвенци и профилима категорија на протеинским секвенцама

- Имплементација и објављивање сервиса за примену метода преко интернета који подразумева аутоматско проширење база секвенци

Очекује се да ће даља обрада предложене теме допринети бољој и ефикаснијој класификацији биолошких секвенци, филогенетској анализи, и анализи веза између секвенци и њихових функција.

6. Додатак

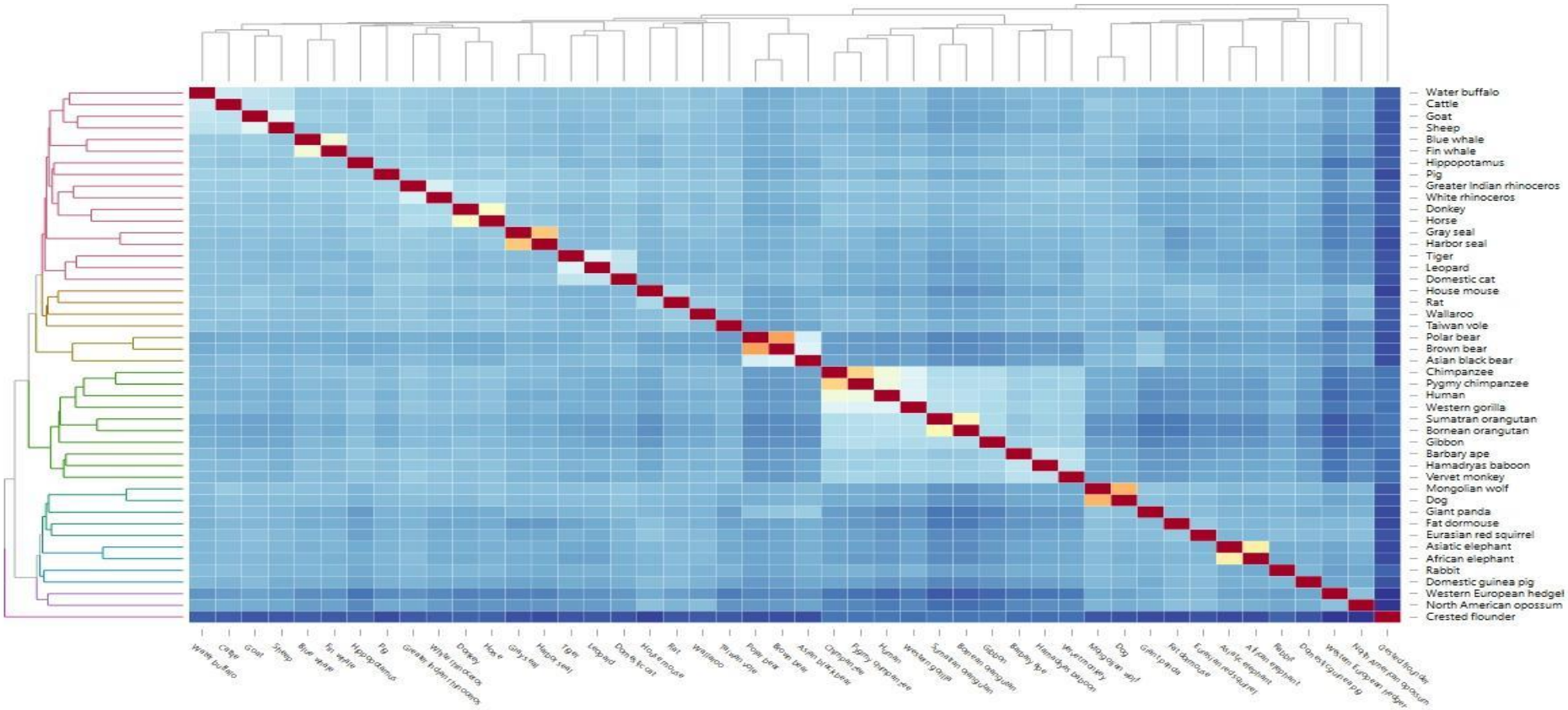
У овом поглављу су приказане додатне табеле и слике које садрже резултате за различите типове поновака који нису приказани у четвртом поглављу. Такође су приказане и табеле са више информација о тестним секвенцама и резултатима.

6.1 Додатак резултатима методе засноване на позицији и локалној учесталости поновака

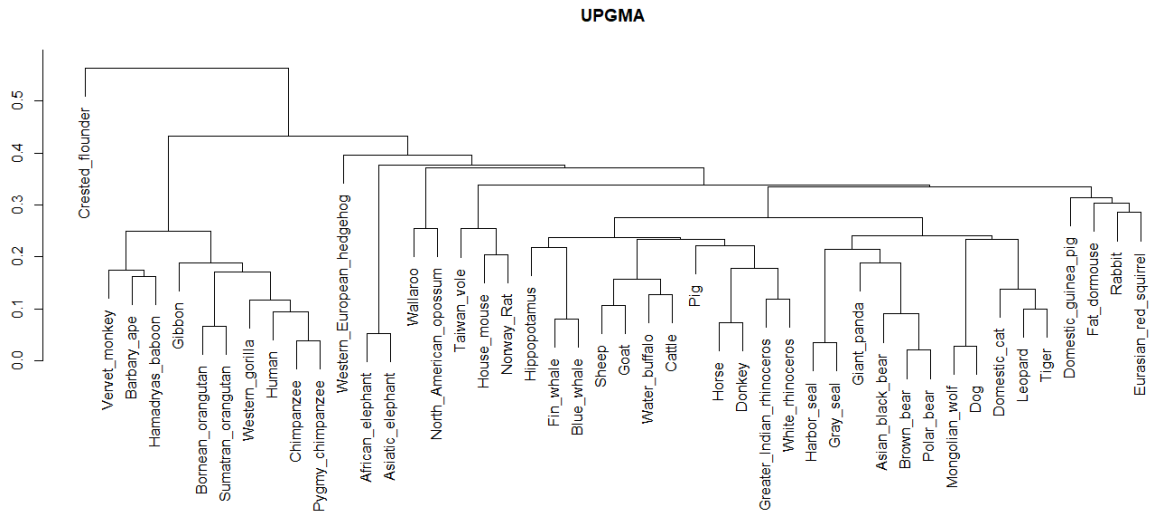
Табела додатак 6.1.1. Списак секвенци, идентификатора секвнеци и дужине секвенци митохондријалне ДНК коришћене у анализи

Редни број	**Име врсте**	**NCBI референтни број**	**Дужина секвенце (бр. нуклеотида)**
1	African elephant	AJ224821	16 866
2	Barbary ape	NC_002764	16 586
3	Asiatic elephant	DQ316068	16 902
4	Asian black bear	DQ402478	16 868
5	Blue whale	X72204	16 402
6	Bornean orangutan	D38115	16 389
7	Brown bear	AF303110	17 020
8	Water buffalo	AY488491	16 355
9	Domestic cat	U20753	17 009
10	Chimpanzee	D38113	16 554
11	Cattle	V00654	16 338
12	Dog	U96639	16 727
13	Domestic guinea pig	AJ222767	16 801
14	Donkey	X97337	16 670
15	Eurasian red squirrel	AJ238588	16 507
16	Fat dormouse	AJ001562	16 602
17	Fin whale	X61145	16 398
18	Giant panda	EF212882	16 805
19	Gibbon	X99256	16 472
20	Goat	KP231536	16 813
21	Western gorilla	D38114	16 364
22	Gray seal	X72004	16 797
23	Greater Indian rhinoceros	X97336	16 829
24	Hamadryas baboon	Y18001	16 521
25	Harbor seal	X63726	16 826
26	Western European hedgehog	X88898	17 447
27	Hippopotamus	AJ010957	16 407
28	Human	V00662	16 569
29	Horse	X79547	16 660
30	House mouse	V00711	16 295
31	Leopard	EF551002	16 964
32	North American opossum	Z29573	17 084
33	Norway Rat	X14848	16 300
34	Pig	AJ002189	16 680

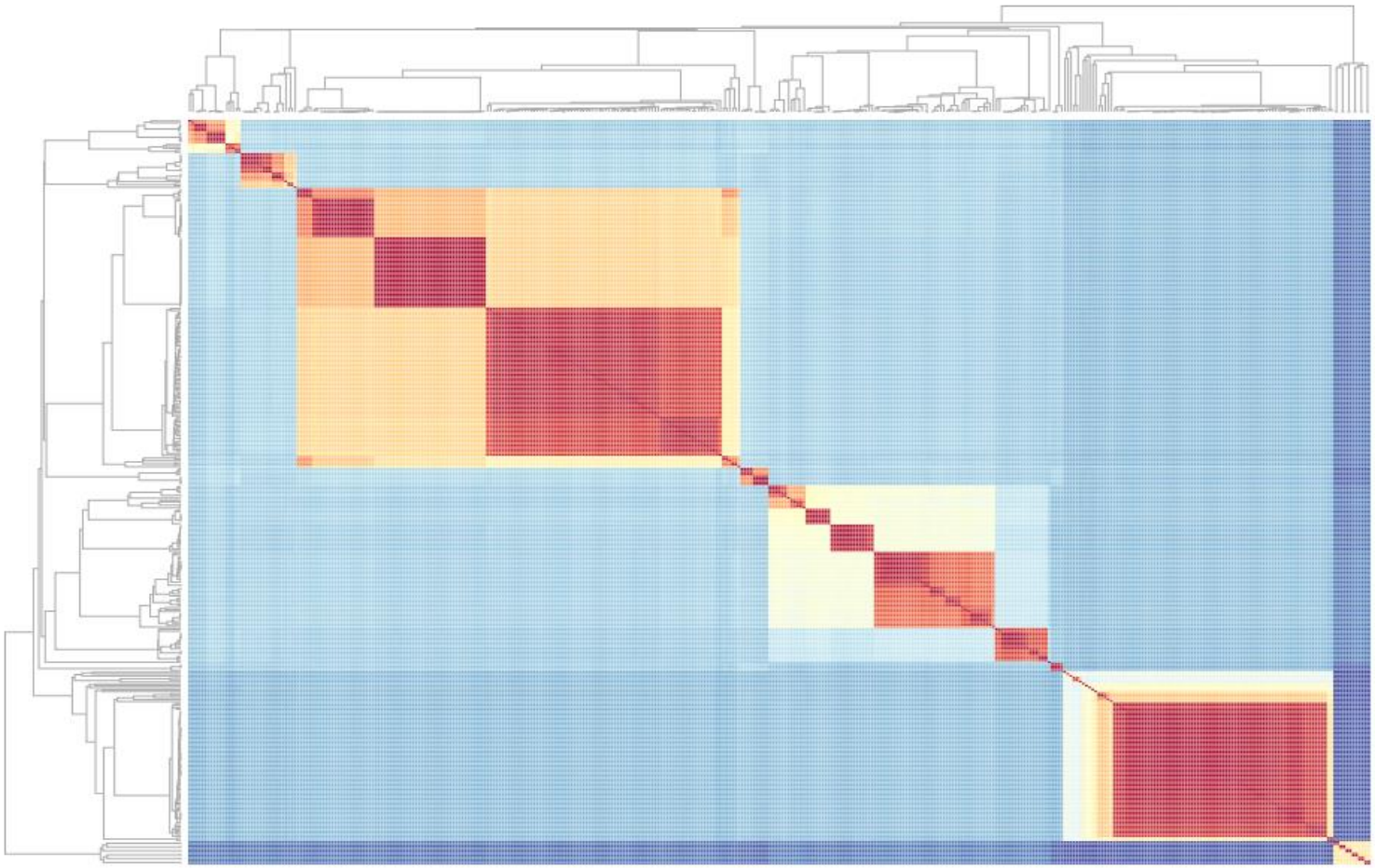
35	Polar bear	AF303111	17 017
36	Pygmy chimpanzee	D38116	16 563
37	Rabbit	AJ001588	17 245
38	Sheep	AF010406	16 616
39	Sumatran orangutan	NC_002083	16 499
40	Tiger	EF551003	16 990
41	Vervet monkey	AY863426	16 389
42	Taiwan vole	AF348082	16 312
43	Wallaroo	Y10524	16 896
44	White rhinoceros	Y07726	16 832
45	Mongolian wolf	EU442884	16 774
46	Crested flounder	KJ433567	18 642



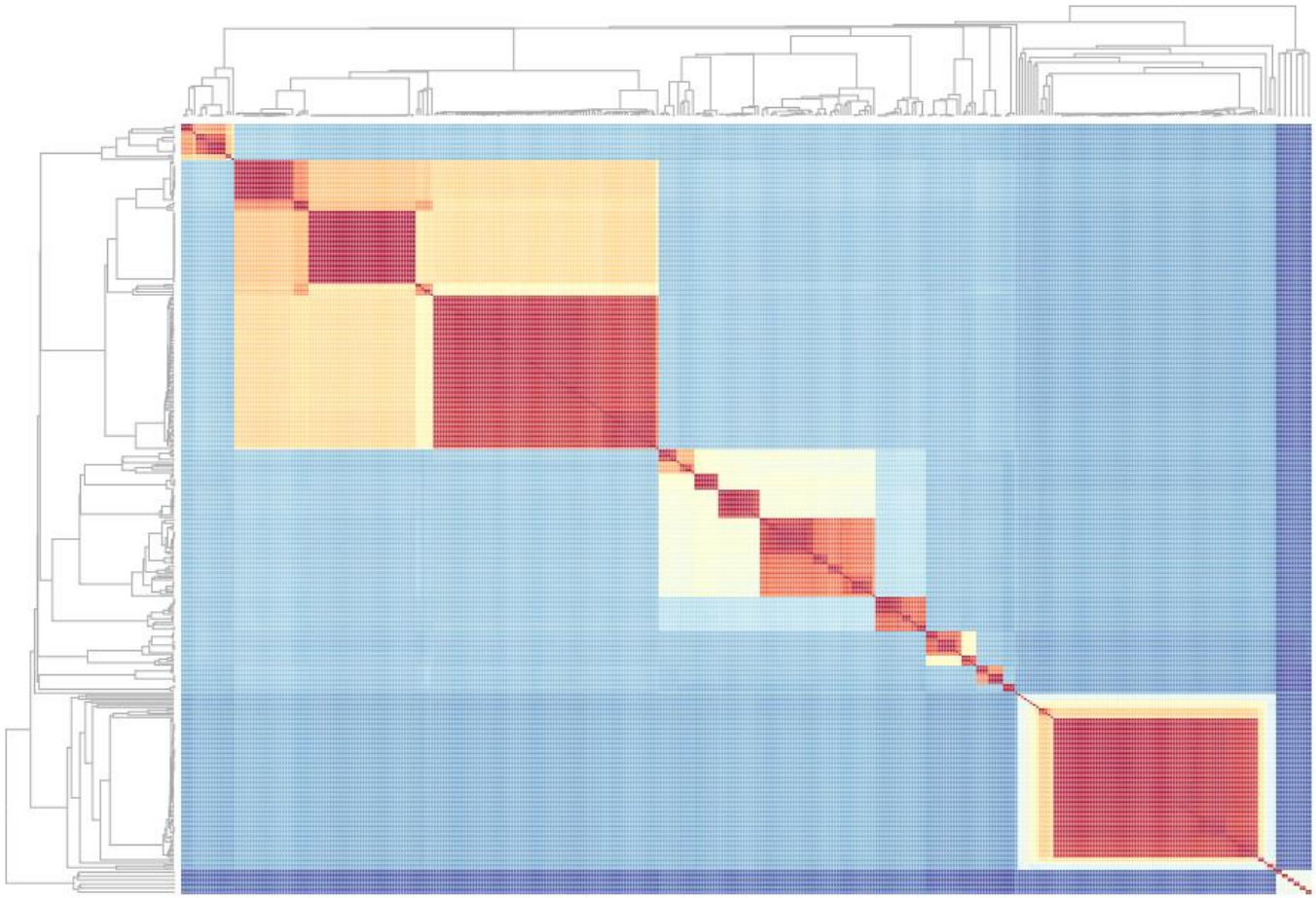
Слика додатак 6.1.1. Матрица сличности добијена као резултат R-P/F методе примењене на скуп митохондријалних ДНК секвенци за DN тип поновака приказана топлотном мапом. Сличност секвенци је наглашена интензитетом боје. Секвенце су поређана на основу денограма који је добијен као резултат методе хијерархијског кластеровања.



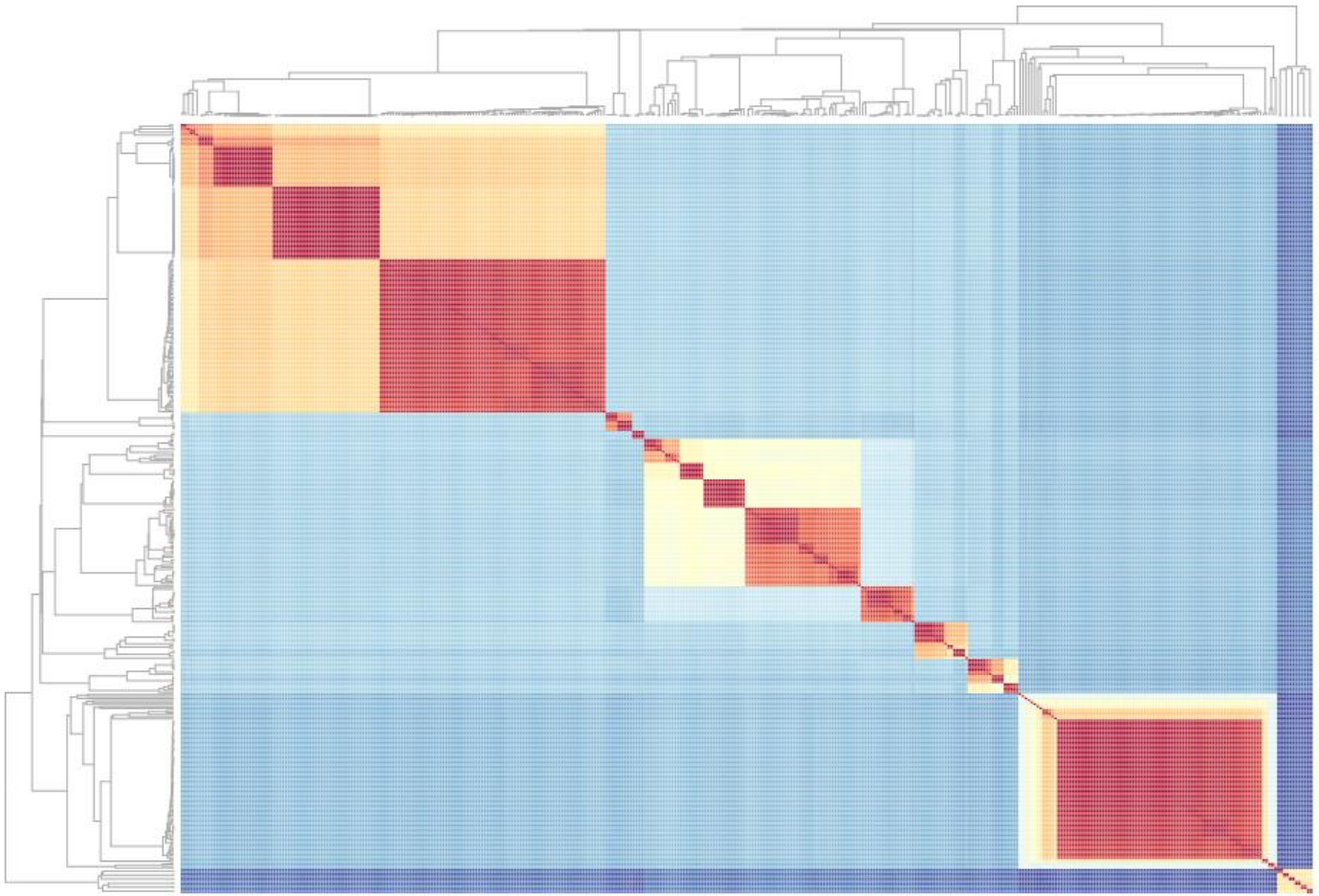
Слика додатак 6.1.2. Дендрограм који приказује резултате кластеровања Upgma алгоритмом примењеним на резултате вишеструког поравнања Clustal Omega методе за 46 митохондријалне секвенце



Слика додатак 6.1.3. Матрица сличности добијена као резултат R-P/F методе примењене на скуп нуклеотидних секвенци изолата РНК вируса еболе, марбург, и бетакоронавируса за DC тип поновака приказана топлотном мапом. Сличност секвенци је наглашена интензитетом боје. Секвенце су поређана на основу дендрограма који је добијен као резултат методе хијерархијског кластеровања.



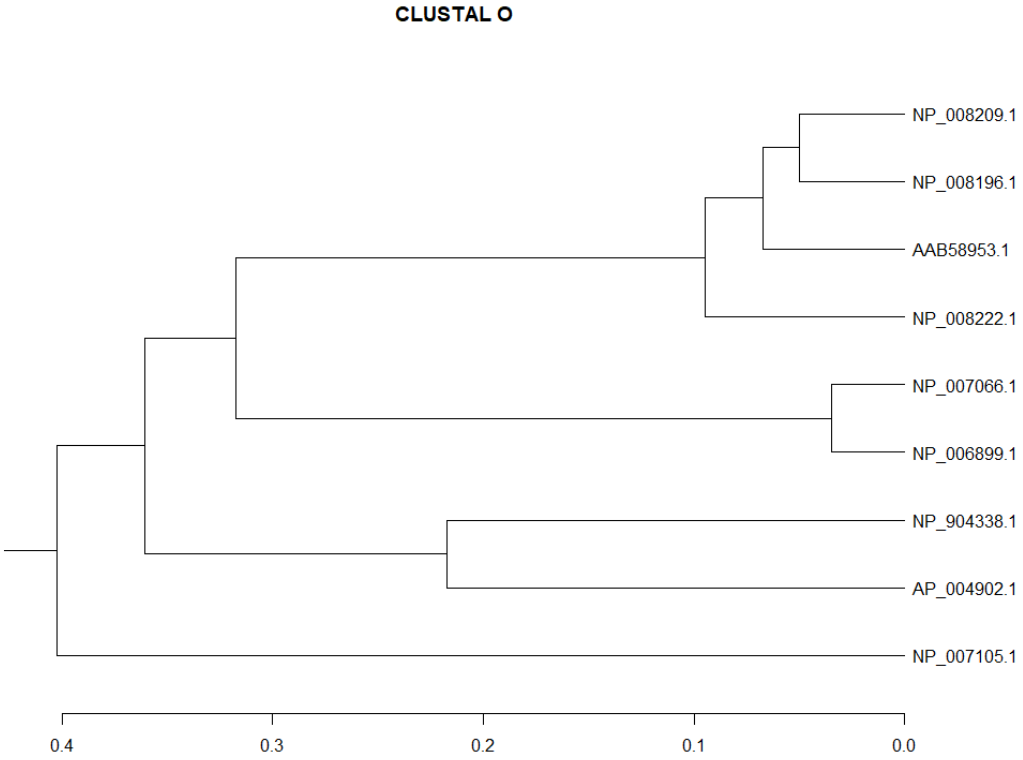
Слика додатак 6.1.4. Матрица сличности добијена као резултат R-P/F методе примењене на скуп нуклеотидних секвенци изолата РНК вируса еболе, марбург, и бетакоронавируса за IN тип поновака приказана топлотном мапом. Сличност секвенци је наглашена интензитетом боје. Секвенце су поређана на основу дендрограма који је добијен као резултат методе хијерархијског кластеровања.



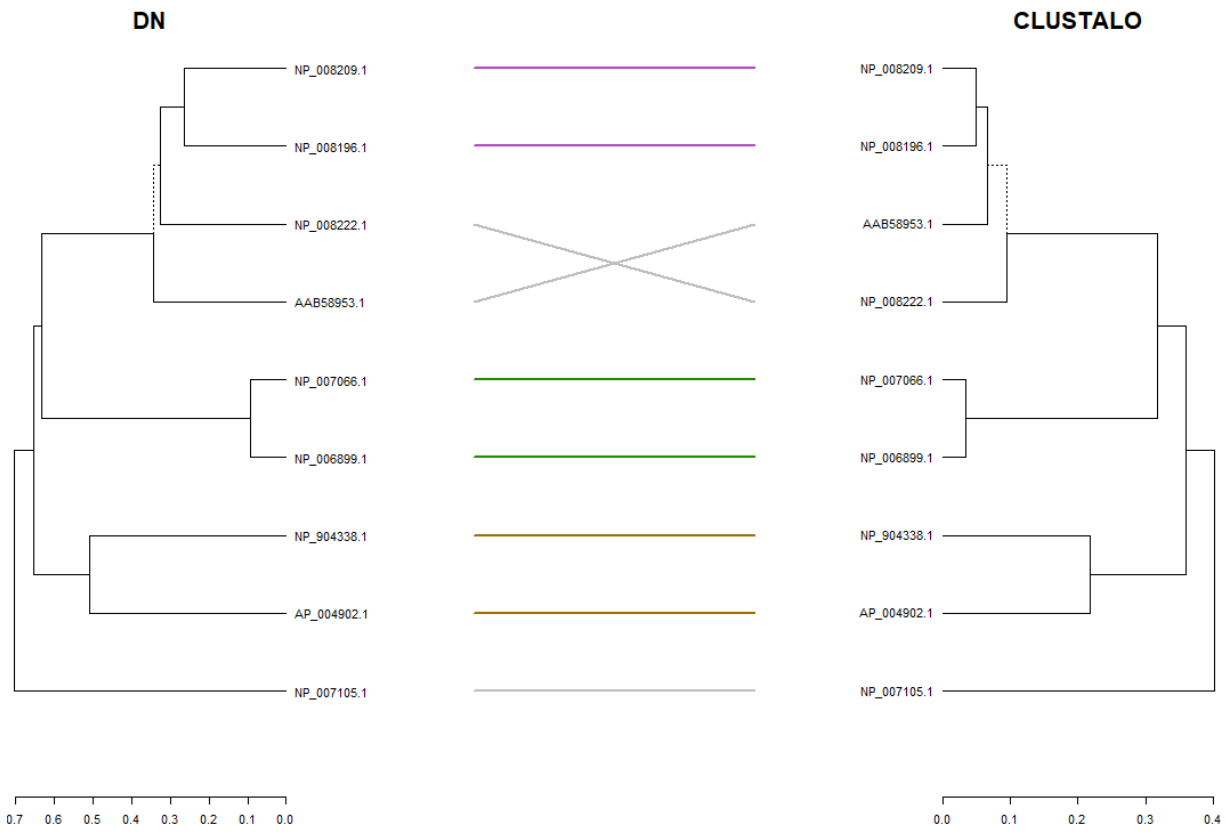
Слика додатак 6.1.5. Матрица сличности добијена као резултат R-P/F методе примењене на скуп нуклеотидних секвенци изолата РНК вируса еболе, марбург, и бетакоронавируса за IC тип поновака приказана топлотном мапом. Сличност секвенци је наглашена интензитетом боје. Секвенце су поређана на основу дендрограма који је добијен као резултат методе хијерархијског кластеровања.

Табела додатак 6.1.2. Сажете информације о 9 MT-ND5 протеинских секвенци

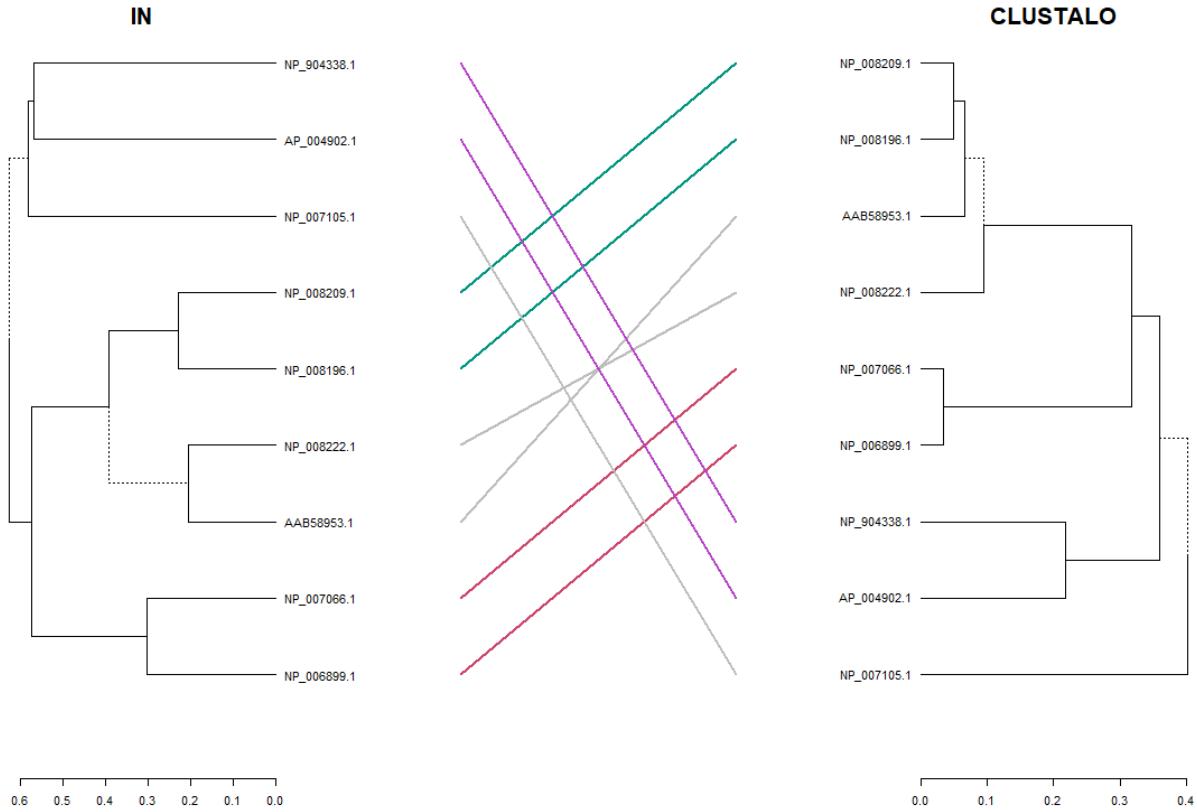
ID	Врста	Дужина
AAB58953.1	(primates) human	603
NP_008222.1	(primates) western gorilla	603
NP_008196.1	(primates) chimpanzee	603
NP_008209.1	(primates) pygmy chimpanzee	603
NP_006899.1	(whales & dolphins) fin whale	606
NP_007066.1	(whales & dolphins) blue whale	606
AP_004902.1	(rodents) Norway rat	610
NP_904338.1	(rodents) house mouse	607
NP_007105.1	(marsupials) North American opossum	602



Слика додатак 6.1.6. Дендрограмски приказ секвенци протеина NADH дехидрогеназе 5 (MT-ND5) конструисано UPGMA методом примењеном на матрицу растојања добијену из Clustal Omega програма.



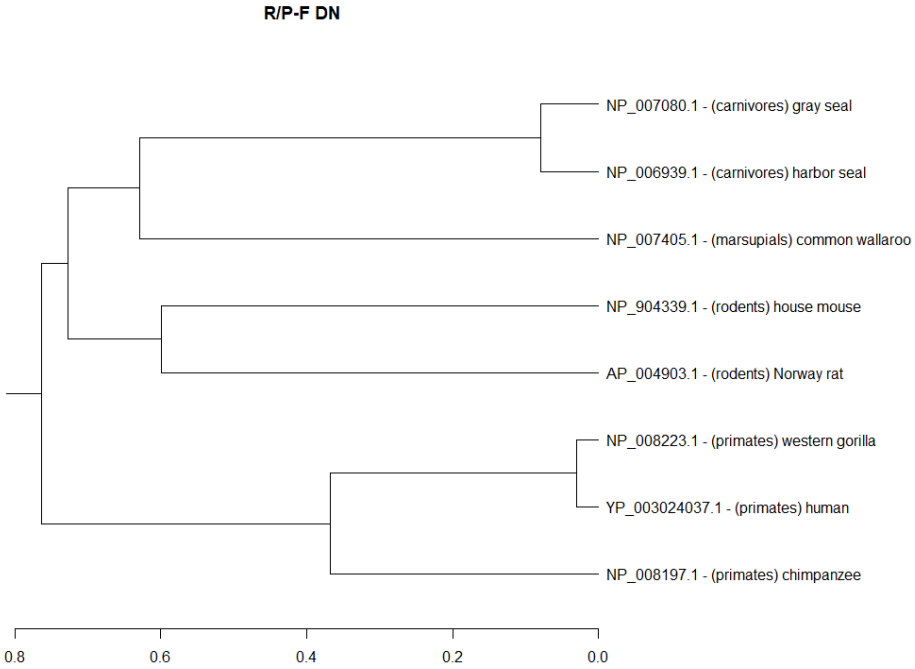
Слика додатак 6.1.7. Компаративни приказ дендрограма секвенци протеина NADH дехидрогеназе 5 (MT-ND5) конструисано R-P/F методом за DN тип поновака и UPGMA методом примењеном на матрицу растојања добијену из Clustal Omega програма.



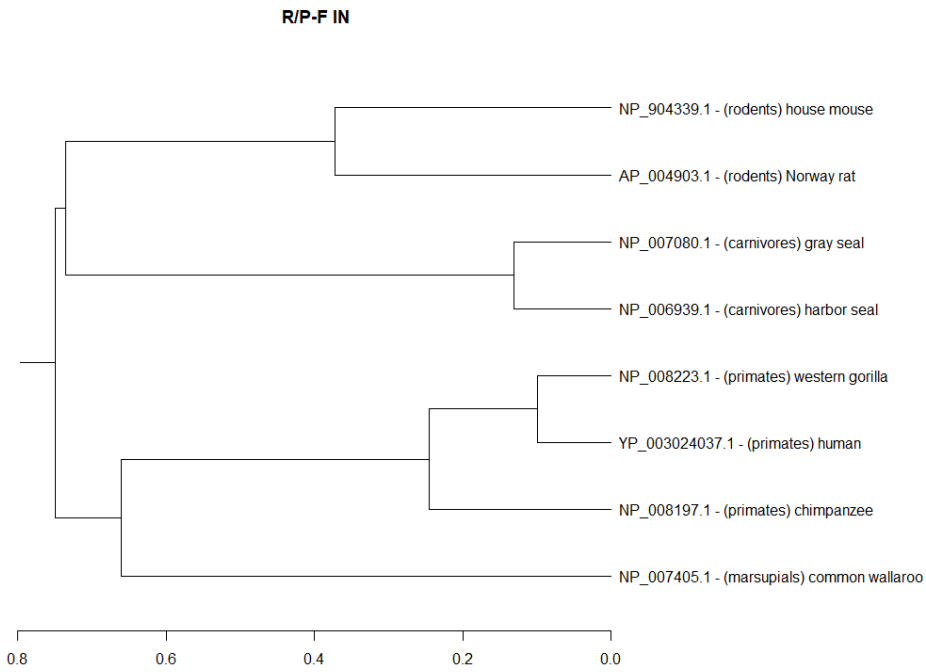
Слика додатак 6.1.8. Компаративни приказ дендрограма секвенци протеина NADH дехидрогеназе 5 (MT-ND5) конструисано R-P/F методом за IN тип поновака и UPGMA методом примењеном на матрицу растојања добијену из Clustal Omega програма.

Табела додатак 6.1.3. Сажете информације о 8 MT-ND6 протеинских секвенци

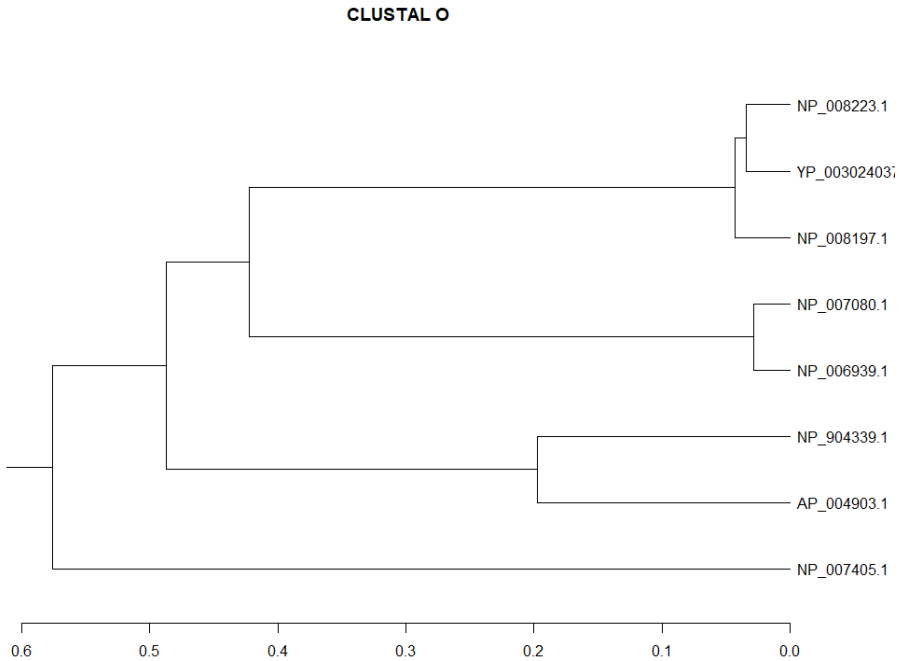
ID	Врста	Дужина
YP_003024037.1	(primates) human	174
NP_008223.1	(primates) western gorilla	174
NP_008197.1	(primates) chimpanzee	174
NP_006939.1	(carnivores) harbor seal	175
NP_007080.1	(carnivores) gray seal	175
AP_004903.1	(rodents) Norway rat	172
NP_904339.1	(rodents) house mouse	172
NP_007405.1	(marsupials) common wallaroo	167



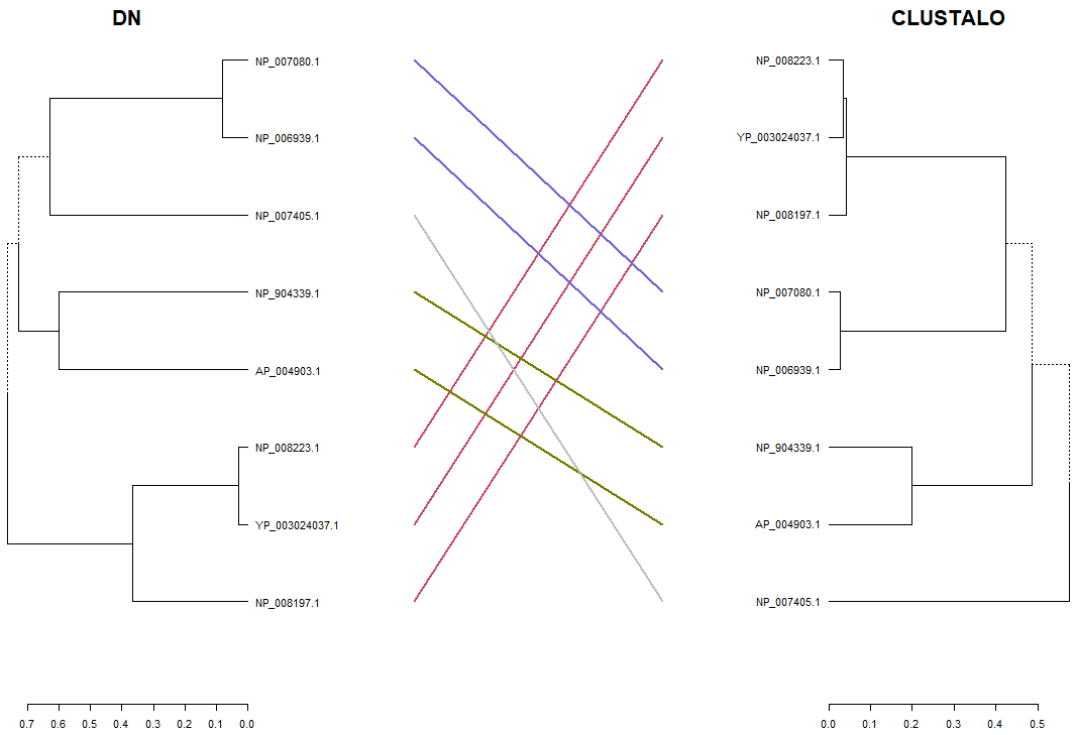
Слика додатак 6.1.9. Дендрограмски приказ секвенци протеина NADH дехидрогеназу субјединица 6 (MT-ND6) конструисано R-P/F методом за DN тип поновака. У листовима су приказани идентификатори секвенци, иза којих следи информација о врсти којој та секвенца припада.



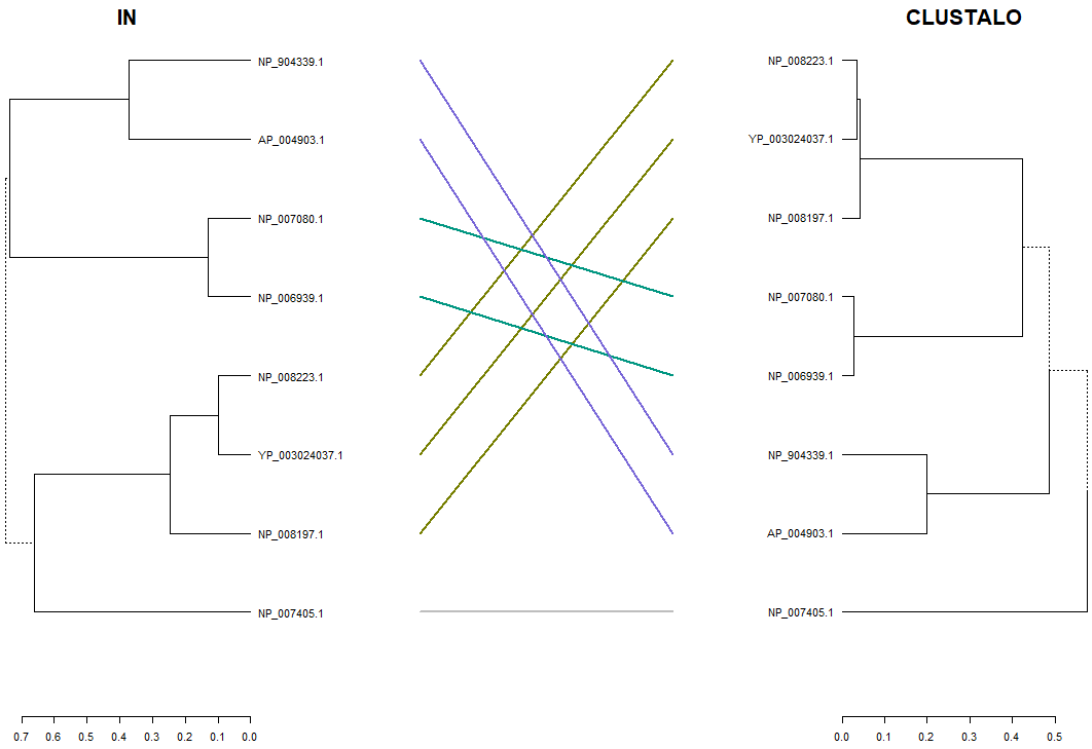
Слика додатак 6.1.10. Дендрограмски приказ секвенци протеина NADH дехидрогеназу субјединица 6 (MT-ND6) конструисано R-P/F методом за IN тип поновака. У листовима су приказани идентификатори секвенци, иза којих следи информација о врсти којој та секвенца припада.



Слика додатак 6.1.11. Дендрограмски приказ секвенци протеина NADH дехидрогеназе 6 (MT-ND6) конструисано UPGMA методом примењеном на матрицу растојања добијену из Clustal Omega програма.



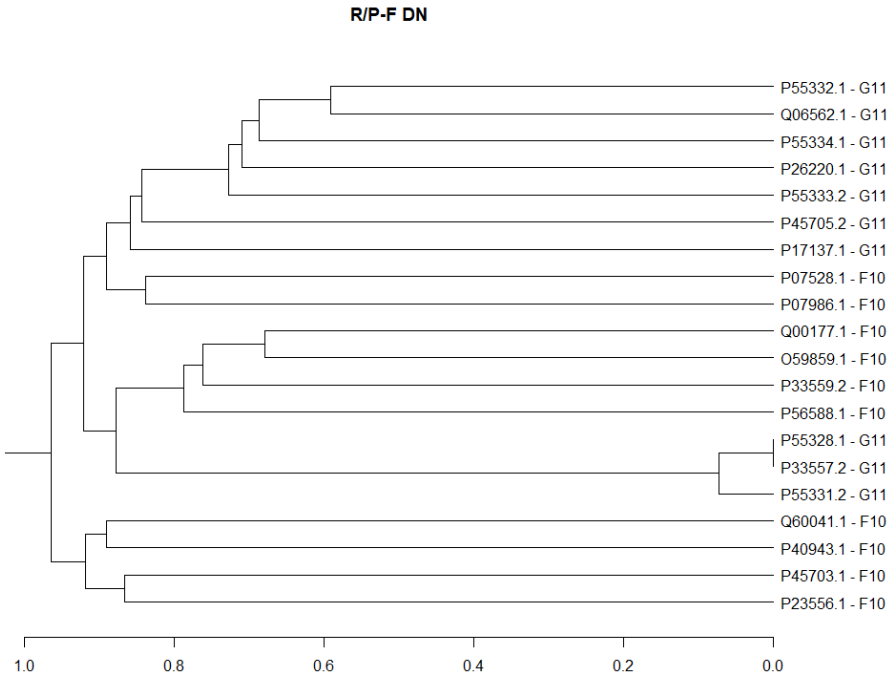
Слика додатак 6.1.12. Компаративни приказ дендрограма секвенци протеина NADH дехидрогеназе 6 (MT-ND6) конструисано R-P/F методом за DN тип поновака и UPGMA методом примењеном на матрицу растојања добијену из Clustal Omega програма.



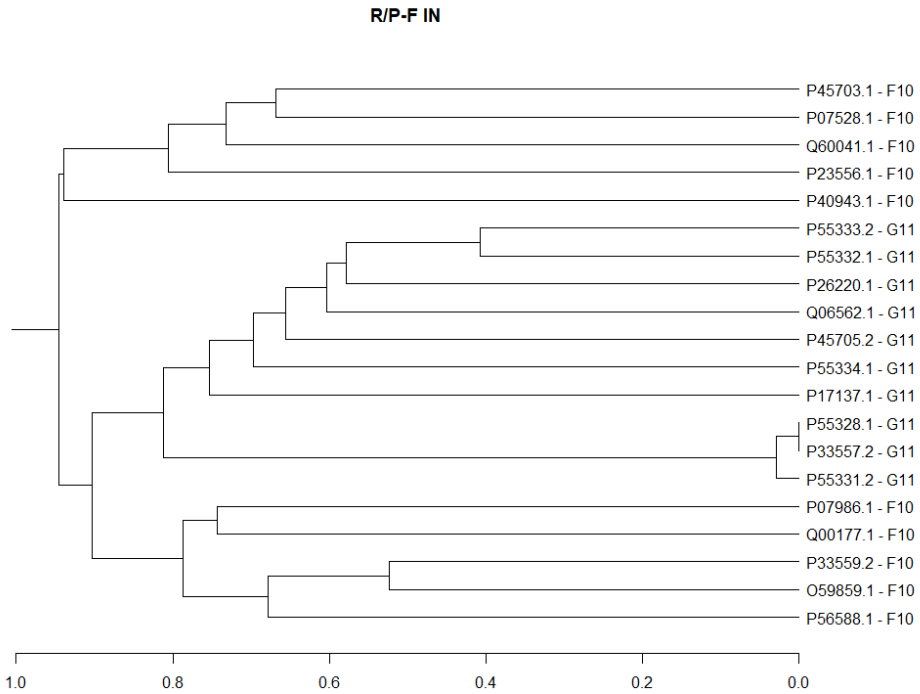
Слика додатак 6.1.13. Компаративни приказ дендрограма секвенци протеина NADH дехидрогеназе 6 (MT-ND6) конструисано R-P/F методом за IN тип поновака и UPGMA методом примењеном на матрицу растојања добијену из Clustal Omega програма.

Табела додатак 6.1.4. Сажете информације о 20 протеинских секвенци скупа ксиланазе

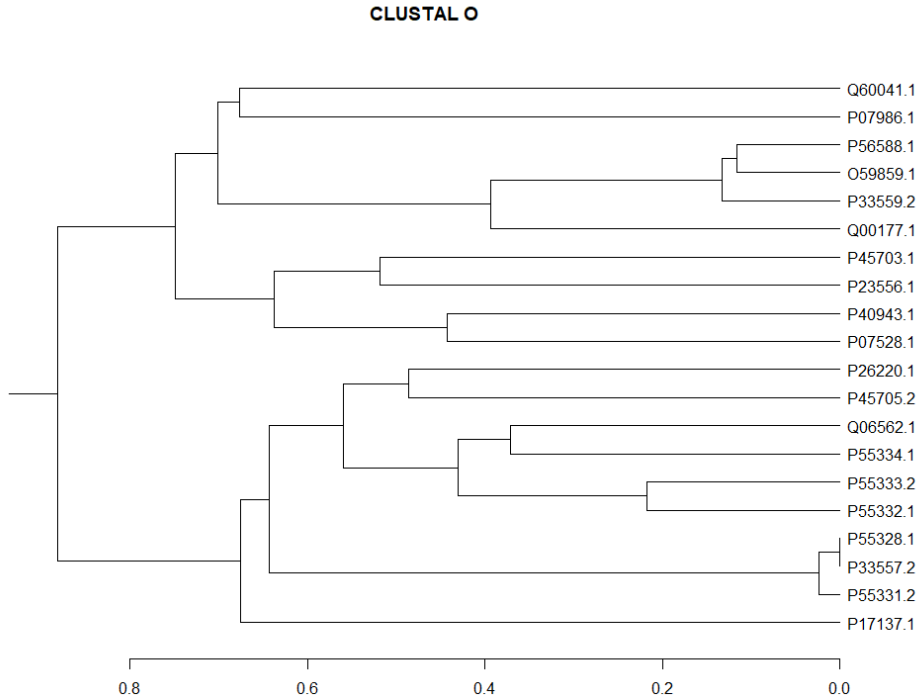
ID	Група	Дужина	ID	Група	Дужина
O59859.1	F10	327	P33557.2	G11	211
P56588.1	F10	302	P55328.1	G11	211
P33559.2	F10	327	P55331.2	G11	211
Q00177.1	F10	327	P45705.2	G11	210
P07986.1	F10	484	P26220.1	G11	240
P07528.1	F10	396	P55334.1	G11	227
P40943.1	F10	407	Q06562.1	G11	221
P23556.1	F10	342	P55332.1	G11	225
P45703.1	F10	330	P55333.2	G11	221
Q60041.1	F10	346	P17137.1	G11	261



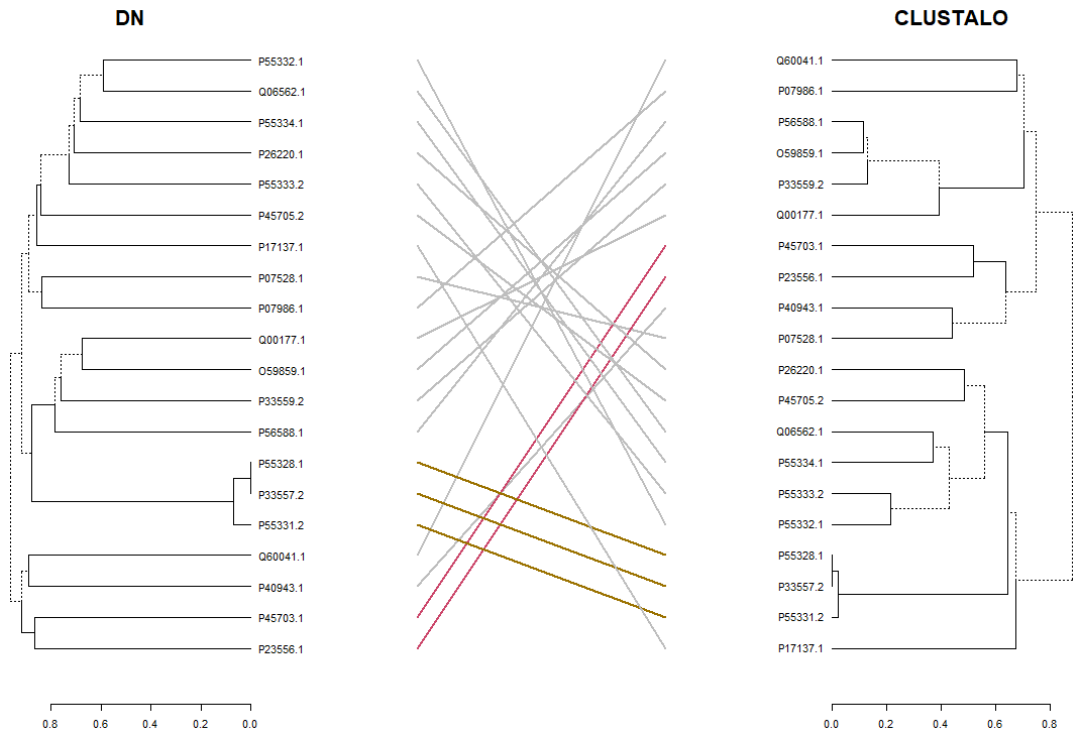
Слика додатак 6.1.14. Дендрограмски приказ секвенци протеина ксиланазе конструисано R-P/F методом за DN тип поновака. У листовима су приказани идентификатори секвенци, иза којих следи информација о групи којој та секвенца припада.



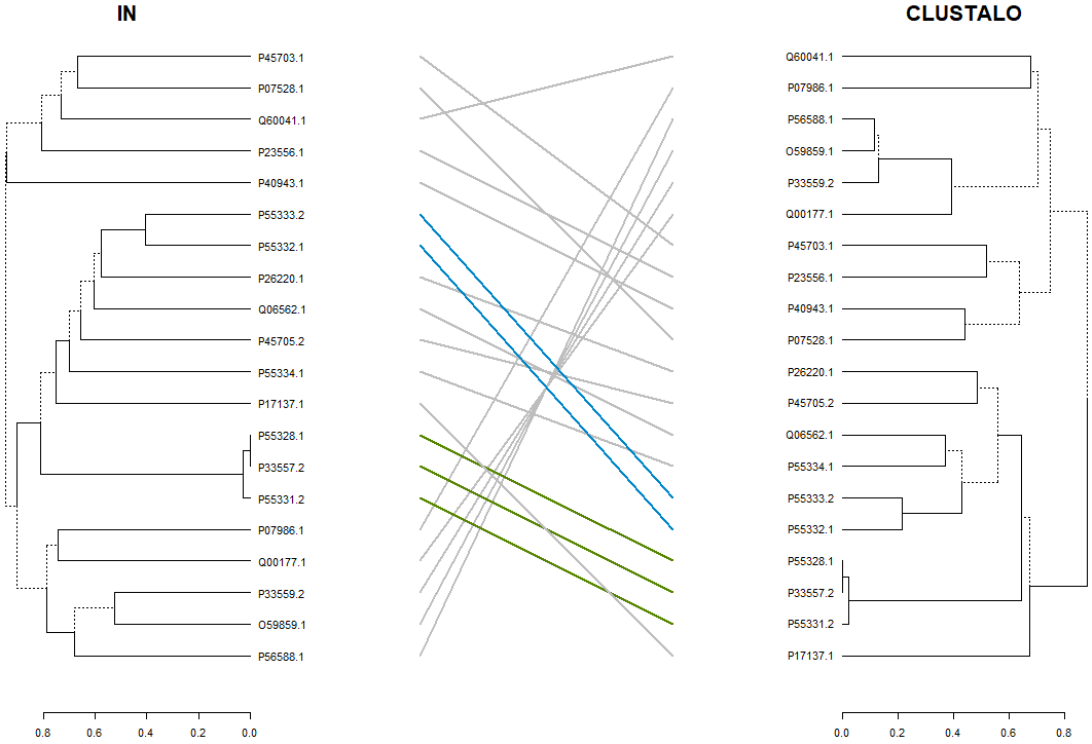
Слика додатак 6.1.15. Дендрограмски приказ секвенци протеина ксиланазе конструисано R-P/F методом за IN тип поновака. У листовима су приказани идентификатори секвенци, иза којих следи информација о групи којој та секвенца припада.



Слика додатак 6.1.16. Дендрограмски приказ секвенци протеина ксиланазе конструисано UPGMA методом примењеном на матрицу растојања добијену из Clustal Omega програма.



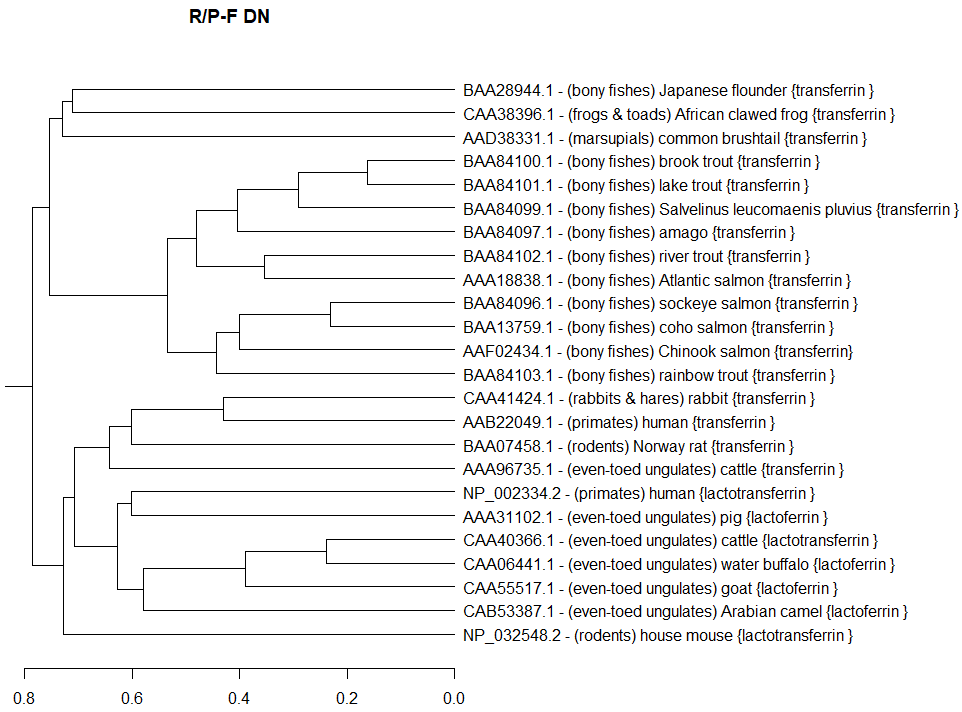
Слика додатак 6.1.17. Компаративни приказ дендрограма секвенци протеина ксиланазе конструисано R-P/F методом за DN тип поновака и UPGMA методом примењеном на матрицу растојања добијену из Clustal Omega.



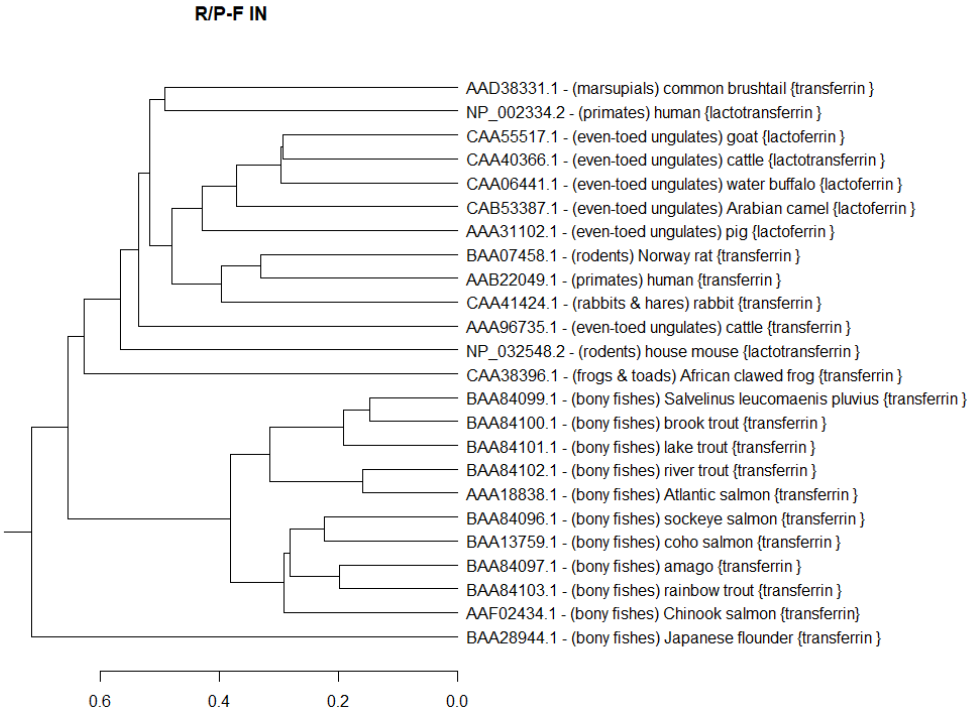
Слика додатак 6.1.18. Компаративни приказ дендрограма секвенци протеина ксиланазе конструисано R-P/F методом за IN тип поновака и UPGMA методом примењеном на матрицу растојања добијену из Clustal Omega програма.

Табела додатак 6.1.5. Сажете информације о 24 протеинских секвенци скупа протеина трансферина

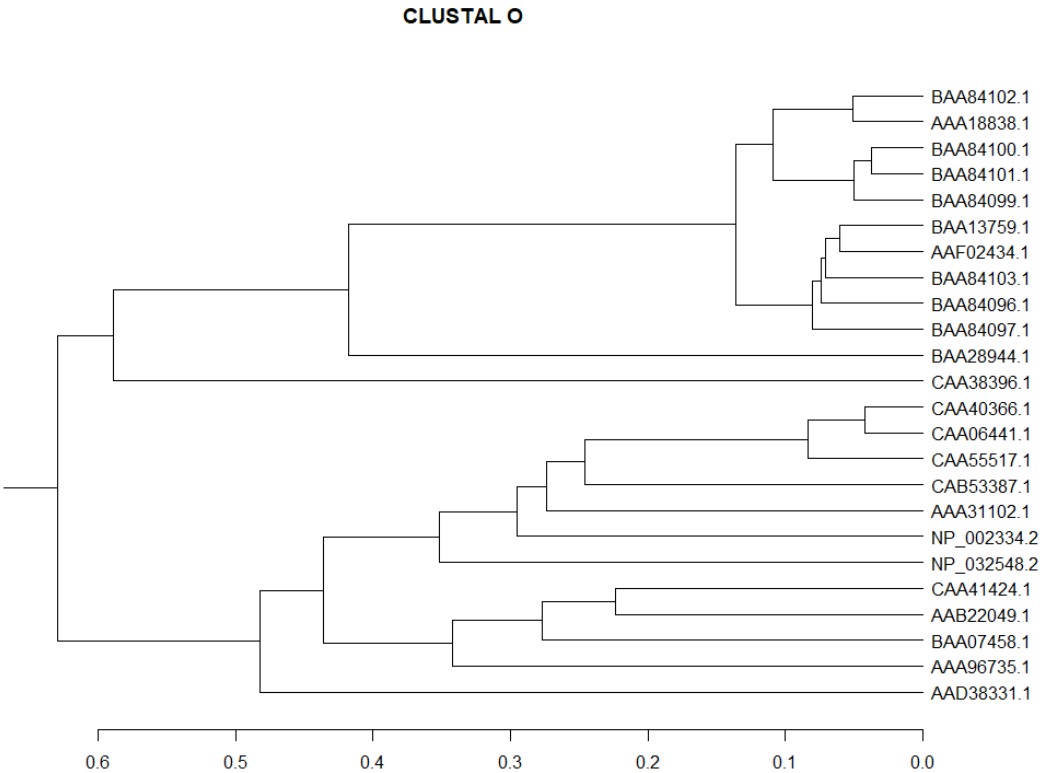
ID	Протеин	Ред	Врста	Дужина
AAB22049.1	transferrin	primates	human	698
CAA41424.1	transferrin	rabbits & hares	rabbit	694
BAA07458.1	transferrin	rodents	Norway rat	698
AAA96735.1	transferrin	even-toed ungulates	cattle	704
CAA06441.1	lactoferrin	even-toed ungulates	water buffalo	708
CAA40366.1	lactotransferrin	even-toed ungulates	cattle	708
CAA55517.1	lactoferrin	even-toed ungulates	goat	708
CAB53387.1	lactoferrin	even-toed ungulates	Arabian camel	708
AAA31102.1	lactoferrin	even-toed ungulates	pig	686
NP_002334.2	lactotransferrin	primates	human	710
NP_032548.2	lactotransferrin	rodents	house mouse	707
AAD38331.1	transferrin	marsupials	common brushtail	711
CAA38396.1	transferrin	frogs & toads	African clawed frog	717
BAA28944.1	transferrin	bony fishes	Japanese flounder	685
AAA18838.1	transferrin	bony fishes	Atlantic salmon	690
BAA84102.1	transferrin	bony fishes	river trout	691
BAA84101.1	transferrin	bony fishes	lake trout	691
BAA84100.1	transferrin	bony fishes	brook trout	691
BAA84099.1	transferrin	bony fishes	Salvelinus leucomaenis pluvius	691
AAF02434.1	transferrin	bony fishes	Chinook salmon	677
BAA13759.1	transferrin	bony fishes	coho salmon	687
BAA84096.1	transferrin	bony fishes	sockeye salmon	691
BAA84103.1	transferrin	bony fishes	rainbow trout	691
BAA84097.1	transferrin	bony fishes	amago	691



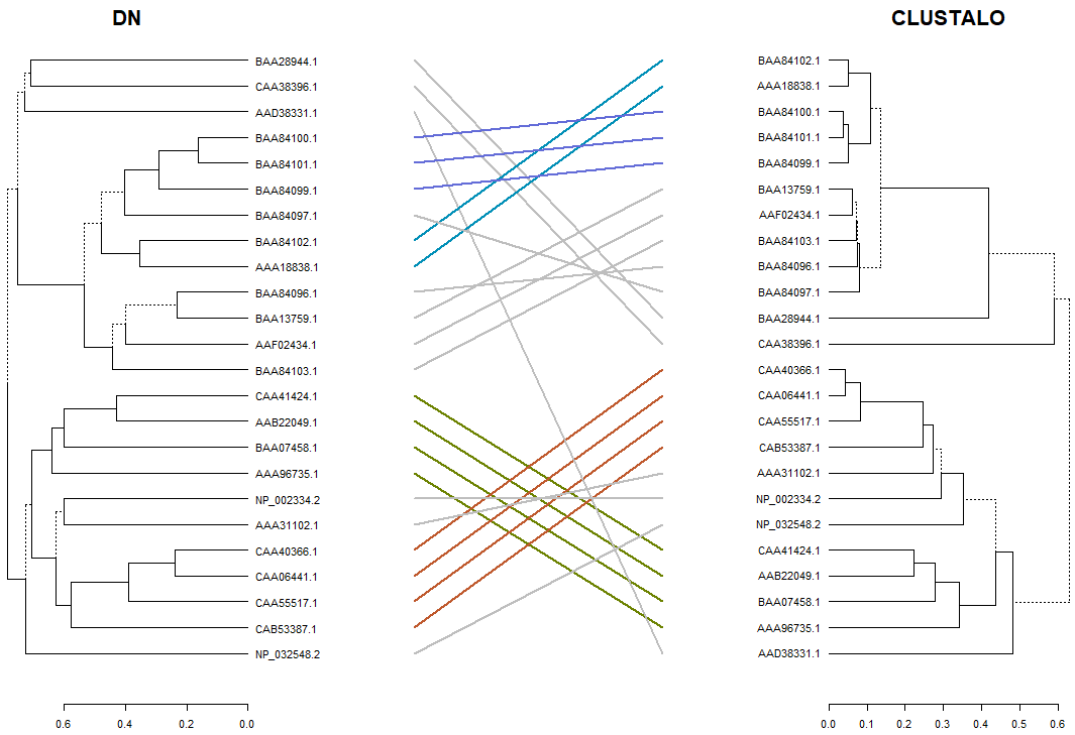
Слика додатак 6.1.19. Дендрограмски приказ секвенци протеина трансферина конструисано R-P/F методом за DN тип поновака. У листовима су приказани идентификатори секвенци, иза којих следе додатне информације о врсти протеина којој те секвенца припадају.



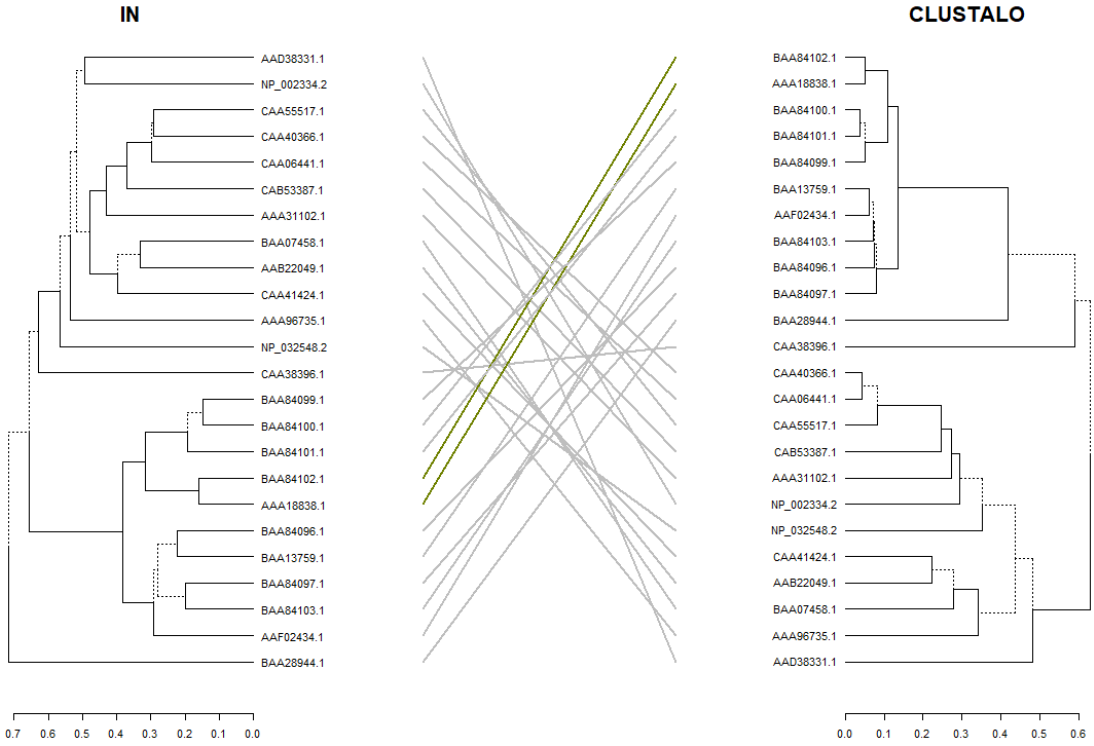
Слика додатак 6.1.20. Дендрограмски приказ секвенци протеина трансферина конструисано R-P/F методом за IN тип поновака. У листовима су приказани идентификатори секвенци, иза којих следе додатне информације о врсти протеина којој те секвенца припадају.



Слика додатак 6.1.21. Дендрограмски приказ секвенци протеина трансферина конструисано UPGMA методом примењеном на матрицу растојања добијену из Clustal Omega програма.



Слика додатак 6.1.22. Компаративни приказ дендрограма секвенци протеина ксиланазе конструисано R-P/F методом за DN тип поновака и UPGMA методом примењеном на матрицу растојања добијену из Clustal Omega програма.

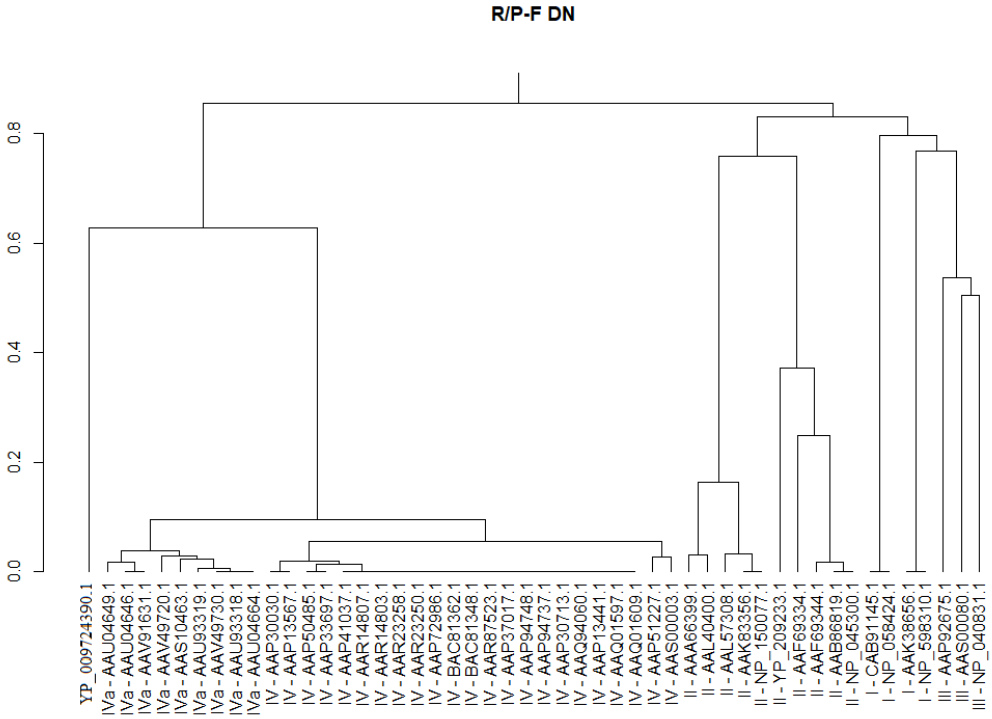


Слика додатак 6.1.23. Компаративни приказ дендрограма секвенци протеина трансферина конструисано R-P/F методом за IN тип поновака и UPGMA методом примењеном на матрицу растојања добијену из Clustal Omega програма.

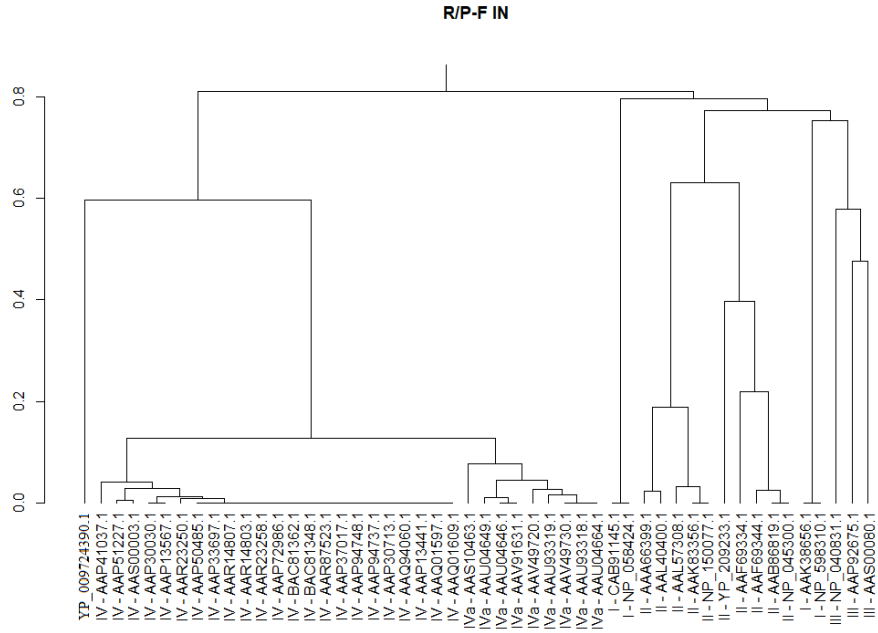
Табела додатак 6.1.6. Сажете информације о 50 протеинских секвенци скупа *Spike* протеина корона вируса

ID	Група	Дужина	ID	Група	Дужина
CAB91145.1	I	1447	AAV49730.1	IVa	1255
NP_058424.1	I	1447	AAP51227.1	IV	1255
AAK38656.1	I	1383	AAS00003.1	IV	1255
NP_598310.1	I	1383	AAP30030.1	IV	1255
AAK83356.1	II	1363	AAP13567.1	IV	1255
AAL57308.1	II	1363	AAP50485.1	IV	1255
AAA66399.1	II	1363	AAP41037.1	IV	1255
AAL40400.1	II	1363	AAQ01597.1	IV	1255
NP_150077.1	II	1363	AAQ01609.1	IV	1255
AAB86819.1	II	1324	AAP13441.1	IV	1255
YP_209233.1	II	1376	AAQ94060.1	IV	1255
AAF69334.1	II	1321	AAP30713.1	IV	1255
AAF69344.1	II	1324	AAP33697.1	IV	1255
NP_045300.1	II	1324	AAP94737.1	IV	1255
AAP92675.1	III	1169	AAP94748.1	IV	1255
AAS00080.1	III	1169	AAP37017.1	IV	1255
NP_040831.1	III	1162	AAR87523.1	IV	1255
AAS10463.1	IVa	1255	BAC81348.1	IV	1255
AAU93318.1	IVa	1255	BAC81362.1	IV	1255
AAV49720.1	IVa	1255	AAP72986.1	IV	1255
AAU93319.1	IVa	1255	AAR23250.1	IV	1255
AAU04646.1	IVa	1255	AAR23258.1	IV	1255
AAU04649.1	IVa	1255	AAR14803.1	IV	1255
AAU04664.1	IVa	1255	AAR14807.1	IV	1255
AAV91631.1	IVa	1255	YP_009724390.1	*	1273

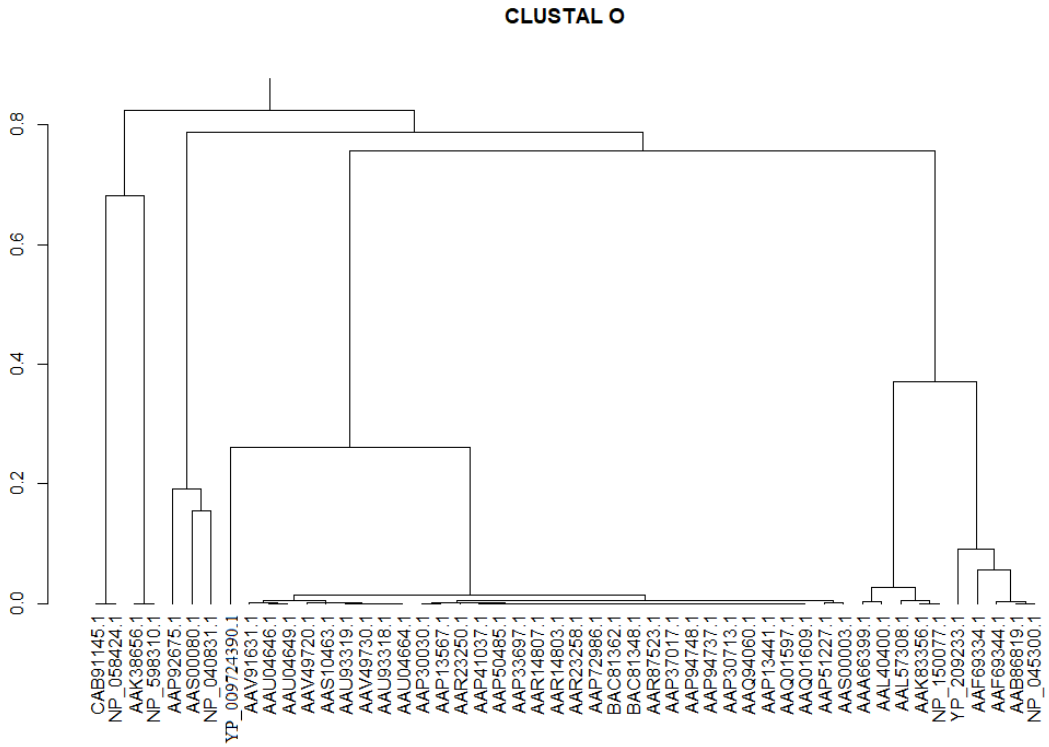
grupe I i II sadrže koranaviruse sisara, grupa III sadrži korona viruse ptica i grupa IV sadrži SARS-CoVs viruse
* spike protein Covid-19 SARS-CoV-2 virusa



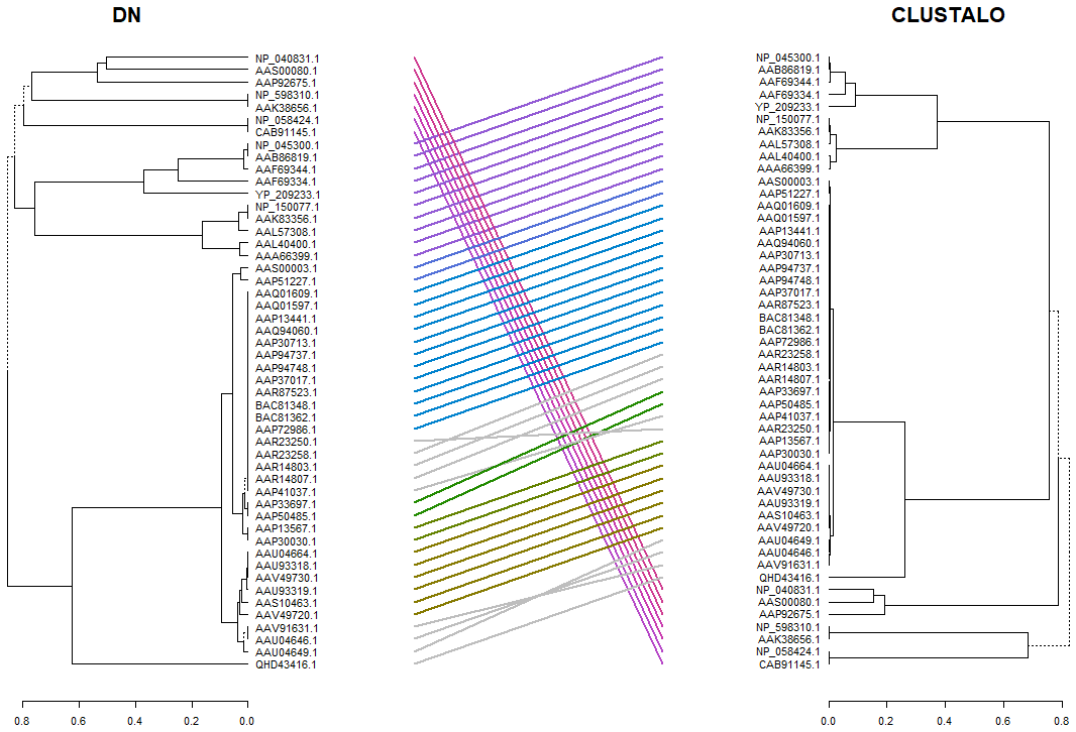
Слика додатак 6.1.24. Дендрограмски приказ секвенци *Spike* протеина корона вируса конструисано R-P/F методом за DN тип поновака. У листовима су приказани идентификатори секвенци, иза којих следе додатне информације о групи којој та секвенца припада.



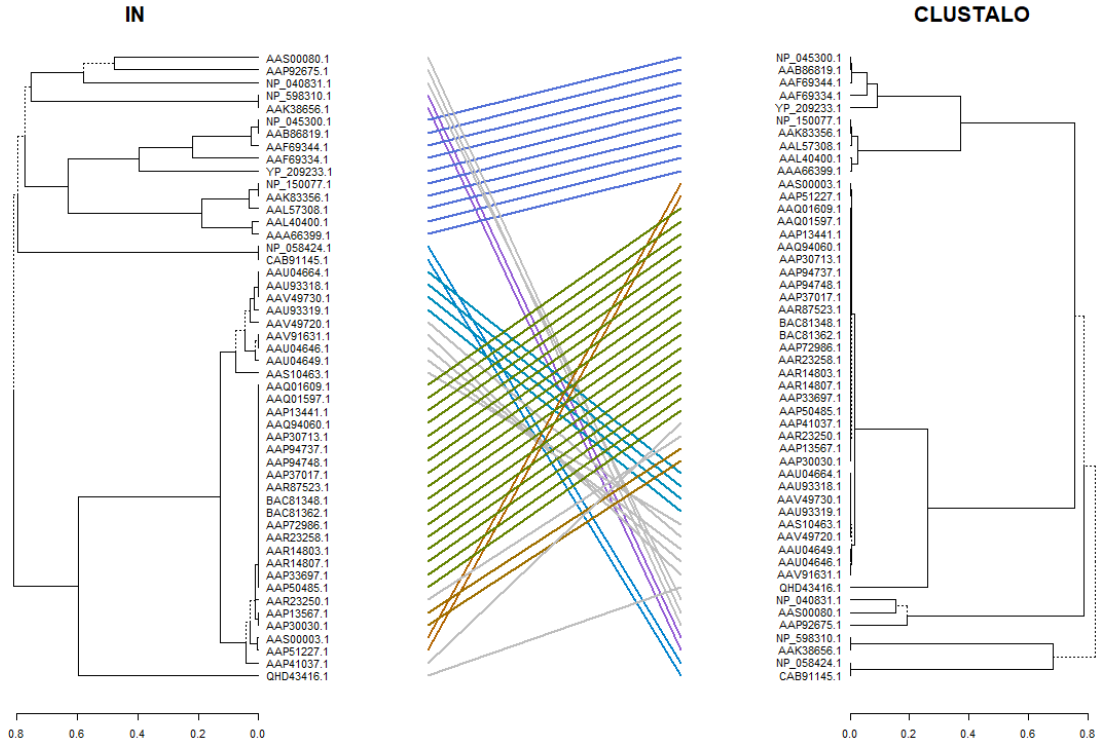
Слика додатак 6.1.25. Дендрограмски приказ секвенци *Spike* протеина корона вируса конструисано R-P/F методом за IN тип поновака. У листовима су приказани идентификатори секвенци, иза којих следе додатне информације о групи којој та секвенца припада.



Слика додатак 6.1.26. Дендрограмски приказ секвенци *Spike* протеина корона вируса конструисано UPGMA методом примењеном на матрицу растојања добијену из Clustal Omega програма.



Слика додатак 6.1.27. Компаративни приказ дендрограма секвенци *Spike* протеина корона вируса конструисано R-P/F методом за DN тип поновака и UPGMA методом примењеном на матрицу растојања добијену из Clustal Omega програма.

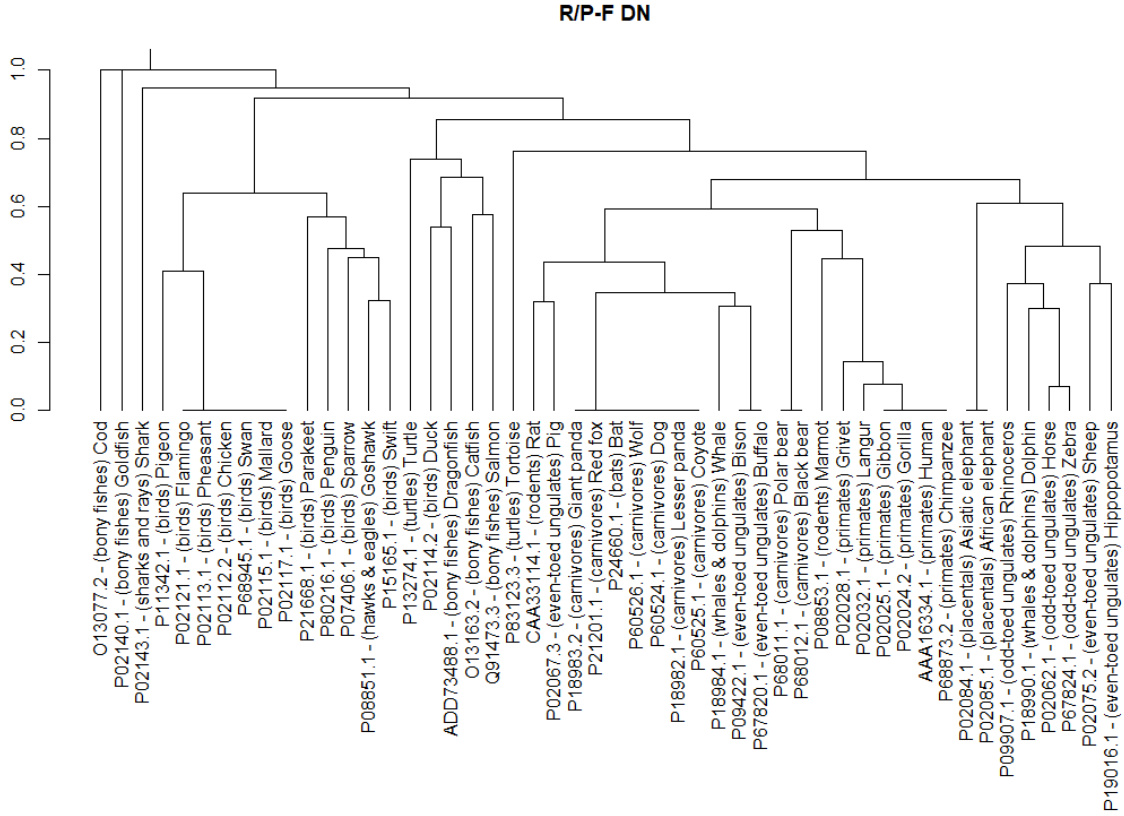


Слика додатак 6.1.28. Компаративни приказ дендрограма секвенци *Spike* протеина корона вируса конструисано R-P/F методом за IN тип поновака и UPGMA методом примењеном на матрицу растојања добијену из Clustal Omega програма.

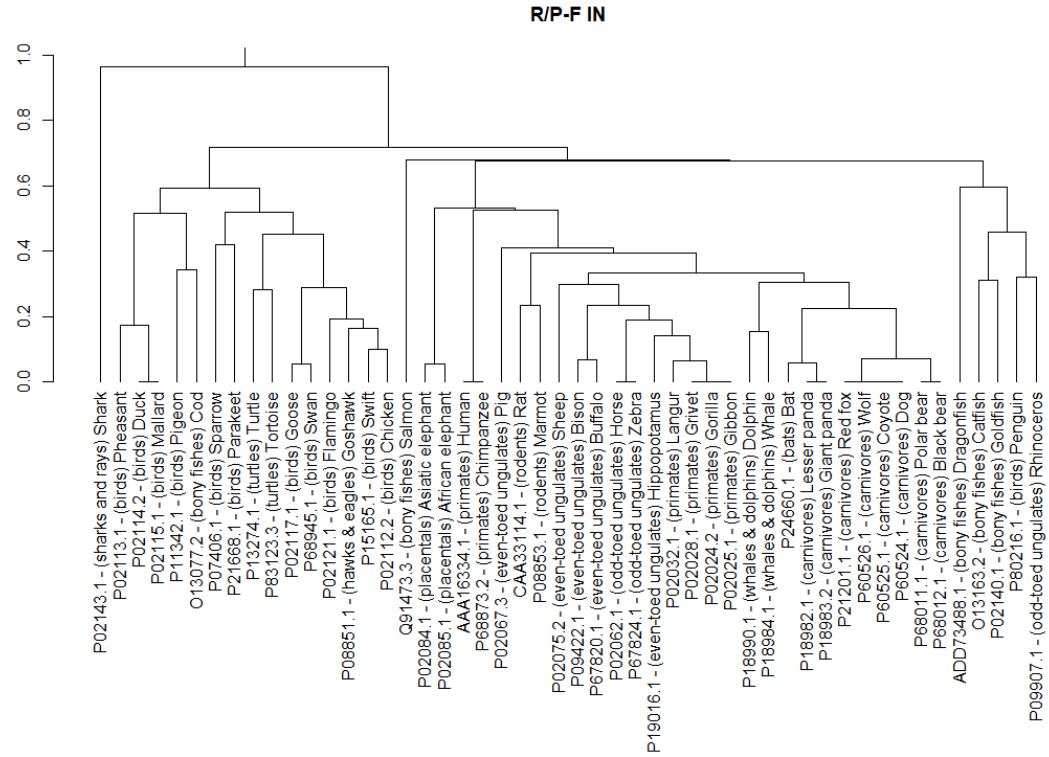
Табела додатак 6.1.7. Сажете информације о 50 протеинских секвенци протеина бета глобина

ID	Врста	Дужина
AAA16334.1	(primates) Human	147
P08851.1	(hawks & eagles) Goshawk	146
P18982.1	(carnivores) Lesser panda	146
P18983.2	(carnivores) Giant panda	147
P02075.2	(even-toed ungulates) Sheep	145
P02114.2	(birds) Duck	147
P02115.1	(birds) Mallard	146
P02117.1	(birds) Goose	146
CAA33114.1	(rodents) Rat	147
P80216.1	(birds) Penguin	146
P15165.1	(birds) Swift	146
P60525.1	(carnivores) Coyote	146
O13163.2	(bony fishes) Catfish	148
P09422.1	(even-toed ungulates) Bison	145
P68945.1	(birds) Swan	146
P67820.1	(even-toed ungulates) Buffalo	145
P60524.1	(carnivores) Dog	146
P68873.2	(primates) Chimpanzee	147
P18990.1	(whales & dolphins) Dolphin	146
P02140.1	(bony fishes) Goldfish	147
P68011.1	(carnivores) Polar bear	146
P09907.1	(odd-toed ungulates) Rhinoceros	146
P02112.2	(birds) Chicken	147
P60526.1	(carnivores) Wolf	146
P13274.1	(turtles) Turtle	146
P11342.1	(birds) Pigeon	146
P68012.1	(carnivores) Black bear	146
P02084.1	(placentals) Asiatic elephant	146
P02085.1	(placentals) African elephant	146
P83123.3	(turtles) Tortoise	147
P02028.1	(primates) Grivet	146
P02024.2	(primates) Gorilla	147
P02143.1	(sharks and rays) Shark	141
P19016.1	(even-toed ungulates) Hippopotamus	146
P02062.1	(odd-toed ungulates) Horse	146
P02025.1	(primates) Gibbon	146
P18984.1	(whales & dolphins) Whale	146
P24660.1	(bats) Bat	146
P21201.1	(carnivores) Red fox	146
P08853.1	(rodents) Marmot	146
Q91473.3	(bony fishes) Salmon	148
P07406.1	(birds) Sparrow	146
P02113.1	(birds) Pheasant	146

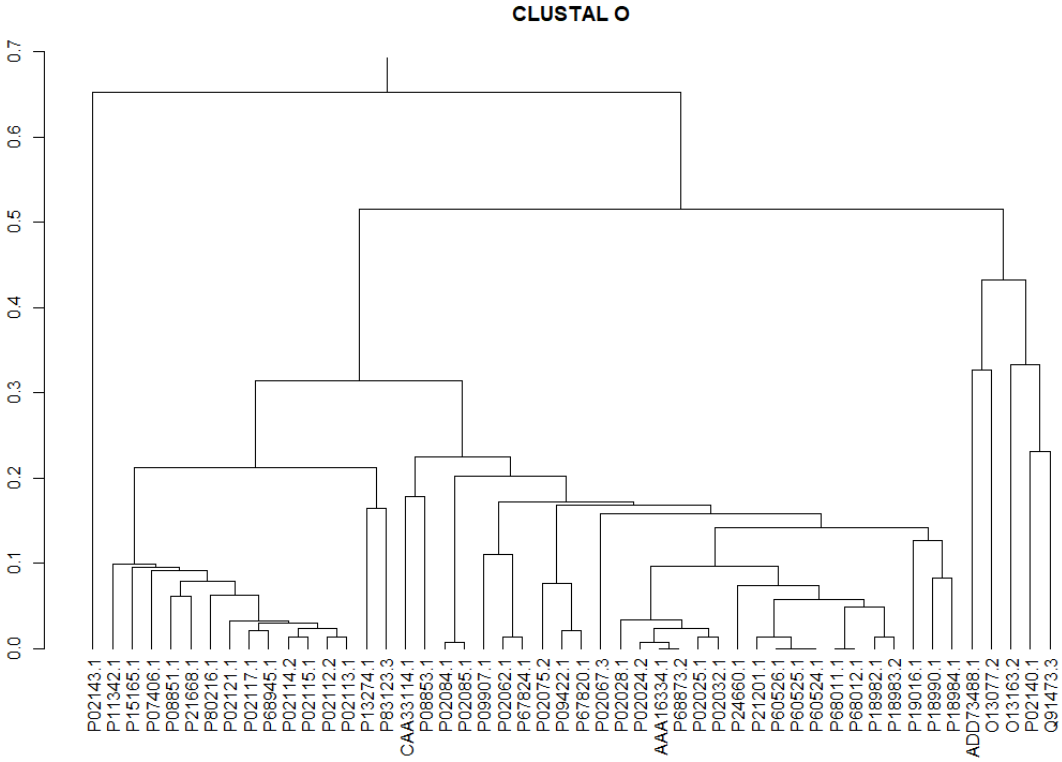
P02121.1	(birds) Flamingo	146
P02067.3	(even-toed ungulates) Pig	147
ADD73488.1	(bony fishes) Dragonfish	141
P21668.1	(birds) Parakeet	146
P67824.1	(odd-toed ungulates) Zebra	146
O13077.2	(bony fishes) Cod	147
P02032.1	(primates) Langur	146



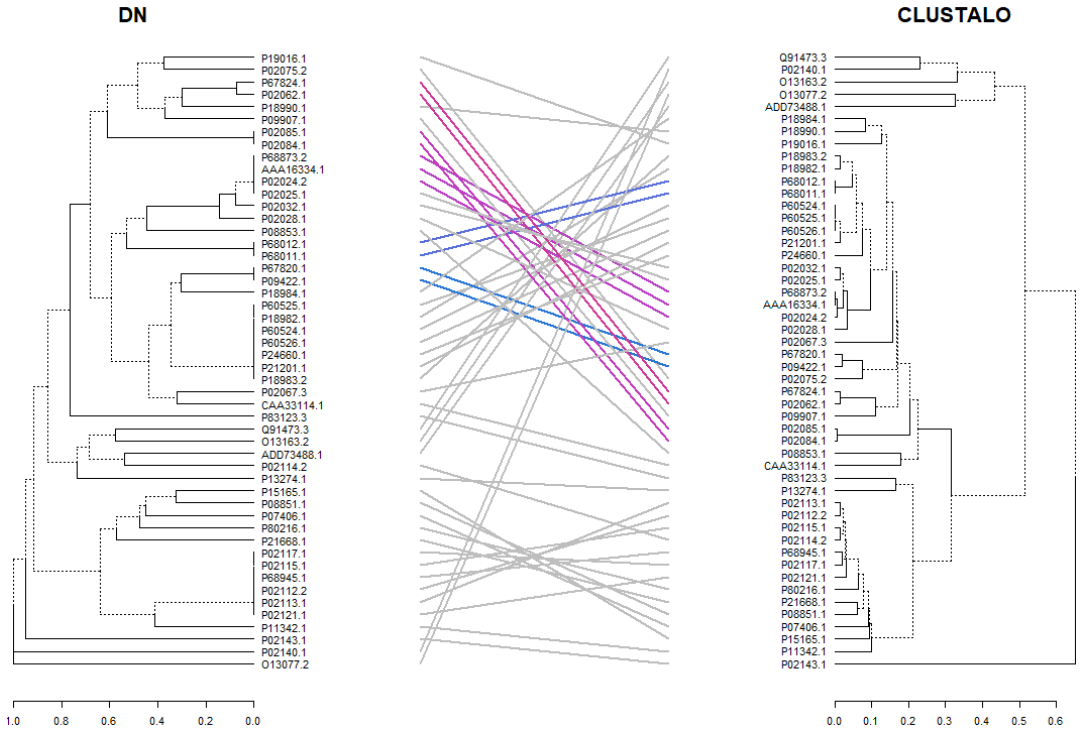
Слика додатак 6.1.29. Дендрограмски приказ секвенци бета глобина конструисано R-P/F методом за DN тип поновака. У листовима су приказани идентификатори секвенци, иза којих следе додатне информације о врсти којој та секвенца припада



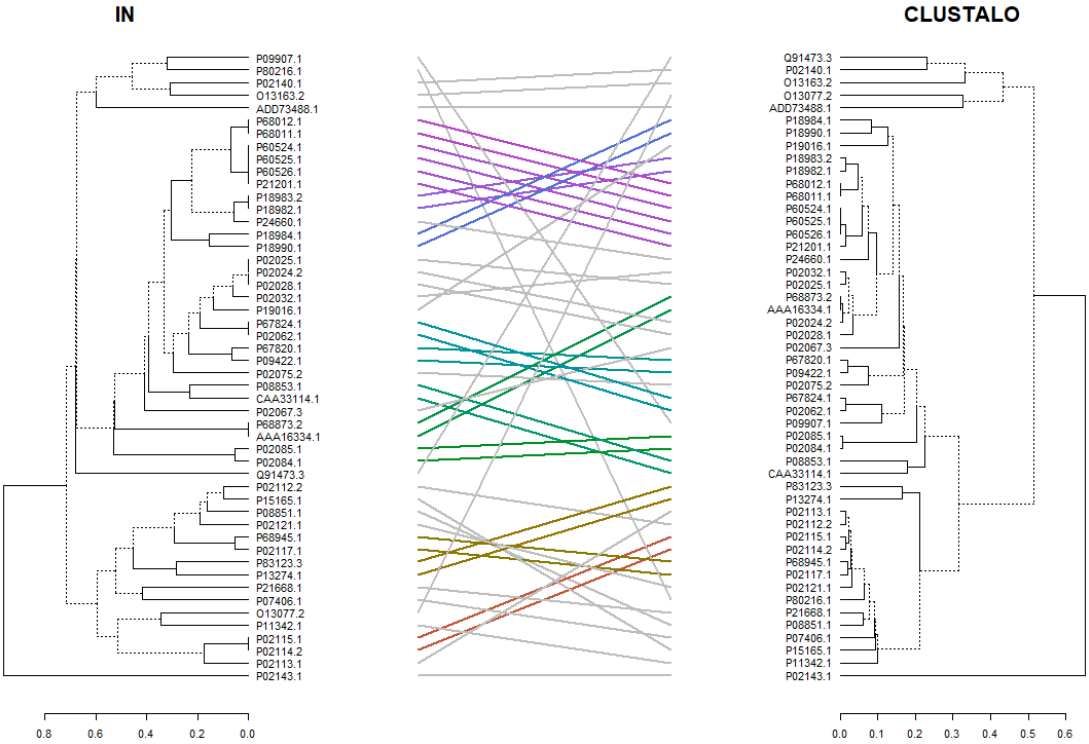
Слика додатак 6.1.30. Дендрограмски приказ секвенци бета глобина конструисано R-P/F методом за IN тип поновака. У листовима су приказани идентификатори секвенци, иза којих следе додатне информације о врсти којој та секвенца припада



Слика додатак 6.1.31. Дендрограмски приказ секвенци протеина бета глобина конструисано UPGMA методом примењеном на матрицу растојања добијену из Clustal Omega програма.



Слика додатак 6.1.32. Компаративни приказ дендрограма секвенци протеина бета глобина конструисано R-P/F методом за DN тип поновака и UPGMA методом примењеном на матрицу растојања добијену из Clustal Omega програма.



Слика додатак 6.1.33. Компаративни приказ дендрограма секвенци протеина бета глобина конструисано R-P/F методом за IN тип поновака и UPGMA методом примењеном на матрицу растојања добијену из Clustal Omega програма.

6.2 Додатак резултатима методе засноване на потписима секвенци и профилима категорије

Табела додатак 6.2.1. Вредности додељене категорије и сличности секвенци: KF514396.1, MF599507.1, JX458829.1 са профилом категорије коришћењем метода класификација секвенци на основу профила категорије применом различитих параметара модела. Слогови су подебљани у случају када је додељена вредност категорије једнака стварној вредности категорије.

Секвенца	Категорија	Тип понов ка	Јачина категорије	Права категорија	Додељена категорија	Косинус на сличност	Коефицијент Танимотоа
JX458829.1	**Ред**	**DC**	**50**	Mononegavirales	Blattodea	0.02318	0.01153
JX458829.1	**Ред**	**DC**	**75**	Mononegavirales	Blattodea	0.02519	0.01247
JX458829.1	**Ред**	**DC**	**90**	Mononegavirales	Blattodea	0.02527	0.01251
JX458829.1	**Ред**	**DC**	**95**	Mononegavirales	Blattodea	0.02527	0.01251
JX458829.1	**Ред**	**DN**	**50**	Mononegavirales	Blattodea	0.02242	0.01113
JX458829.1	**Ред**	**DN**	**75**	Mononegavirales	Blattodea	0.02484	0.01226
JX458829.1	**Ред**	**DN**	**90**	Mononegavirales	Blattodea	0.02497	0.01232
JX458829.1	**Ред**	**DN**	**95**	Mononegavirales	Blattodea	0.02497	0.01232
JX458829.1	**Ред**	**IC**	**50**	Mononegavirales	Blattodea	0.02304	0.01152
JX458829.1	**Ред**	**IC**	**75**	Mononegavirales	Blattodea	0.02508	0.01249
JX458829.1	**Ред**	**IC**	**90**	Mononegavirales	Blattodea	0.02513	0.01251
JX458829.1	**Ред**	**IC**	**95**	Mononegavirales	Blattodea	0.02513	0.01251
JX458829.1	**Ред**	**IN**	**50**	Mononegavirales	Blattodea	0.02402	0.01183
JX458829.1	**Ред**	**IN**	**75**	Mononegavirales	Blattodea	0.02665	0.01303
JX458829.1	**Ред**	**IN**	**90**	Mononegavirales	Blattodea	0.02673	0.01307
JX458829.1	**Ред**	**IN**	**95**	Mononegavirales	Blattodea	0.02674	0.01308
JX458829.1	**Род**	**DC**	**50**	**Marburgvirus**	**Marburgvirus**	**0.44500**	**0.20849**
JX458829.1	**Род**	**DC**	**75**	**Marburgvirus**	**Marburgvirus**	**0.33503**	**0.11496**
JX458829.1	**Род**	**DC**	**90**	**Marburgvirus**	**Marburgvirus**	**0.08688**	0.00776
JX458829.1	**Род**	**DC**	**95**	Marburgvirus	Blaptica		0.00517
JX458829.1	**Род**	**DC**	**95**	**Marburgvirus**	**Marburgvirus**	**0.06900**	
JX458829.1	**Род**	**DN**	**50**	**Marburgvirus**	**Marburgvirus**	**0.37961**	**0.15333**
JX458829.1	**Род**	**DN**	**75**	**Marburgvirus**	**Marburgvirus**	**0.28929**	**0.08544**
JX458829.1	**Род**	**DN**	**90**	**Marburgvirus**	**Marburgvirus**	**0.07485**	0.00572
JX458829.1	**Род**	**DN**	**95**	Marburgvirus	Alphamononivirus		0.00388
JX458829.1	**Род**	**DN**	**95**	**Marburgvirus**	**Marburgvirus**	**0.05851**	
JX458829.1	**Род**	**IC**	**50**	**Marburgvirus**	**Marburgvirus**	**0.45569**	**0.21795**
JX458829.1	**Род**	**IC**	**75**	**Marburgvirus**	**Marburgvirus**	**0.35512**	**0.12876**
JX458829.1	**Род**	**IC**	**90**	**Marburgvirus**	**Marburgvirus**	**0.09654**	0.00954
JX458829.1	**Род**	**IC**	**95**	**Marburgvirus**	**Marburgvirus**	**0.07538**	0.00569
JX458829.1	**Род**	**IN**	**50**	**Marburgvirus**	**Marburgvirus**	**0.41448**	**0.18220**
JX458829.1	**Род**	**IN**	**75**	**Marburgvirus**	**Marburgvirus**	**0.31243**	**0.09989**
JX458829.1	**Род**	**IN**	**90**	**Marburgvirus**	**Marburgvirus**	**0.07696**	0.00617
JX458829.1	**Род**	**IN**	**95**	Marburgvirus	Blaptica		0.00476
JX458829.1	**Род**	**IN**	**95**	**Marburgvirus**	**Marburgvirus**	**0.06014**	
JX458829.1	**Фамилија**	**DC**	**50**	Filoviridae	Lavidaviridae	0.01620	0.00795
JX458829.1	**Фамилија**	**DC**	**75**	Filoviridae	Blaberidae	0.01662	0.00835
JX458829.1	**Фамилија**	**DC**	**90**	Filoviridae	Blaberidae	0.01721	0.00864
JX458829.1	**Фамилија**	**DC**	**95**	Filoviridae	Blaberidae	0.01723	0.00865
JX458829.1	**Фамилија**	**DN**	**50**	Filoviridae	Lavidaviridae	0.01270	0.00627
JX458829.1	**Фамилија**	**DN**	**75**	Filoviridae	Mononiviridae	0.01953	0.00609
JX458829.1	**Фамилија**	**DN**	**90**	Filoviridae	Mononiviridae	0.02041	0.00634
JX458829.1	**Фамилија**	**DN**	**95**	Filoviridae	Mononiviridae	0.02046	0.00635
JX458829.1	**Фамилија**	**IC**	**50**	Filoviridae	Lavidaviridae	0.01801	0.00880
JX458829.1	**Фамилија**	**IC**	**75**	Filoviridae	Blaberidae	0.01725	0.00869

JX458829.1	**Фамилија**	**IC**	**90**	Filoviridae	Blaberidae	0.01779	0.00896
JX458829.1	**Фамилија**	**IC**	**95**	Filoviridae	Blaberidae	0.01780	0.00897
JX458829.1	**Фамилија**	**IN**	**50**	Filoviridae	Lavidaviridae	0.01483	0.00726
JX458829.1	**Фамилија**	**IN**	**75**	Filoviridae	Blaberidae	0.01649	0.00825
JX458829.1	**Фамилија**	**IN**	**90**	Filoviridae	Blaberidae	0.01713	0.00856
JX458829.1	**Фамилија**	**IN**	**95**	Filoviridae	Blaberidae	0.01713	0.00856
KF514396.1	**Ред**	**DC**	**50**	**Nidovirales**	**Nidovirales**	**0.90913**	**0.82662**
KF514396.1	**Ред**	**DC**	**75**	**Nidovirales**	**Nidovirales**	**0.22286**	0.04967
KF514396.1	**Ред**	**DC**	**90**	Nidovirales	Blattodea	0.01441	0.00709
KF514396.1	**Ред**	**DC**	**95**	Nidovirales	Blattodea	0.01441	0.00709
KF514396.1	**Ред**	**DN**	**50**	**Nidovirales**	**Nidovirales**	**0.89668**	**0.80434**
KF514396.1	**Ред**	**DN**	**75**	**Nidovirales**	**Nidovirales**	**0.25877**	**0.06696**
KF514396.1	**Ред**	**DN**	**90**	Nidovirales	Caudovirales	0.02065	0.00997
KF514396.1	**Ред**	**DN**	**95**	Nidovirales	Caudovirales	0.02066	0.00997
KF514396.1	**Ред**	**IC**	**50**	**Nidovirales**	**Nidovirales**	**0.90632**	**0.82153**
KF514396.1	**Ред**	**IC**	**75**	**Nidovirales**	**Nidovirales**	**0.22357**	0.04998
KF514396.1	**Ред**	**IC**	**90**	Nidovirales	Blattodea	0.01397	0.00683
KF514396.1	**Ред**	**IC**	**95**	Nidovirales	Blattodea	0.01397	0.00683
KF514396.1	**Ред**	**IN**	**50**	**Nidovirales**	**Nidovirales**	**0.90171**	**0.81319**
KF514396.1	**Ред**	**IN**	**75**	**Nidovirales**	**Nidovirales**	**0.22758**	**0.05179**
KF514396.1	**Ред**	**IN**	**90**	Nidovirales	Blattodea	0.01342	0.00667
KF514396.1	**Ред**	**IN**	**95**	Nidovirales	Blattodea	0.01343	0.00668
KF514396.1	**Род**	**DC**	**50**	**Betacoronavirus**	**Betacoronavirus**	**0.78580**	**0.61813**
KF514396.1	**Род**	**DC**	**75**	**Betacoronavirus**	**Betacoronavirus**	**0.76938**	**0.59197**
KF514396.1	**Род**	**DC**	**90**	**Betacoronavirus**	**Betacoronavirus**	**0.21397**	0.04578
KF514396.1	**Род**	**DC**	**95**	**Betacoronavirus**	**Betacoronavirus**	0.03605	
KF514396.1	**Род**	**DC**	**95**	Betacoronavirus	Paguronivirus		0.00503
KF514396.1	**Род**	**DN**	**50**	**Betacoronavirus**	**Betacoronavirus**	**0.72046**	**0.52132**
KF514396.1	**Род**	**DN**	**75**	**Betacoronavirus**	**Betacoronavirus**	**0.71879**	**0.51669**
KF514396.1	**Род**	**DN**	**90**	**Betacoronavirus**	**Betacoronavirus**	**0.21474**	0.04611
KF514396.1	**Род**	**DN**	**95**	Betacoronavirus	Alphamononivirus		0.00654
KF514396.1	**Род**	**DN**	**95**	**Betacoronavirus**	**Betacoronavirus**	**0.05704**	
KF514396.1	**Род**	**IC**	**50**	**Betacoronavirus**	**Betacoronavirus**	**0.77761**	**0.60538**
KF514396.1	**Род**	**IC**	**75**	**Betacoronavirus**	**Betacoronavirus**	**0.76829**	**0.59030**
KF514396.1	**Род**	**IC**	**90**	**Betacoronavirus**	**Betacoronavirus**	**0.21445**	0.04599
KF514396.1	**Род**	**IC**	**95**	**Betacoronavirus**	**Betacoronavirus**	0.03444	
KF514396.1	**Род**	**IC**	**95**	Betacoronavirus	Paguronivirus		0.00517
KF514396.1	**Род**	**IN**	**50**	**Betacoronavirus**	**Betacoronavirus**	**0.74802**	**0.56045**
KF514396.1	**Род**	**IN**	**75**	**Betacoronavirus**	**Betacoronavirus**	**0.74238**	**0.55115**
KF514396.1	**Род**	**IN**	**90**	**Betacoronavirus**	**Betacoronavirus**	**0.20437**	0.04177
KF514396.1	**Род**	**IN**	**95**	**Betacoronavirus**	**Betacoronavirus**	0.03871	
KF514396.1	**Род**	**IN**	**95**	Betacoronavirus	Bostovirus		0.00680
KF514396.1	**Фамилија**	**DC**	**50**	**Coronaviridae**	**Coronaviridae**	**0.87822**	**0.77158**
KF514396.1	**Фамилија**	**DC**	**75**	**Coronaviridae**	**Coronaviridae**	**0.68497**	**0.46919**
KF514396.1	**Фамилија**	**DC**	**90**	**Coronaviridae**	**Coronaviridae**	0.01862	
KF514396.1	**Фамилија**	**DC**	**90**	Coronaviridae	Ovaliviridae		0.00587
KF514396.1	**Фамилија**	**DC**	**95**	Coronaviridae	Ovaliviridae	0.01186	0.00587
KF514396.1	**Фамилија**	**DN**	**50**	**Coronaviridae**	**Coronaviridae**	**0.83446**	**0.69772**
KF514396.1	**Фамилија**	**DN**	**75**	**Coronaviridae**	**Coronaviridae**	**0.65795**	**0.43291**
KF514396.1	**Фамилија**	**DN**	**90**	**Coronaviridae**	**Coronaviridae**	**0.06122**	
KF514396.1	**Фамилија**	**DN**	**90**	Coronaviridae	Mononiviridae		0.00886
KF514396.1	**Фамилија**	**DN**	**95**	Coronaviridae	Mononiviridae	0.02117	0.00896
KF514396.1	**Фамилија**	**IC**	**50**	**Coronaviridae**	**Coronaviridae**	**0.87531**	**0.76648**
KF514396.1	**Фамилија**	**IC**	**75**	**Coronaviridae**	**Coronaviridae**	**0.68658**	**0.47140**

KF514396.1	**Фамилија**	**IC**	**90**	**Coronaviridae**	**Coronaviridae**	0.01671	
KF514396.1	**Фамилија**	**IC**	**90**	Coronaviridae	Ovaliviridae		0.00493
KF514396.1	**Фамилија**	**IC**	**95**	Coronaviridae	Blaberidae	0.01046	
KF514396.1	**Фамилија**	**IC**	**95**	Coronaviridae	Ovaliviridae		0.00494
KF514396.1	**Фамилија**	**IN**	**50**	**Coronaviridae**	**Coronaviridae**	**0.84899**	**0.72119**
KF514396.1	**Фамилија**	**IN**	**75**	**Coronaviridae**	**Coronaviridae**	**0.66987**	**0.44873**
KF514396.1	**Фамилија**	**IN**	**90**	**Coronaviridae**	**Coronaviridae**	0.03317	
KF514396.1	**Фамилија**	**IN**	**90**	Coronaviridae	Ovaliviridae		0.00539
KF514396.1	**Фамилија**	**IN**	**95**	Coronaviridae	Ovaliviridae	0.01094	0.00540
MF599507.1	**Ред**	**DC**	**50**	**Mononegavirales**	**Mononegavirales**	**0.68126**	**0.46559**
MF599507.1	**Ред**	**DC**	**75**	Mononegavirales	Blattodea	0.01992	0.00928
MF599507.1	**Ред**	**DC**	**90**	Mononegavirales	Blattodea	0.01997	0.00930
MF599507.1	**Ред**	**DC**	**95**	Mononegavirales	Blattodea	0.01997	0.00930
MF599507.1	**Ред**	**DN**	**50**	**Mononegavirales**	**Mononegavirales**	**0.67292**	**0.45315**
MF599507.1	**Ред**	**DN**	**75**	Mononegavirales	Blattodea	0.02030	0.00966
MF599507.1	**Ред**	**DN**	**90**	Mononegavirales	Blattodea	0.02045	0.00972
MF599507.1	**Ред**	**DN**	**95**	Mononegavirales	Blattodea	0.02045	0.00972
MF599507.1	**Ред**	**IC**	**50**	**Mononegavirales**	**Mononegavirales**	**0.70925**	**0.50391**
MF599507.1	**Ред**	**IC**	**75**	Mononegavirales	Blattodea	0.01901	0.00913
MF599507.1	**Ред**	**IC**	**90**	Mononegavirales	Blattodea	0.01907	0.00915
MF599507.1	**Ред**	**IC**	**95**	Mononegavirales	Blattodea	0.01907	0.00915
MF599507.1	**Ред**	**IN**	**50**	**Mononegavirales**	**Mononegavirales**	**0.65804**	**0.43472**
MF599507.1	**Ред**	**IN**	**75**	Mononegavirales	Blattodea	0.02151	0.01010
MF599507.1	**Ред**	**IN**	**90**	Mononegavirales	Blattodea	0.02167	0.01017
MF599507.1	**Ред**	**IN**	**95**	Mononegavirales	Blattodea	0.02168	0.01017
MF599507.1	**Род**	**DC**	**50**	**Ebolavirus**	**Ebolavirus**	**0.62694**	**0.39853**
MF599507.1	**Род**	**DC**	**75**	**Ebolavirus**	**Ebolavirus**	**0.54894**	**0.30304**
MF599507.1	**Род**	**DC**	**90**	**Ebolavirus**	**Ebolavirus**	**0.18380**	0.03378
MF599507.1	**Род**	**DC**	**95**	Ebolavirus	Blaptica	0.00869	0.00436
MF599507.1	**Род**	**DN**	**50**	**Ebolavirus**	**Ebolavirus**	**0.56017**	**0.31606**
MF599507.1	**Род**	**DN**	**75**	**Ebolavirus**	**Ebolavirus**	**0.48585**	**0.23642**
MF599507.1	**Род**	**DN**	**90**	**Ebolavirus**	**Ebolavirus**	**0.15738**	0.02477
MF599507.1	**Род**	**DN**	**95**	**Ebolavirus**	**Ebolavirus**	0.00850	
MF599507.1	**Род**	**DN**	**95**	Ebolavirus	Infratovirus		0.00243
MF599507.1	**Род**	**IC**	**50**	**Ebolavirus**	**Ebolavirus**	**0.62347**	**0.39296**
MF599507.1	**Род**	**IC**	**75**	**Ebolavirus**	**Ebolavirus**	**0.56486**	**0.32054**
MF599507.1	**Род**	**IC**	**90**	**Ebolavirus**	**Ebolavirus**	**0.20103**	0.04041
MF599507.1	**Род**	**IC**	**95**	Ebolavirus	Blaptica		0.00377
MF599507.1	**Род**	**IC**	**95**	**Ebolavirus**	**Ebolavirus**	0.01430	
MF599507.1	**Род**	**IN**	**50**	**Ebolavirus**	**Ebolavirus**	**0.55124**	**0.31111**
MF599507.1	**Род**	**IN**	**75**	**Ebolavirus**	**Ebolavirus**	**0.48738**	**0.23954**
MF599507.1	**Род**	**IN**	**90**	**Ebolavirus**	**Ebolavirus**	**0.15278**	0.02334
MF599507.1	**Род**	**IN**	**95**	Ebolavirus	Blaptica	0.00718	0.00360
MF599507.1	**Фамилија**	**DC**	**50**	**Filoviridae**	**Filoviridae**	**0.65515**	**0.43449**
MF599507.1	**Фамилија**	**DC**	**75**	**Filoviridae**	**Filoviridae**	**0.54706**	**0.29938**
MF599507.1	**Фамилија**	**DC**	**90**	Filoviridae	Blaberidae		0.00675
MF599507.1	**Фамилија**	**DC**	**90**	**Filoviridae**	**Filoviridae**	0.01480	
MF599507.1	**Фамилија**	**DC**	**95**	Filoviridae	Blaberidae	0.01398	0.00676
MF599507.1	**Фамилија**	**DN**	**50**	**Filoviridae**	**Filoviridae**	**0.60638**	**0.36953**
MF599507.1	**Фамилија**	**DN**	**75**	**Filoviridae**	**Filoviridae**	**0.52433**	**0.27492**
MF599507.1	**Фамилија**	**DN**	**90**	Filoviridae	Blaberidae		0.00464
MF599507.1	**Фамилија**	**DN**	**90**	Filoviridae	Mononiviridae	0.01477	
MF599507.1	**Фамилија**	**DN**	**95**	Filoviridae	Blaberidae		0.00465
MF599507.1	**Фамилија**	**DN**	**95**	Filoviridae	Mononiviridae	0.01480	

MF599507.1	**Фамилија**	**IC**	**50**	**Filoviridae**	**Filoviridae**	**0.67425**	**0.45809**
MF599507.1	**Фамилија**	**IC**	**75**	**Filoviridae**	**Filoviridae**	**0.59424**	**0.35314**
MF599507.1	**Фамилија**	**IC**	**90**	Filoviridae	Blaberidae		0.00637
MF599507.1	**Фамилија**	**IC**	**90**	**Filoviridae**	**Filoviridae**	0.01837	
MF599507.1	**Фамилија**	**IC**	**95**	Filoviridae	Blaberidae	0.01292	0.00638
MF599507.1	**Фамилија**	**IN**	**50**	**Filoviridae**	**Filoviridae**	**0.59320**	**0.35866**
MF599507.1	**Фамилија**	**IN**	**75**	**Filoviridae**	**Filoviridae**	**0.52886**	**0.27970**
MF599507.1	**Фамилија**	**IN**	**90**	Filoviridae	Blaberidae	0.01360	0.00663
MF599507.1	**Фамилија**	**IN**	**95**	Filoviridae	Blaberidae	0.01360	0.00663

Табела додатак 6.2.2. Вредности мера перформанси израчунавања класификације приликом примене косинусног растојања

Тип поновка	Категорија	Јачина категорије	макро прецизност	макро одзив	макро ф1	микро прецизност, одзив, ф1 мера	тежинска прецизност	тежински одзив	тежинска ф1	губитак
DC	ред	50	0.06880	0.07176	0.06265	0.38158	0.48429	0.38158	0.37374	0.02290
DC	род	50	0.01623	0.01691	0.01637	0.31933	0.34034	0.31933	0.32783	0.00147
DC	род	90	0.01285	0.01285	0.01285	0.31933	0.31933	0.31933	0.31933	0.00147
DC	род	75	0.01475	0.01488	0.01478	0.31933	0.31565	0.31933	0.31587	0.00147
DC	фамилија	50	0.02084	0.03742	0.02300	0.25210	0.26773	0.25210	0.24609	0.00613
DC	фамилија	90	0.01880	0.01878	0.01818	0.20168	0.26090	0.20168	0.22285	0.00654
DC	фамилија	75	0.01710	0.02103	0.01732	0.23529	0.25029	0.23529	0.21759	0.00627
DC	род	95	0.01067	0.01051	0.01058	0.18487	0.18547	0.18487	0.18471	0.00176
DC	ред	75	0.05797	0.05291	0.05072	0.18421	0.13043	0.18421	0.14699	0.03021
DC	фамилија	95	0.00546	0.00820	0.00656	0.01681	0.01120	0.01681	0.01345	0.00806
DC	ред	90	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.03704
DC	ред	95	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.03704
DN	ред	50	0.09612	0.08912	0.08760	0.38158	0.54010	0.38158	0.41167	0.02290
DN	род	50	0.02706	0.02593	0.02604	0.36975	0.40756	0.36975	0.38441	0.00136
DN	род	75	0.01732	0.01691	0.01709	0.31933	0.34454	0.31933	0.33063	0.00147
DN	род	90	0.01488	0.01502	0.01494	0.32500	0.32604	0.32500	0.32514	0.00147
DN	род	95	0.01500	0.01475	0.01486	0.30579	0.32113	0.30579	0.31295	0.00153
DN	фамилија	50	0.03125	0.02922	0.02966	0.24370	0.30725	0.24370	0.26274	0.00620
DN	фамилија	90	0.02231	0.01711	0.01638	0.14286	0.28898	0.14286	0.13935	0.00703
DN	фамилија	75	0.01995	0.02103	0.01907	0.23333	0.25528	0.23333	0.22060	0.00634
DN	ред	75	0.02293	0.03439	0.02751	0.17105	0.11404	0.17105	0.13684	0.03070
DN	фамилија	95	0.01230	0.00883	0.00929	0.02521	0.07143	0.02521	0.03137	0.00799
DN	ред	90	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.03704
DN	ред	95	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.03704
IC	ред	50	0.06723	0.07755	0.06763	0.44737	0.45999	0.44737	0.42407	0.02047
IC	род	90	0.02128	0.02367	0.02194	0.36134	0.35994	0.36134	0.35841	0.00138

IC	род	95	0.02128	0.02337	0.02177	0.35294	0.35994	0.35294	0.35388	0.00140
IC	род	50	0.02237	0.02340	0.02250	0.33884	0.35813	0.33884	0.34583	0.00146
IC	род	75	0.02424	0.02570	0.02459	0.35537	0.35207	0.35537	0.35128	0.00142
IC	фамилија	50	0.02771	0.03742	0.02770	0.25210	0.30212	0.25210	0.26381	0.00613
IC	фамилија	90	0.02294	0.02075	0.02165	0.23529	0.28676	0.23529	0.25641	0.00627
IC	фамилија	75	0.01940	0.02138	0.01919	0.24370	0.24454	0.24370	0.22528	0.00620
IC	ред	75	0.02534	0.03439	0.02918	0.17105	0.12604	0.17105	0.14514	0.03070
IC	фамилија	95	0.00820	0.00820	0.00820	0.01681	0.01681	0.01681	0.01681	0.00806
IC	ред	90	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.03704
IC	ред	95	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.03704
IN	ред	50	0.09171	0.07523	0.07808	0.42105	0.48997	0.42105	0.43502	0.02144
IN	род	50	0.02273	0.02340	0.02286	0.34454	0.36555	0.34454	0.35304	0.00142
IN	род	75	0.01587	0.01691	0.01601	0.31667	0.33611	0.31667	0.32371	0.00149
IN	род	90	0.01371	0.01502	0.01400	0.32773	0.33053	0.32773	0.32760	0.00146
IN	фамилија	50	0.03314	0.03742	0.03119	0.25210	0.32810	0.25210	0.27474	0.00613
IN	фамилија	90	0.02295	0.01683	0.01711	0.14286	0.29748	0.14286	0.15505	0.00703
IN	фамилија	75	0.02124	0.02103	0.02021	0.23529	0.27464	0.23529	0.23755	0.00627
IN	род	95	0.01139	0.01299	0.01182	0.20168	0.18828	0.20168	0.19343	0.00173
IN	ред	75	0.02832	0.03439	0.03106	0.17105	0.14087	0.17105	0.15450	0.03070
IN	фамилија	95	0.00820	0.00820	0.00820	0.01681	0.01681	0.01681	0.01681	0.00806
IN	ред	90	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.03704
IN	ред	95	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.03704

Табела додатак 6.2.3. Вредности мера перформанси израчунавања класификације приликом примене коефицијента Танимотоа

Тип поновка	Категорија	Јачина категорије	макро прецизност	макро одзив	макро ф1	микро прецизност, одзив, ф1 мера	тежинска прецизност	тежински одзив	тежинска ф1	губитак
DC	фамилија	50	0.02084	0.03742	0.02300	0.25210	0.26773	0.25210	0.24609	0.00613
DC	фамилија	75	0.01710	0.02103	0.01732	0.23529	0.25029	0.23529	0.21759	0.00627
DC	фамилија	90	0.01880	0.01878	0.01818	0.20168	0.26090	0.20168	0.22285	0.00654
DC	фамилија	95	0.00546	0.00820	0.00656	0.01681	0.01120	0.01681	0.01345	0.00806
DN	фамилија	75	0.01995	0.02103	0.01907	0.23333	0.25528	0.23333	0.22060	0.00634
DN	фамилија	90	0.02231	0.01711	0.01638	0.14286	0.28898	0.14286	0.13935	0.00703
DN	фамилија	95	0.01230	0.00883	0.00929	0.02521	0.07143	0.02521	0.03137	0.00799
DN	фамилија	50	0.03125	0.02922	0.02966	0.24370	0.30725	0.24370	0.26274	0.00620
IC	фамилија	50	0.02771	0.03742	0.02770	0.25210	0.30212	0.25210	0.26381	0.00613
IC	фамилија	75	0.01940	0.02138	0.01919	0.24370	0.24454	0.24370	0.22528	0.00620
IC	фамилија	90	0.02294	0.02075	0.02165	0.23529	0.28676	0.23529	0.25641	0.00627
IC	фамилија	95	0.00820	0.00820	0.00820	0.01681	0.01681	0.01681	0.01681	0.00806
IN	фамилија	50	0.03314	0.03742	0.03119	0.25210	0.32810	0.25210	0.27474	0.00613
IN	фамилија	75	0.02124	0.02103	0.02021	0.23529	0.27464	0.23529	0.23755	0.00627
IN	фамилија	90	0.02295	0.01683	0.01711	0.14286	0.29748	0.14286	0.15505	0.00703
IN	фамилија	95	0.00820	0.00820	0.00820	0.01681	0.01681	0.01681	0.01681	0.00806
DC	род	75	0.01475	0.01488	0.01478	0.31933	0.31565	0.31933	0.31587	0.00147
DC	род	90	0.01285	0.01285	0.01285	0.31933	0.31933	0.31933	0.31933	0.00147
DC	род	95	0.01067	0.01051	0.01058	0.18487	0.18547	0.18487	0.18471	0.00176
DC	род	50	0.01623	0.01691	0.01637	0.31933	0.34034	0.31933	0.32783	0.00147
DN	род	50	0.02706	0.02593	0.02604	0.36975	0.40756	0.36975	0.38441	0.00136
DN	род	75	0.01732	0.01691	0.01709	0.31933	0.34454	0.31933	0.33063	0.00147
DN	род	90	0.01488	0.01502	0.01494	0.32500	0.32604	0.32500	0.32514	0.00147
DN	род	95	0.01500	0.01475	0.01486	0.30579	0.32113	0.30579	0.31295	0.00153
IC	род	75	0.02424	0.02570	0.02459	0.35537	0.35207	0.35537	0.35128	0.00142
IC	род	90	0.02128	0.02367	0.02194	0.36134	0.35994	0.36134	0.35841	0.00138
IC	род	95	0.02128	0.02337	0.02177	0.35294	0.35994	0.35294	0.35388	0.00140
IC	род	50	0.02237	0.02340	0.02250	0.33884	0.35813	0.33884	0.34583	0.00146
IN	род	75	0.01587	0.01691	0.01601	0.31667	0.33611	0.31667	0.32371	0.00149

IN	род	90	0.01371	0.01502	0.01400	0.32773	0.33053	0.32773	0.32760	0.00146
IN	род	95	0.01139	0.01299	0.01182	0.20168	0.18828	0.20168	0.19343	0.00173
IN	род	50	0.02273	0.02340	0.02286	0.34454	0.36555	0.34454	0.35304	0.00142
DC	ред	75	0.05797	0.05291	0.05072	0.18421	0.13043	0.18421	0.14699	0.03021
DC	ред	90	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.03704
DC	ред	95	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.03704
DC	ред	50	0.06880	0.07176	0.06265	0.38158	0.48429	0.38158	0.37374	0.02290
DN	ред	75	0.02293	0.03439	0.02751	0.17105	0.11404	0.17105	0.13684	0.03070
DN	ред	90	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.03704
DN	ред	95	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.03704
DN	ред	50	0.09612	0.08912	0.08760	0.38158	0.54010	0.38158	0.41167	0.02290
IC	ред	50	0.06723	0.07755	0.06763	0.44737	0.45999	0.44737	0.42407	0.02047
IC	ред	75	0.02534	0.03439	0.02918	0.17105	0.12604	0.17105	0.14514	0.03070
IC	ред	90	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.03704
IC	ред	95	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.03704
IN	ред	50	0.09171	0.07523	0.07808	0.42105	0.48997	0.42105	0.43502	0.02144
IN	ред	75	0.02832	0.03439	0.03106	0.17105	0.14087	0.17105	0.15450	0.03070
IN	ред	90	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.03704
IN	ред	95	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.03704

Табела додатак 6.2.4. Резултати класификације методе засноване на профилима категорија приликом примене косинусног растојања као мере сличности за секвенцу NC_045512.2. Слогови су подебљани у случају када је додељена вредност категорије једнака стварној вредности категорије.

Тип поновка	Категорија	Јачина категорије	Оригинална категорија	Додељена категорија	Вредност сличности
DN	ред	75	**Nidovirales**	**Nidovirales**	0.07290
IN	фамилија	75	**Coronaviridae**	**Coronaviridae**	0.06800
DC	фамилија	75	**Coronaviridae**	**Coronaviridae**	0.06658
IN	ред	75	**Nidovirales**	**Nidovirales**	0.06637
DN	фамилија	75	**Coronaviridae**	**Coronaviridae**	0.06636
DN	ред	50	**Nidovirales**	**Nidovirales**	0.06596
IN	ред	50	**Nidovirales**	**Nidovirales**	0.06432
DC	ред	50	**Nidovirales**	**Nidovirales**	0.06267
IC	фамилија	75	**Coronaviridae**	**Coronaviridae**	0.06232
DC	ред	75	**Nidovirales**	**Nidovirales**	0.05987
IC	ред	50	**Nidovirales**	**Nidovirales**	0.05911
IC	ред	75	**Nidovirales**	**Nidovirales**	0.05481
IN	род	90	**Betacoronavirus**	**Betacoronavirus**	0.05077
DN	род	90	**Betacoronavirus**	**Betacoronavirus**	0.04936
DN	фамилија	50	**Coronaviridae**	**Coronaviridae**	0.04903
DC	фамилија	50	**Coronaviridae**	**Coronaviridae**	0.04886
IN	фамилија	50	**Coronaviridae**	**Coronaviridae**	0.04877
DC	род	90	**Betacoronavirus**	**Betacoronavirus**	0.04734
IC	фамилија	50	**Coronaviridae**	**Coronaviridae**	0.04664
DC	род	75	**Betacoronavirus**	**Betacoronavirus**	0.04640
IN	род	75	**Betacoronavirus**	**Betacoronavirus**	0.04552
IC	род	90	**Betacoronavirus**	**Betacoronavirus**	0.04421
IC	род	75	**Betacoronavirus**	**Betacoronavirus**	0.04362
DN	род	75	**Betacoronavirus**	**Betacoronavirus**	0.04328
DC	род	50	**Betacoronavirus**	**Betacoronavirus**	0.03564
IN	род	50	**Betacoronavirus**	**Betacoronavirus**	0.03445
IC	род	50	**Betacoronavirus**	**Betacoronavirus**	0.03407
DN	род	50	**Betacoronavirus**	**Betacoronavirus**	0.03185
DN	фамилија	90	**Coronaviridae**	**Coronaviridae**	0.03033
DN	фамилија	95	Coronaviridae	Mononiviridae	0.02829
DN	род	95	**Betacoronavirus**	**Betacoronavirus**	0.02626
DN	ред	95	Nidovirales	Caudovirales	0.02345
DN	ред	90	Nidovirales	Caudovirales	0.02345
DC	ред	90	Nidovirales	Blattodea	0.02107
DC	ред	95	Nidovirales	Blattodea	0.02107
IN	ред	95	Nidovirales	Blattodea	0.02077
IN	ред	90	Nidovirales	Blattodea	0.02077
IC	ред	90	Nidovirales	Blattodea	0.02036
IC	ред	95	Nidovirales	Blattodea	0.02036
IN	род	95	**Betacoronavirus**	**Betacoronavirus**	0.02024
IN	фамилија	90	**Coronaviridae**	**Coronaviridae**	0.01822

IC	род	95	**Betacoronavirus**	**Betacoronavirus**	0.01658
DC	род	95	**Betacoronavirus**	**Betacoronavirus**	0.01625
DC	фамилија	95	Coronaviridae	Blaberidae	0.01585
DC	фамилија	90	Coronaviridae	Blaberidae	0.01584
IC	фамилија	95	Coronaviridae	Blaberidae	0.01524
IC	фамилија	90	Coronaviridae	Blaberidae	0.01523
IN	фамилија	95	Coronaviridae	Ovaliviridae	0.01465

Табела додатак 6.2.5. Резултати класификације методе засноване на профилима категорија приликом примене коефицијента Танимотоа као мере сличности за секвенцу NC_045512.2. Слогови су подебљани у случају када је додељена вредност категорије једнака стварној вредности категорије.

Тип поновка	**Категорија**	**Јачина категорије**	**Оригинална категорија**	**Додељена категорија**	**Вредност сличности**
DN	ред	50	**Nidovirales**	**Nidovirales**	0.03349
IN	ред	50	**Nidovirales**	**Nidovirales**	0.03257
DC	ред	50	**Nidovirales**	**Nidovirales**	0.03169
IN	фамилија	75	**Coronaviridae**	**Coronaviridae**	0.03115
DC	фамилија	75	**Coronaviridae**	**Coronaviridae**	0.03064
DN	фамилија	75	**Coronaviridae**	**Coronaviridae**	0.03033
IC	ред	50	**Nidovirales**	**Nidovirales**	0.02991
IC	фамилија	75	**Coronaviridae**	**Coronaviridae**	0.02887
DC	фамилија	50	**Coronaviridae**	**Coronaviridae**	0.02435
DN	фамилија	50	**Coronaviridae**	**Coronaviridae**	0.02432
IN	фамилија	50	**Coronaviridae**	**Coronaviridae**	0.02417
IC	фамилија	50	**Coronaviridae**	**Coronaviridae**	0.02329
DC	род	75	**Betacoronavirus**	**Betacoronavirus**	0.02219
IN	род	75	**Betacoronavirus**	**Betacoronavirus**	0.02156
IC	род	75	**Betacoronavirus**	**Betacoronavirus**	0.02095
DN	род	75	**Betacoronavirus**	**Betacoronavirus**	0.02034
DC	род	50	**Betacoronavirus**	**Betacoronavirus**	0.01710
DN	ред	75	**Nidovirales**	**Nidovirales**	0.01674
IC	род	50	**Betacoronavirus**	**Betacoronavirus**	0.01637
IN	род	50	**Betacoronavirus**	**Betacoronavirus**	0.01630
DN	род	50	**Betacoronavirus**	**Betacoronavirus**	0.01496
IN	ред	75	**Nidovirales**	**Nidovirales**	0.01334
DN	фамилија	95	Coronaviridae	Mononiviridae	0.01253
DN	фамилија	90	Coronaviridae	Mononiviridae	0.01247
DC	ред	75	**Nidovirales**	**Nidovirales**	0.01167
DN	ред	95	Nidovirales	Caudovirales	0.01157
DN	ред	90	Nidovirales	Caudovirales	0.01157
IC	ред	75	**Nidovirales**	**Nidovirales**	0.01087
IN	ред	95	Nidovirales	Blattodea	0.01017
IN	ред	90	Nidovirales	Blattodea	0.01016
DC	ред	90	Nidovirales	Blattodea	0.01011

DC	ред	95	Nidovirales	Blattodea	0.01011
IC	ред	90	Nidovirales	Blattodea	0.00973
IC	ред	95	Nidovirales	Blattodea	0.00973
DN	род	90	**Betacoronavirus**	**Betacoronavirus**	0.00950
DN	род	95	Betacoronavirus	Alphamononivirus	0.00927
IN	род	90	**Betacoronavirus**	**Betacoronavirus**	0.00920
IN	род	95	Betacoronavirus	Bostovirus	0.00886
DC	род	90	**Betacoronavirus**	**Betacoronavirus**	0.00886
IC	род	90	**Betacoronavirus**	**Betacoronavirus**	0.00842
DC	фамилија	90	Coronaviridae	Ovaliviridae	0.00741
DC	фамилија	95	Coronaviridae	Ovaliviridae	0.00741
IC	фамилија	95	Coronaviridae	Ovaliviridae	0.00741
IC	фамилија	90	Coronaviridae	Ovaliviridae	0.00740
IN	фамилија	95	Coronaviridae	Ovaliviridae	0.00707
IN	фамилија	90	Coronaviridae	Ovaliviridae	0.00707
IC	род	95	Betacoronavirus	Alphaovalivirus	0.00562
DC	род	95	Betacoronavirus	Paguronivirus	0.00546

Литература

1 Singh G. Fundamentals of Bioinformatics and Computational Biology. Vol 6. Springer; 2015.

2 Baxevanis A, Ouellette F. Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. John Wiley & Sons, Inc.; 2001.

3 Zhang X, Zhou X, Wang X. Basics for Bioinformatics. In: Basics of Bioinformatics. Beijing: Springer; 2013.

4 Yang A, Zhang W, Wang J, Yang K, Han , Zhang L. Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA. Frontiers in Bioengineering and Biotechnology. 2020 1032.

5 Jelovic A, Mitic N, Eshafah S, Beljanski M. Finding Statistically Significant Repeats in Nucleic Acids and Proteins. Journal of Computational Biology. 2018;25(4):375-387.

6 Jelovic A. RepeatsPlus - program for finding motifs and repeats in data sequences. J Bioinform Comput Biol. 2021;19(3).

7 Duitama J, Zablotskaya A, Gemayel R, Jansen A. Large-scale analysis of tandem repeat variability in the human genome. Nucleic Acids Research. 2014;42:5728–5741.

8 Rogozin I, Milanesi L, Kolchanov N. Gene structure prediction using information on homologous protein sequence. Bioinformatics. 1996;12(3):161–170.

9 Altschul S, Gish W, Miller W, Myers E, Lipman D. Basic local alignment search tool. J Mol Biol. 1990;215:403-10.

10 Thompson J, Higgins D, Gibson T. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22:4673–80.

11 Rigden D, Ferna´ndez X́. The 2022 Nucleic Acids Research database issue and the online molecular biology database collection. Nucleic Acids Research. 2022;50(D1):D1–D10.

12 NAR Database Summary Paper Category List. [Internet]. [cited 2021 January 8]. Available from: https://www.oxfordjournals.org/nar/database/cat/1.

13 National Center for Biotechnology Information (NCBI)[Internet]. [Internet]. 2017 Available from: https://www.ncbi.nlm.nih.gov/.

14 Sayers E, Cavanaugh M, Clark K, Pruitt K, Schoch C, Sherry S, Karsch-Mizrachi I. GenBank. Nucleic Acids Research. 2021;49:D92–D96.

15 Pickett B, Sadat E, Zhang Y, Noronha J, Squires B, Hunt V, Liu M, Kumar S, Zaremba S, Gu Z, et al. ViPR: an open bioinformatics database and analysis resource for virology research. Nucleic Acids Research. 2012;40:D593–D598.

16 Saeed U, Usman Z. Biological Sequence Analysis. Computational Biology. 2019 55-69.

17 MR: Virus Metadata Resource. [Internet]. 2019 [cited 2020 Feb]. Available from: https://talk.ictvonline.org/taxonomy/vmr/m/vmr-file-repository/9341.

18 Needleman S, Wunsch C. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology. 1970;48(3):443–53.

19 Smith T, Waterman M. Identification of Common Molecular Subsequences. Journal of Molecular Biology. 1981;147(1):195–197.

20 Sievers F, Higgins D. Clustal Omega for making accurate alignments of many protein sequences. Protein Science. 2018;27:135-145.

21 Sievers F, Barton G, Higgins D. Multiple Sequence Alignment. In: Bioinformatics. Wiley; 2020. p. 227-250.

22 Sievers F, Wilm A, Dineen D, Gibson T, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011 7:539.

23 Zielezinski A, Girgis H, Bernard G, Leimeister CA, Tang K, Dencker T, Lau AK, Röhling S, Choi JJ, Waterman M, et al. Benchmarking of alignment-free sequence comparison methods. Genome Biology. 2019 20:144.

24 Zielezinski A, Vinga S, Almeida J, Karlowski W. Alignment-free sequence comparison: benefits, applications, and tools. Genome Biology volume. 2017 186.

25 Vinga S, Almeida J. Alignment-free sequence comparison—a review. Bioinformatics. 2003;19(4):513–523.

26 Ren J, Bai X, Lu YY, Tang K, Wang Y, Reinert G, Sun F. Alignment-Free Sequence Analysis and Applications. Annu Rev Biomed Data Sci. 2018 93-114.

27 Luczak B, James , Girgis H. A survey and evaluations of histogram-based statistics in alignment-free sequence comparison. Briefings in Bioinformatics. 2019;20(4):1222–1237.

28 Shannon C. A Mathematical Theory of Communication. The Bell System Technical Journal. 1948;27:623-656.

29 Wei D, Jiang Q, Wei Y, Wang. A novel hierarchical clustering algorithm for gene sequences. BMC Bioinformatics volume. 2012.

30 Bao J, Yuan R, Bao. An improved alignment-free model for dna sequence similarity metric. BMC Bioinformatics volume. 2014.

31 Wei D, Jiang Q. A DNA sequence distance measure approach for phylogenetic tree construction. In: 2010 IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA); 2010. p. 204-212.

32 Li C, Wang J. Relative entropy of DNA and its application. Physica A: Statistical Mechanics and its Applications. 2005;347(4):465-471.

33 Dai Q, Liu X, Yao Y, Zhao F. Numerical characteristics of word frequencies and their application to dissimilarity measure for sequence comparison. Journal of Theoretical Biology. 2011;276(1):174-180.

34 Wang J, Zheng X. Wse a new sequence distance measure based on word frequencies. Mathematical Biosciences. 2008;215:78-83.

35 Gupta , Niyogi , Misra. An alignment-free method to find similarity among protein sequences via the general form of Chou's pseudo amino acid composition. SAR and QSAR in Environmental Research. 2013;24(7):597-609.

36 Nandy , Ghosh , Nandy. Numerical Characterization of Protein Sequences and Application to Voltage-Gated Sodium Channel α Subunit Phylogeny. In Silico Biology. 2009;9:77–87.

37 Yao YH, Dai Q, Li C, He PA, Nan XY, Zhang YZ. Analysis of similarity/dissimilarity of protein sequences. PROTEINS : Structure, Function, and Bioinformatics. 2008;73(864–871).

38 Qi ZH, Jin MZ, Li SL, Feng. A protein mapping method based on physicochemical properties and dimension reduction. Computers in Biology and Medicine. 2015;57:1-7.

39 Chen Y, Li KS, Chang S, Yang L. A new 3D graphical representation for similarity/dissimilarity studies of protein sequences. COMPUTER MODELLING & NEW TECHNOLOGIES. 2014;18: 296-303.

40 Yao Yh, Kong , Dai , He Pa. A Sequence-Segmented Method Applied to the Similarity Analysis of Long Protein Sequence. Communications in Mathematical and in Computer Chemistry. 2013;70:431-450.

41 Mahmoodi-Reihani M, Abbasitabar , Zare-Shahabadi V. A novel graphical representation and similarity analysis of protein sequences based on physicochemical properties. Physica A: Statistical Mechanics and its Applications. 2018;510:477-485.

42 Ping P, Zhu X, Wang L. Similarities/dissimilarities analysis of protein sequences based on PCA-FFT. Journal of Biological Systems. 2017 Feb;25(1):29-45.

43 Wu C, Gao R, De Marinis Y, Zhang Y. A novel model for protein sequence similarity analysis based on spectral radius. Journal of Theoretical Biology. 2018;446:61-70.

44 Bonham-Carter O, Steele J, Bastola D. Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. Briefings in Bioinformatics. 2013;15(6):890-905.

45 Song , Ren , Reinert , Deng , Waterman MS, Sun. New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. Briefings in bioinformatics. 2014;15(3):343-53.

46 Bandyopadphyay S, Maulik U. Data Mining and Knowledge Discovery Methods with Case Examples. In: Basics of Bioinformatics. Beijing: Springer; 2013. p. 258-270.

47 Pearson. An Introduction to Sequence Similarity ("Homology") Searching. Current Protocols in Bioinformatics. 2013.

48 Tan PN, Steinbach M, Karpatne A, Kumar V. Introduction to Data Mining, 2nd ed. Pearson Education; 2019.

49 Zhou C, Cule B, Goethals B. Pattern Based Sequence Classification. IEEE Transactions on Knowledge and Data Engineering. 2016;8(5):1285 - 1298.

50 Xing Z, Pei J, Keogh E. A Brief Survey on Sequence Classification. ACM SIGKDD Explorations Newsletter. 2010;12(1):40-48.

51 Xia P, Zhang L, Li F. Learning similarity with cosine similarity ensemble. Information Sciences. 2015;307:39-52.

52 Hossin M, Sulaiman MN. A Review on Evaluation Metrics for Data Classification Evaluations. International Journal of Data Mining & Knowledge Management Process. 2015;5(2).

53 Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Information Processing and Management. 2009;45:427–437.

54 Behera , Kumaravelan , Kumar.B. Performance Evaluation of Deep Learning Algorithms in Biomedical Document Classification. In: 2019 11th International Conference on Advanced Computing (ICoAC); 2019; Chennai, India.

55 Wu XZ, Zhou ZH. A unified view of multi-label performance measures. In: ICML'17: Proceedings of the 34th International Conference on Machine Learning; 2017; Sydney NSW Australia. p. 3780–3788.

56 Halkidi M, Vazirgiannis M, Batistakis Y. Quality Scheme Assessment in the Clustering Process. Lecture Notes in Computer Science. 2000;1910:265-276.

57 Vukicevic M. Razvoj i projektovanje algoritama za klasterovanje ekspresije gena. Univerzitet u Beogradu, Fakultet organizacionih nauka; 2014.

58 Dimitriadou E, Dolničar , Weingessel. An examination of indexes for determining the number of clusters in binary data sets. Psychometrika. 2002 137–159.

59 Wu J, Chen J, Xiong H, Xie M. External validation measures for K-means clustering: A data distribution perspective. Expert Systems with Applications. 2009;36:6050–6061.

60 Lawson R, Jurs P. New Index for Clustering Tendency and Its Application to Chemical Problems. J. Chem. Inf. Comput. Sci. 1990;30:36-41.

61 Sokal R, Rohlf J. The Comparison of Dendrograms by Objective Methods. Taxon. 1962;11(2):33-40.

62 Böcker S, Canzar S, Klau G. The Generalized Robinson-Foulds Metric. In: International Workshop on Algorithms in Bioinformatics; 2013; Berlin. p. 156-169.

63 Robinson D, Foulds L. Comparison of phylogenetic trees. Mathematical Biosciences. 1981;53(1-2):131-147.

64 Jovanovic J. New Method for Sequence Similarity Analysis Based on the Position and Frequency of Statistically Significant Repeats. Current Bioinformatics. 2021 Aug;16(10):1299 - 1310.

65 IBM DB2 Data Management System. [Internet]. Available from: https://www.ibm.com/analytics/db2.

66 Team RC. R: A Language and Environment for Statistical Computing. [Internet]. 2017 Available from: https://www.R-project.org.

67 Ripley B, Lapsley M. RODBC: ODBC Database Access. [Internet]. 2017 Available from: https://cran.r-project.org/web/packages/RODBC/.

68 Wild F. Latent Semantic Analysis. [Internet]. 2015 Available from: https://cran.r-project.org/web/packages/lsa/.

69 Cheng J, Galili T. Interactive Heat Maps Using 'htmlwidgets' and 'D3.js'. [Internet]. 2016 Available from: http://cran.nexr.com/web/packages/d3heatmap/.

70 Henry L, Wickham H. purrr: Functional Programming Tools. [Internet]. 2017 Available from: https://cran.r-project.org/web/packages/purrr/index.html.

71 Kassambara A, Mundt F. factoextra: Extract and Visualize the Results of Multivariate Data Analyses. [Internet]. 2017 Available from: https://cran.r-project.org/web/packages/factoextra/index.html.

72 Wei T, Simko V. corrplot: Visualization of a Correlation Matrix. [Internet]. 2017 Available from: https://cran.r-project.org/web/packages/corrplot/index.html.

73 Galili T, Jefferis G. dendextend: Extending 'dendrogram' Functionality in R. [Internet]. 2020 Available from: https://cran.r-project.org/web/packages/dendextend/index.html.

74 Schliep. phangorn: phylogenetic analysis in R. Bioinformatics. 2011;27(4):592--593.

75 Schliep K, Potts A, Morrison DA, Grimm GW. Intertwining phylogenetic trees and networks. Methods in Ecology and Evolution. 2017;8(10):1212--1220.

76 Sherrill-Mix. taxonomizr: Functions to Work with NCBI Accessions and Taxonomy. [Internet]. 2021 Available from: https://cran.r-project.org/web/packages/taxonomizr/taxonomizr.pdf.

77 NCBI Taxonomy. [Internet]. [cited 2020 Feb]. Available from: https://www.ncbi.nlm.nih.gov/sites/batchentrez.

78 Saw AK, Tripathy BC, Nandi. Alignment-free similarity analysis for protein sequences based on fuzzy integral. Scientific Reports volume 9. 2019.

79 Lu YY, Tang K, Ren J, Fuhrman J, Waterman M, Sun F. CAFE: aCcelerated Alignment-FrEe sequence analysis. Nucleic Acids Research. 45(W1):W554–W559.

80 Gorbalenya AE, Baker SC, Baric RS, Groot RJd, Drosten , Gulyaeva AA, Haagmans BL, Lauber , Leontovich AM, Neuman BW, et al. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nature Microbiology. 2020 Mar;5:536–544.

81 Gorbalenya A, Snijder E, Spaan W. Severe Acute Respiratory Syndrome Coronavirus Phylogeny. JOURNAL OF VIROLOGY. 2004;78(15):7863–7866.

82 Zhang Z, Schwartz S, Wagner L, Miller W. A Greedy Algorithm for Aligning DNA Sequences. Journal of Computational Biology. 2000;7(1-2):203-214.

83 Yang A, Wei Z, Wang J, Yang K, Han Y, Zhang. B - Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA. Front Bioeng Biotechnol. 2020 8:1032.

84 Pearson W, Lipman D. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. U.S.A.. 1988;85:2444–2448.

85 Sievers F, Wilm A, Dineen D, Gibson T, Karplus K, Weizhong L, Lopez R, McWilliam H, Remmert M, Söding J, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology. 2011;7:539.

86 Galili. dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. Bioinformatics. 2015;31(22):3718–3720.

Биографија аутора

Биографски подаци

Јасмина Јовановић рођена је 19. октобра 1987. године у Пожаревцу. Основну школу "Иво Лола Рибар" и Гимназију у Великом Градишту завршила је као носилац дипломе Вук Караџић и ђак генерације.

Математички факултет у Београду уписала је 2006. године, смер Рачунарство и информатика. Дипломирала је у јунском испитном року 2010. године са просечном оценом 9.82. Мастер студије, на студијском програму Математика, модул Рачунарство и информатика, завршила је 2011. године са просечном оценом 10.00. Мастер рад под називом "Дизајн и имплементација апликације за мобилна плаћања рачуна путем SMS и USSD сервиса" одбранила је под менторством др Владимира Филиповића, редовног професора Математичког факултета Универзитета у Београду. Добитник је стипендије "Доситеја", фонда за младе таленте Републике Србије за студенте завршних година студија 2009. године, као и 2010. године као студент мастер студија.

Докторске студије студијског програма Информатика уписала је 2011. године. Све испите предвиђене планом студија положила је са просечном оценом 9.83.

Основна област интересовања је истраживање података у биоинформатици. Током докторских студија била је на стручном усавршавању у Лондону на Унивезитету "Imperial College London", где је добила сертификат за "Студије асоцијације целокупног генома".

Библиографија
Радови у међународним часописима са СЦИ листе

1. Jovanović J. *New Method for Sequence Similarity Analysis Based on the Position and Frequency of Statistically Significant Repeats*. Current Bioinformatics. 2021; 16(10):1299-1310, doi: 10.2174/1574893616999210805165628, (M21; IF2020= 3.543).
2. Zeljic K, Jovanovic I, Jovanovic J, Magic Z, Stankovic A, Supic G. *MicroRNA meta-signature of oral cancer: evidence from a meta-analysis*. Ups J Med Sci. 2018; 123(1): 43-49, doi: 10.1080/03009734.2018.1439551, (M21; IF2018= 2.747).
3. Jovanović I, Zivković M, Jovanović J, Djurić T, Stanković A. *The Co-Inertia approach in identification of specific microRNA in early and advanced atherosclerosis plaque*. Med Hypotheses. 2014;83(1):11-5, doi: 10.1016/j.mehy.2014.04.019, (M23; IF2014= 1.074).
4. Stojkovic G; Jovanovic I; Dimitrijevic M; Jovanovic J; Tomanovic N; Stankovic A; Arsovic N; Boricic I; Zeljic K. *The meta-signature guided investigation of miRNA candidates as potential biomarkers of oral cancer.* Oral Diseases, 2022, 1– 15, doi: 10.1111/odi.14185, (M21; IF2014= 3.511).

Саопштења на научним скуповима

1. Jovanović I, Zivković M, Jovanović J, Djurić T, Stanković A. *Could integrative bioinformatic approach predict the circulating miRs that have significant role in pancreatic tissue in type 2 diabetes?*. Belgrade Bioinformatic Conference BelBi 2016 Proceedings, 2017; 82-87.
2. Jovanović I, Zivković M, Jovanović J, Djurić T, Stanković A. *The improvement of microRNA activity prediction: The integration of Co Expression Meta Analysis of microRNA Targets into Co-Inertia analysis*, NGS and non-coding RNA data analysis COST workshop, 15-16 May 2014 Plovdiv, Bulgaria.
3. Jovanović I, Zivković M, Jovanović J, Djurić T, Stanković A. *The prediction of potentially new microRNA biomarkers for advanced atherosclerosis using Co-inertia analysis*, V Congress of the Serbian Genetic Society, September 28th – October 2nd 2014, Kladovo, Serbia.
4. Zeljić K, Jovanović I, Jovanović J, Magić Z, Stanković A, Šupić G. *Identification of miRNA meta-signature for discrimination between oral cancer and normal tissue: meta-analysis approach*. The Third Congress of the Serbian Association for Cancer Research SDIR-3, Belgrade, Serbia, 6-7 October 6th – 7th, 2017.

Изјава о ауторству

Име и презиме аутора Јасмина Јовановић

Број индекса 2022/2011

Изјављујем

да је докторска дисертација под насловом

Развој метода за анализу сличности биолошких секвенци на основу карактеристика поновака

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

У Београду, _____

Потпис аутора

_____

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора: Јасмина Јовановић
Број индекса: 2022/2011
Студијски програм: Информатика
Наслов рада: Развој метода за анализу сличности биолошких секвенци на основу карактеристика поновака
Ментор: Проф. др Ненад Митић

Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла ради похрањивања у **Дигиталном репозиторијуму Универзитета у Београду.**

Дозвољавам да се објаве моји лични подаци везани за добијање академског назива 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.

У Београду, _____

Потпис аутора

_____

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:
Развој метода за анализу сличности биолошких секвенци на основу карактеристика поновака

која је моје ауторско дело.

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

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

1. Ауторство (CC BY)
2. Ауторство – некомерцијално (CC BY-NC)
3. Ауторство – некомерцијално – без прерада (CC BY-NC-ND)
4. Ауторство – некомерцијално – делити под истим условима (CC BY-NC-SA)
5. Ауторство – без прерада (CC BY-ND)
6. Ауторство – делити под истим условима (CC BY-SA)

(Молимо да заокружите само једну од шест понуђених лиценци.
Кратак опис лиценци је саставни део ове изјаве).

У Београду, _____

Потпис аутора

_____

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 прерада**. 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 прерада**. 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 одстране 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.