Filozofická fakulta Univerzity Palackého v Olomouci

Katedra obecné lingvistiky

Lingvistická analýza v typologii proteinů

magisterská diplomová práce

Autor: Bc. Johana Lukovská

Vedoucí práce: Mgr. Dan Faltýnek, Ph.D.

Olomouc

2018

Prohlášení

Prohlašuji, že jsem magisterskou diplomovou práci „Lingvistická analýza v typologii proteinů“ vypracovala samostatně a uvedla jsem veškerou použitou literaturu a veškeré použité zdroje.

V Olomouci dne Podpis

Abstrakt

Název práce: Lingvistická analýza v typologii proteinů

Autor práce: Bc. Johana Lukovská

Vedoucí práce: Mgr. Dan Faltýnek, Ph.D.

Počet stran a znaků: 117, 170 723

Počet příloh: 0

Abstrakt: Cílem této práce je experimentální popis charakteristických znaků genetického textu proteinů. Řešení tohoto problému vyžaduje stanovení výchozí typologie proteinů na základě vhodně zvolené proteomeické literatury, v tomto případě se jedná o proteiny klasifikované podle biologické funkce. Dalším krokem bylo shromáždění dostatečného množství vzorků, které pocházejí z databáze Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) a jejich anotace. Data jsou zpracována kvantitativně lingvistickou analýzou pomocí programu Quantitative Index Text Analyzer (Quita), s jehož pomocí jsou sledovány n-gramy stringů a jejich projevy v modelu Bag of Words. Jednotlivé typy stringů jsou analyzovány ve skupinách s podobnou délkou tak, aby byl zohledněn vliv délky na jednotlivé lingvistické metriky, přičemž pro zajištění přesnějších výsledků je v analýze použita kosinovská vzdálenost. Analýza pracuje s vybranými indexy kvantitativní lingvistiky a zaměřuje se zejména na entropii, type token ratio, repeat rate atd. Jednotlivé lingvistické metriky dále slouží jako vlastnosti textu v data miningové analýze (Hclust, PCA, MDS atd.).

Klíčová slova: protein, klasifikace, typologie proteinů, aminokyseliny, funkce proteinů, Quita, Bag of Words, n-gramy, entropie, TTR, PCA, MDS

Abstract

Title: Linguistic Analysis in Protein Typology

Author: Bc. Johana Lukovská

Supervisor: Mgr. Dan Faltýnek, Ph.D.

Number of pages and characters:117, 170 727

Number of appendices: 0

Abstract:

The purpose of this experimental master theses is to describe a characteristics of genetic text of protein. To solve this problem, it is necessary to determine typology of protein based on chosen proteomic literature, in this case the classification is based on thein biological function. Next step was to gather appropriate number of protein samples using the database Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) and annotate the data. Obtained data are processed by quantitative linguistics analysis via Quantitative Index Text Analyzer (Quita) software. The analysis looks at n-grmas of protein strings and their demonstration shown Bag of Words model. The individual types of strings are analyzed in groups classified by lenght of single linguistics metrics and to ensure the most accurate outcome, the cosine distance is used. This analysis focuses especially on chosen indicies of quantitative linguistic, such as entropy, type token ratio, repeat rate etc. Single linguistics metrics than serve as features of the text in data mining analysis (Hclust, PCA, MDS).

Keywords: protein, typology of protein, amino acids, protein function, Quita, n-grams, Bag of Words, entropy, TTR, PCA, MDS

Obsah

Úvod

Propojení biologie a lingvistiky skýtá nové možnosti výzkumu pro obě tyto disciplíny. Aplikace formálních pravidel lingvistiky při popisu struktury genomu slouží například k popisu gramatických pravidel pro báze DNA, molekul RNA a podobně. Lingvistickým studiem proteinů se však zabývá minorita biolingvistické a bioinformatické komunity, a podle dosavadních výsledků má právě tento výzkum signifikantní potenciál. Tzv. „proteinová lingvistika“ (Protein Linguistic) reflektuje jazykové plány a promítá je na hierarchii molekulární biologie tak, že sekvence odpovídá lexikální rovině, struktura rovině syntaktické, funkce rovině sémantické a role pak rovině pragmatické.[1] Předmětem této magisterské práce je kvantitativní analýza typologie proteinů klasifikovaných na základě jejich funkce, což odpovídá sémantické rovině textu.

Tato experimentální magisterská práce si klade za cíl odpovědět na výzkumnou otázku, zda a jakým způsobem je možné provést kvantitativní lingvistickou analýzu proteinů klasifikovaných podle jejich biologické funkce a popsat tak charakteristické znaky jejich genetického textu pomocí programu QUITA (Quantitative Text Index Analyzer), modelu Bag of Words a vybraných kvantitativně lingvistických indexů. Vzhledem k tomu, že účelem práce je kvantitativní lingvistická analýza, jsou tomu přizpůsobeny kapitoly pojednávající o proteinu, jakožto o základní jednotce tvořící živé organismy. Kapitoly čerpají z teoretického základu zvolené proteomické literatury, a poskytují úvod do problematiky vhodný i pro čtenáře, kteří s tímto tématem nejsou obeznámeni. Složité chemické a biologické procesy, kterými proteiny procházejí, jsou proto popsány obecně s případným doplněním informací v poznámkovém aparátu, neboť práce neusiluje o hloubkový popis na úrovni makromolekulární biologie. Naopak teoretické kapitoly slouží jako podkladový materiál základních znalostí o problematice. Co se týká formálních záležitostí, tato magisterská práce používá citace zdrojů přímo textu, a to zejména kvůli přehlednosti, dalším důvodem pro citování v textu je pak hojné využití poznámkového aparátu. Veškeré citace, jak přímé, tak i nepřímé, jsou čerpány z cizojazyčné literatury a překlad je dílem autora, který nese zodpovědnost za případné nesrovnalosti v interpretaci překladu.

[1] Searls, D. B. Reading the book of life. Bioinformatics, 17, 579–580 (2001).

Celá tato práce je rozdělena do čtyř kapitol, z nichž tři se věnují teoretickému uchopení a popisu proteinu, jakožto základní biologické jednotky, a čtvrtá, nejobsáhlejší, kapitola je pak věnována samotné kvantitativní analýze. První kapitola popisuje základní informace o proteinu, uvádí několik z možných definic, informace o základních vlastnostech proteinu, jeho složení, skládání, o tvaru molekul atd. Zabývá se také aminokyselinami, základními jednotkami tvořícími protein, a v neposlední řadě také peptidy. Teoretický základ této kapitoly se opírá o několik děl proteomické literatury. Jedná se zejména o publikaci Bruce Albertse *Molecular Biology of the Cell* (2007) a knihu David Whitford *Proteins: Structure and Function* (2005), které poskytují základní vhled do problematiky.

Druhá kapitola přináší informace o klasifikaci proteinů a slouží tedy jako podkladový materiál, jehož poznatky jsou využité v následné analýze. Tato práce zmiňuje více možných klasifikací proteinů, nicméně využívá klasifikace J. L. Jiana z publikace *Fundamentals of Biochemistry* (2005), a to zejména kvůli její přehlednosti. Podle této klasifikace se proteiny dělí do čtyř kategorií, podle zdroje molekuly proteinu, podle tvaru molekuly proteinu, podle složení a rozpustnosti nebo na základě jejich biologické funkce. Pro tuto práci je stěžejní klasifikace na základě biologické funkce proteinů, neboť právě proteiny rozdělené do podskupin podle své biologické funkce jsou následně analyzovány. Je nutné zmínit, že práce reflektuje rozmanitost vydělených podskupin, a to jak v teoreticky popisné části, tak v samotné analýze.

Třetí kapitola poskytuje doplňující informace k předchozí teorii popisem struktury proteinů, které mají také částečný vliv na jeho funkci. V samotné analýze je však vzhledem k metodologickému uchopení a také rozsahu práce struktura proteinu zohledněna jen v několika případech. Tato kapitola, stejně jako v širším měřítku i tato diplomová práce, poskytuje podklad pro navazující akademickou činnost, týkající se popisu syntaxe genetického textu proteinů, tedy právě struktury proteinů. Tato kapitola tedy obsahuje základní informace o primární, sekundární, terciární a kvartérní struktuře a pro jejich popis využívá mezi jinými publikaci od autorů Trudy a Jamese McKee, *Biochemistry: The Molecular Basis of Life* (2015).

Čtvrtou kapitolu tvoří analytická část, která obsahuje popis metodologického uchopení analýzy, popis vybraných vzorků proteinů klasifikovaných do skupin podle biologické funkce, jejich charakterizace na základě modelu Bag of Words a vybraných indexů kvantitativní lingvistiky, a v neposlední řad celkové analýze všech výše vytyčených

skupin proteinů. Celá tato kapitola se snaží odpovědět na výzkumnou otázku vytyčenou výše, přičemž celkové shrnutí analýzy pak poskytuje závěr. Kvantitativní lingvistická analýza používá software QUITA, s jehož pomocí budou popsány n-gramy stringů a jejich projevy v modelu Bag of Words. Analýza se dále zaměřuje na indexy vybrané indexy, které dále slouží jako vlastnosti textu v data miningové analýze, zejména v hierarchickém shlukování, multidimenzionálním škálování a v analýze hlavních komponent.

1 Proteiny

1.1 Definice a základní vlastnosti proteinu

Název protein pochází z řeckého slova proteios, tedy nejvýznamnější nebo první. Termín protein představil roku 1838 holandský chemik Gerardus Johannes Mulder. *„Proteiny jsou komplexní organické substance založené na dusíkové bázi obsažené v buňkách živých organismů.“* (Jain 2005:132) Jedná se o vnitrobuněčné makromolekuly, které tvoří více než polovinu čisté hmotnosti (hmotnosti sušiny) většiny organismů. Proteiny zaujímají ústřední postavení ve složení a chodu živé hmoty. Jsou těsně propojeny se všemi fázemi chemických a fyzických aktivit, které ovlivňují život buňky. (Jain 2005:132–133) Jedná se o molekulární nástroje, které mohou nabývat rozličných funkcí. Krom toho, že proteiny slouží jako strukturní jednotky živých organismů, například aktin a myosin v živočišných svalových buňkách, zapojují se také do různorodých funkcí jako je katalýza, metabolická regulace, transport a obrana apod. (McKee 2015: Chapter 5) Proteiny mohou dále sloužit například jako hormony, které regulují růst rostlin a živočichů, stejně jako kontrolují průběh mnoha dalších fyziologických funkcí. Některé proteiny tvoří endorfiny, které napomáhají potlačování bolesti, vytváření euforie a pocitu štěstí apod. (Jain 2005: 198 - 199)

Proteiny také determinují tvar a strukturu buňky a rovněž slouží jako hlavní nástroj buněčného děleni a molekulární katalýzy. Ačkoli informace nutné k vytvoření buňky jsou uloženy v DNA, genetická informace má minimální přímý vliv na buněčné procesy.[2] (Alberts 2007: 202) Buňka totiž produkuje proteiny na základě svého dědičného materiálu nebo genotypu. Proteiny v sobě mají zabudovanou informaci o své katalytické aktivitě, vnitrobuněčné struktuře a interakci s ostatními proteiny a mají proto proto zásadní vliv nejen na buněčnou strukturu, ale i na funkci buňky. (Jain 2005: 133)

Díky své vysoké molekulární hmotnosti[3] jsou proteiny makromolekulami. To jsou polymery, řetězově uspořádané molekuly produkované spojením malých jednotek aminokyselin, nazývaných monomery. (Jain 2005: 133) *„[...]protein je tedy lineárním*

[2] Například gen pro hemoglobinu není schopen nést kyslík, tato schopnost je závislá na proteinu, který je geneticky specifikován. (Alberts 2007: 202)

[3] Molekulární hmotnost je jednou z fyzikálních vlastností proteinů. Vzhledem k mimořádné velikosti, špatné stabilitě, specifickým podmínkám rozpustnosti a vysoké reaktivitě je stanovení molekulární hmotnosti obtížné. Obecně však má protein vysokou molekulární hmostnost, která se pohybuje v rozmezí 5 x 10^3 až 1x $10^{6.}$. (Jian 2005: 215)

polymerem několika stovek aminokyselin, které se při syntéze řadí za sebou do lineárního, nevětveného řetězce." (Markoš 2014:32) Aminokyseliny se do řetězce řadí procesem translace, k němuž se používá 20 aminokyselin. Řazení je v naprosté většině případů neperiodické, ale nenáhodné. Probíhá tedy podle předpisu, nicméně pořadí aminokyselin nelze vypočítat.[4] (Markoš 2014:32) *„Každá z aminokyselin tvořící protein má rozdílnou chemickou podstatu. Právě tato rozmanitost umožňuje nesmírnou versatilitu chemických vlastností různých proteinů a také vysvětluje, proč namísto molekul RNA katalyzují většinu chemických reakcí v buňce právě proteiny."* (Alberts 2007: 202) Nové proteiny se obvykle vyvýjejí alternacemi již existujících proteinů. Buňky mají genetický mechanismus, který umožňuje genům, aby se v průběhu evoluce duplikovaly, modifikovaly a přeskupovaly. Následkem toho může být základní struktura proteinu s užitečnými vlastnostmi začleněna do mnoha dalších proteinů. Proteiny s odlišnými ale příbuznými funkcemi mají často stejnou sekvenci aminokyselin. Takovéto rodiny proteinů se pravděpodobně vyvinuly z jednoho prapůvodního genu, který byl duplikován během evoluce. Tímto způsobem vznikly další geny, ve kterých se postupně nahromadily mutace, umožňující produkci příbuzných proteinů s odlišnými novými funkcemi. Některé ze změn aminokyselin, odlišujících tyto proteiny, byly bezpochyby vybrány v průběhu evoluce, protože vedly k změnám v biologické aktivitě. To dává jednotlivým rodinám členové různé funkční vlastnosti, které mají dnes. Jiné změny aminokyselin byly neutrální, bez prospěšného, či škodlivého vlivu na základní strukturu a funkce proteinu. Vzhledem k tomu, že mutace je náhodný proces, mnohé změny musely mít škodlivý efekt, který změnil trojrozměrnou strukturu proteinů tak, aby je deaktivoval. Takové neaktivní proteiny by byly ztraceny vždy, když jednotlivé organismy byly eliminovány přirozeným výběrem. Není tedy překvapující, že buňky obsahují celé soubory strukturně příbuzných polypeptidových řetězců, které mají a společné předky, ale různé funkce. (Alberts 2007: 207-208)

Při tvorbě proteinu mohou být aminokyseliny spojeny do sekvencí jakýchkoli představitelných rozměrů. Ze všech myslitelných kombinací se pouze jejich zlomek vyskytuje v živých organismech[5]. Důvod lze ilustrovat komplexním setem strukturálních a

[4] Pokud známe pořadí 299 aminokyselin v řetězci, nevíme jaká aminokyselina se v řetězci zařadí na pozici 300. (Markoš 2014:32)

[5] Například, pokud by byl hypotetický protein složen ze 100 aminokyselin, vzniklo by tak 20^{100} možných kombinací sekvení aminokysellin. Z trilionu možných sekvencí se však v živých organismech pravděpodobně nevyskytuje více bež dva miliony kombinací. (McKee 2015: Chapter 5)

funkčních vlastností v přírodě se vyskytujících proteinů, které se vyvinuly v průběhu evoluce. Jedná se o strukturální vlastnosti, které umožňují skládání (folding) proteinu jakožto relativně rychlého a úspěšného procesu. Dále o přítomnost vazeb specifických pro jednu molekulu nebo malou skupinu molekul. Další vlastností je odpovídající vyváženost strukturní flexibility tak, aby byla zachována funkce proteinu. Rovněž povrchová struktura proteinu je přiměřená prostředí, ve kterém se protein vyskytuje [6]. A v neposlední řadě také zranitelnost proteinů vůči degradačním reakcím, pokud jsou poškozeny, nebo již déle nepotřebné. (McKee 2015: Chapter 5)

Sekvence aminokyselin rovněž determinuje tvar molekuly proteinu, kterou tvoří. Mnoho vazeb v dlouhém peptidovém řetězci umožňuje libovolnou rotaci atomů, které se na sebe vážou, což umožňuje flexibilitu proteinu. *„Principiálně je tak každá proteinová molekula schopna nabýt téměř nelimitovaného počtu tvarů, tedy konfirmací. Většina polypeptidových řetězců přesto zaujímá konkrétní konfirmaci, která je determinována sekvencí aminokyselin.“* (Alberts 2007: 203) To se děje proto, že páteřní struktura proteinu (backbone protein structure) a boční řetězce (side chains) aminokyselin spolu asocijují za přítomnosti vody tak, aby vytvořily různé typy slabých nekovalentních[7] vazeb. Pokud jsou na klíčových pozicích přítomny patřičné boční řetězce, umožňuje to vznik konkrétní stabilní konfirmace. Většina proteinů může samovolně nabýt svého správného tvaru. Při použití určitých rozpouštědel lze protein rozložit nebo denaturovat na flexibilní polypeptidový řetězec, který tak ztratí svou přirozenou konfirmaci. Pokud je rozpouštědlo odstraněno, protein většinou nabude svého původního přirozeného tvaru, což indikuje to, že veškeré informace potřebné ke specifikaci tvaru proteinu jsou obsaženy v samotné sekvenci aminokyselin. (Alberts 2007: 203)

Jedním z nejdůležitějších faktorů rozhodujícím o skládání (folding) proteinu je distribuce jeho polárních a nepolárních[8] bočních řetězců. Mnoho hydrofobních bočních řetězců v proteinu má sklon seskupit se uvnitř buňky, což jim umožňuje vyhnout se kontaktu s vodním prostředím. Stejně jako se olejové kapky shlukují poté co jsou mechanicky

[6] Například hydrofobní prostředí v membránách a hydrofilní v cytoplasmě. (McKee 2015: Chapter 5)
[7] Existují dva typy vazeb, kovalentní a nekovalentní. Kovalentní vazba je stabilní chemická vazba vniklá sdílením jednoho nebo více párů elektronů mezi atomy v molekule. (Jian 2005: 1106) Naproti tomu v nekovalentní vazbě žádné elektrony sdíleny nejsou. Jedná se o relativně slabé vazby, které se však mohou shlukovat tak, aby umožnily vznik silných a specifických interakcí mezi molekulami. (Jian 2005: 1132)
[8] Polaritou bočních řetězců je míněna jejich schopnost vzájemně reagovat s vodou při přirozeném pH (pH 7.0). Boční řetězce mohou být nepolární, odpuzující vodu (hydrofobické) , nebo naopak polární, tedy schopné s vodou interagovat (hydrofilické). (Jain 2005: 141)

rozpuštěny ve vodě. Naproti tomu polární boční řetězce se uspořádávají vně proteinu, kde mohou vzájemně reagovat s vodou a dalšími polárními řetězci tak, aby vytvořily hydrogenové (vodíkové) vazby[9], přičemž téměř všechny polární rezidua náležící proteinu se párují tímto způsobem. Hydrogenové vazby mají klíčovou úlohu při tvorbě mnoha dalších vazeb, které se vytvářejí na povrchu proteinu, jejich hlavní úlohou je však držet pohromadě různé části polypeptidového řetězce[10] ve skládaném (foldovaném) proteinu (Alberts 2007: 203)

Po zapojení do polypeptidového řetězce jednotlivé aminokyseliny podstupují chemické modifikace, což může zahrnovat buď jednoduchou chemickou modifikaci, nebo přírůstek velké chemické skupiny**.** Malé modifikace mohou být permanentní a nezbytné pro správné skládání a funkci proteinu, častěji však jsou minoritní modifikace vratné a umožňují tak regulaci proteinové aktivity. Kovalentní připojení objemných chemických molekulárních sloučenin je obvykle modifikace permanentní, která může zahrnovat pohyby a procesy uvnitř buňky, stejně jako připojení konjugovaného[11] proteinu k jeho prostetické skupině[12]. (Twymann 1999: 287) *„Nově syntetizovaný polypeptid pak musí nabýt své přirozené konfirmace, tedy té konfirmace[13], ve které je biologicky aktivní.“* (Twymann 1999: 287) Mnoho let bylo předpokládáno, že protein při svém skládání prochází všemi možnými konfirmacemi, než dosáhne té, která je pro něj přirozená. Navzdory vysoké rychlosti molekulárního pohybu v proteinu však existuje příliš velké množství konfirmací, kterých je protein schopen nabýt, a než je možné vyzkoušet v průběhu několika málo sekund, kdy

[9] Hxdrogenová, neboli vodíková vazba vzniká , když skupina obsahující atom vodíku, který je kovalentně vázán na atom s negativním elektrickým nábojem (kyslík, dusík), sousedí s druhou skupinou obsahující další atom s negativním elektrickým nábojem. V takovém to případě nastává energeticky výhodná interakce, která se nazývá hydrogenová vazba, která vzniká kvůli tendenci vodíkového atomu sdílet elektrony se dvěma sousedícími atomy kyslíku a dusíku. (Jain 2005: 152)

[10] Každý protein se skládá z jednoho nebo více polypeptidových řetězců, přičemž polypeptidy se nazývají molekuly s molekulární hmostnostní pohybující se mezi několika tisíci a několika miliony daltonů (jednotka používaná k vyjádření molekulové hmotnosti proteinů, která je ekvivaletní jednotce atomové hmotnosti) Molekuly s menší hmotností se typicky skládají z méně než 50 aminokyselin a nazývají se peptidy. Termín protein tedy označuje molekuly s více než 50 aminokyselinami. (McKee 2015: Chapter 5) Pro upřesnění budou níže v textu uvedeny podkapitoly týkající se jak aminokyselin, tak peptidů.

[11] Velká skupina proteinů je označována jako konjugované proteiny, protože taková to komplexní molekula se skládá z proteinové a neproteinové části, přičemž neproteinová část se nazývá prostetická skupina. (Encyclopedia Britannica)

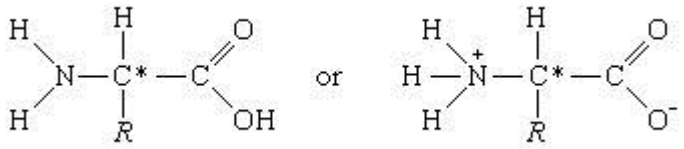
[12] Prostetické skupiny jsou spojovány s proteiny kovalentními nebo nekovalentními vazbami. Může se jednat o mnohé biomolekuly jako jsou lipidy, karbohydráty, nukleové kyseliny, fosfátové skupin, flaviny, heme skupiny a ionty kovů. Komplexní jednotky tvořeny lipidy a proteiny se nazývají lipoproteiny, jednotky obsahující karbohydráty jsou glykoproteiny, a ty s ionty kovu nazýváme metaloproteiny atd. (Whitford 2005:7)

[13] V tomto stavu může mít protein několik typů strukturní organizace, sekundární, terciární a kvartérní. (Twymann 1999: 287) Pro účely této práce bude níže uvedena samostatná kapitola, věnující se struktuře proteinu.

obvykle skládání (folding) proteinu probíhá. Krom toho existence zmutovaných proteinů se specifickými defekty ve skládání indikuje to, že sekvence aminokyselin v proteinech byla selektována během evoluce zejména díky své schopnosti rychle nabýt přirozené konfirmace, stejně jako díky vlastnostem své finální struktury (Alberts 2007: 382)

1.2 Aminokyseliny

Proteiny se skládají z jednoho nebo více polypeptidů, rozvětvených polymerů tvořených 20 různými aminokyselinami. Genom většiny organismů specifikuje aminokyselinovou sekvenci až deseti tisíců proteinů. Hydrolýzou[14] každého polypeptidu je možné získat skupinu aminokyselin, nazývanou aminokyselinovou kompozicí. Existuje řada kombinací 20 aminokyselin tvořících polypeptidy, které se nacházejí v přírodě. Jedná se o tzv. standardní aminokyseliny, které sdílejí stejnou základní strukturu. Obsahují centrální atom uhlíku (α - uhlík), na něj je vázána aminokyselina, karboxylová skupina, atom vodíku a postranní řetězec. (McKee 2015: Chapter 5)



Obr. 1 Obecná struktura aminokyselin, kde R je označení postranního řetězce. (Zdroj: Encyklopedie Britannica)

Kyseliny s takovou to strukturou se nazývají aminokyselinami proto, že atom α – uhlíku v molekule nese aminoskupinu ($-NH_2$). V kyselých roztocích , kde je pH nižší než 4, se karboxylová skupina ($-COO$) kombinuje s iontem vodíku (H^+) a mění se tak do nenabité formy ($-COOH$). V zásaditých roztocích , kde je pH vyšší než 9, amoniová skupina ($-NH^+_3$) ztrácí iont vodíku a mění v aminoskupinu ($-NH_2$). (viz. Obr. 1) (Encyklopedie Britannica) Pro účely této práce seznam standardních aminokyselin spolu s jejich vzorci neuvádím, obojí je možné nalézt ve vybrané proteomické literatuře.

Nestandardní aminokyseliny se skládají z aminokyselinových reziduí, které byly chemicky modifikovány poté, co byly včleněny do polypeptidu, nebo z aminokyselin, které se se vyskytují v živých organismech, ale ne v proteinech. (McKee 2015: Chapter 5) Jako

[14] Hydrolýza je jednou z chemických vlastností proteinů. Jedná se o rozštěpení kovalentní vazby (například peptidové) přidáním vody, čímž vznikne jeden nebo dva výsledné produkty. Hydrolýza je opakem dehydratace.

příklad uveďme hydroxyproline, jehož výskyt je přírodě limitován, ale tvoří asi 12 % složení kolagenu, který je důležitým strukturálním proteinem organismu živočichů. Dalším nestandardním proteinem je pak N-methyllsin, který je možné nalézt v myosinu, proteinu umožňující stahování svalů. Důležitou Nestandardní aminokyselinou je γ-carboxyglutamát, který se vyskytuje v proteinu prothrombinu, který napomáhá srážení krve, dále také v jiných proteinech, které váží $Ca2^{+}$ ve svých biologických funkcích. (Jian 2005: 145)

Sekvence aminokyselin determinuje konečnou trojdimenzionální konfiguraci každého proteinu, je proto nutné popsat jejich strukturu. Aminokyseliny lze klasifikovat podle jejich schopnosti interagovat s vodou, podle tohoto kritéria lze aminokyseliny rozdělit do čtyř skupin, na nepolární, polární, acidické a základní. Nepolární aminokyseliny obsahují z velké většiny hydrokarbonovou R skupinu, která nemá pozitivní ani negativní náboj. Nepolární (hydrofobické) aminokyseliny hrají klíčovou roli v zachování trojdimenzionální struktury proteinu, protože s vodou nereagují, jsou hydrofobní. Naproti tomu polární aminokyseliny s vodou reagují a to zejména proto, že jejich funkční skupiny jsou schopny tvořit hydrogenové vazby. Polární aminokyseliny jsou proto popisovány jako hydrofilické.

Acidické aminokyseliny jsou aminokyseliny s negativně nabitými (acidickými) R skupinami (skupiny v postranních řetězcích). Jejich postranní řetězec obsahuje karboxylovou skupinu s oddělitelným protonem. Výsledný negativní náboj způsobuje elektrochemické chování proteinů. (Jian 2005: 144) Co se týká základních aminokyselin, ty nesou při fyziologickém pH pozitivní náboj, proto mohou tvořit ionické vazby s acidickými aminokyselinami. (McKee 2015: Chapter 5)

1.3 Peptidy

Funkční skupiny organických molekul determinují typy reakcí, kterým tyto molekuly podléhají. Vzhledem k tomu, že aminokyseliny obsahují karboxylovou skupinu, aminoskupinu a odlišné R řetězce, mohou podstupovat četné chemické reakce. Nejdůležitější z těchto reakcí je pak schopnost tvořit peptidové vazby a disulfidické můstky, vzhledem k jejich schopnosti ovlivnit strukturu proteinu. (McKee 2015: Chapter 5)

Jednotky aminokyselin jsou spojeny prostřednictvím karboxylové a amino skupiny tak, aby produkovaly primární strukturu proteinového řetězce. Vazba mezi takto sousedícími

aminokyselinami je specifickým typem amidové vazby[15]. (Jian 2005:146) *„Peptidové vazby jsou prostřední vazby, které jsou formovány, když nesdílený elektronový pár α-amino vodíkového atomu jedné aminokyseliny napadne α-karboxylový uhlík další aminokyseliny a vytvoří tak nukleofilní acylovou substituci.*[16]*“* (McKee 2015: Chapter 5) Takto spojené aminokyseliny jsou poté nazývány aminokyselinovými rezidui, protože peptidová vazba způsobuje dehydrataci, tedy odstranění molekuly vody. (McKee 2015: Chapter 5) Seskupením mnoha takto propojených aminokyselin vzniká řetězec, který se nazývá peptidový. Každý peptidový řetězec se skládá z 50 až několika milionů aminokyselinových jednotek. V závislosti na počtu aminokyselinových molekul tvořících řetězec mohou být peptidy označovány jako dipeptidy, obsahující dvě aminokyselinové jednotky, tripeptidy, obsahující tři aminokyselinové jednotky atd. Pokud není peptid složen z více než 10 aminokyselin, nazýváme ho oligopeptidem, skládá-li se z více jednotek, jedná se o polypeptid. Pokud polypeptid obsahuje více než 100 aminokyselin, bývá označován jako makropeptid. *„Většina proteinů nicméně obsahuje 20 aminokyselin […] a ačkoli počet i způsob jakým jsou aminokyseliny propojeny je výrazně variabilní, možný počet proteinů se blíží nekonečnu.“* (Jian 2005: 147) Takové to skládání pak můžeme připodobnit k nekonečnému množství slov, které mohou být vytvořeny například z 26 anglických písmen. Takto tvořená slova jsou však omezena možnou délkou, zatímco na počet řetězících se aminokyselin se toto omezení nevztahuje[17]. (Jian 2005: 147)

Pokud aminokyseliny vytvoří dostatečně dlouhý řetězec, takovéto polypeptidy mají většinou jasně definovanou trojdimenzionální strukturu. Jedná se o přirozenou konfirmaci molekul, která je důsledkem sekvence aminokyselin, tedy pořadí, v jakém jsou aminokyseliny propojeny. Protože veškerá propojení mezi aminokyselinovými rezidui se skládají z jednotlivých vazeb, každý polypeptid může prodělat konfirmační změny způsobené rotací kolem těchto singulárních vazeb. Většina polypeptidů se však přirozeně skládá do formy, ve které jsou biologicky aktivní. Kvůli neohybnosti peptidové vazby není

[15] Amidová vazba vzniká mezi dvěma α-aminokyselinami nebo peptidy. (Montalbetti, Falque 2005: 10828) Jedná se o organickou složku obsahující skupinu (—C(O)NH) derivovanou z amoniaku tím, že je atom vodíku nahrazen acylovou skupinou. (Oxford Dictionary)

[16] Nukleofilní acylová substituce je adičně- eliminačním mechanismem, kdy funkční deriváty kyselin i karboxylové kyseliny reagují s nukleofily. Touto reakcí dochází k přenosu acylové skupiny R–C=O na jiný nukleofil procesem acylace a funkční deriváty kyselin se tak řadí mezi acylační činidla. (Svoboda 2005: 273)

[17] Tento problém rozpracovala například Erlene Cunningham (1978), která tvrdila, že pokud by se každá proteinová molekula skládala pouze z 250 aminokyselinových reziduí, utilizace všech 20 odlišných monomerů by umožnila zformovat 10325 rozdílných proteinů. Ve skutečnosti proteiny často obsahují více, než 250 jednotek a proto je možné dosáhnout vyššího počtu odlišných proteinových molekul. (Jian 2005: 147)

třetina ze všech vazeb v polypeptidovém páteřním řetězci schopna se volně otáčet, což způsobuje omezené možnosti konfirmací. (McKee 2015: Chapter 5)

Další chemickou reakcí, která mezi aminokyselinami může nastat je disulfidický můstek. Jedná se o běžný mechanismus vyskytující se v přírodě, který je užíván ke stabilizaci mnoha proteinů. *„Disulfidický můstek vzniká, když se vytvoří vazba mezi dvěma cysteinovými rezidui uvnitř proteinu.“* (Britannica) Funkční skupinou cysteinu je skupina thiolů, sirných alkoholů jinak zvaných merkaptany. Jedná se o nejvíce reaktivní boční řetězce v 20 přirozeně se vyskytujících se aminokyselin. Velký atom síry jako součást skupiny thiolů ovlivňuje vlastnosti bočních řetězců a to spolu s disulfidovými vazbami mezi cysteinovými rezidui, které se vyskytují blízko sebe. Takovýmto způsobem se formuje silná kovalentní vazba, disulfidický můstek. (Whitford 2005: 26-27) Takto vytvořené můstky lze nalézt mezi mimobuněčnými preoteiny. V eukaryotických organismech se disulfidické můstky vyskytují uvnitř organel endoplasmatického retikula. (Encyklopedia Britannica)

Peptidy se účastní mnoha biologických aktivit, z nichž nejdůležitější je v tomto případě jejich úloha být meziproduktem v procesu formování proteinů. Peptidy mohou dále sloužit například jako součásti alkaloidů, růstové faktory i hormony. Velká část peptidů má antibakteriální účinky, například Penicilin G je tvořen valinem a cysteinem. (Jian 2005: 150)

2 Klasifikace proteinů

Proteiny lze klasifikovat na základě různých kritérií, nejčastěji se dělí podle zdroje molekuly proteinu, podle tvaru molekuly proteiny, podle složení a rozpustnosti nebo na základě jejich biologické funkce. (Jain 2005: 204) Mnozí autoři, například Whitford v publikaci *Proteins: Structure and Function* (2005) uvádí, že by nemělo být usilováno o formální rozdělení proteinů do tříd za každou cenu, vzhledem k tomu, že proteiny zastávají mnohé role a funkce, čímž se vzpírají jednotné klasifikaci. Přesto dále rozděluje proteiny do několika skupin podle jejich funkcí. Jedná se o enzymy a katalytické proteiny (například trávicí enzym trypsin), kontraktilní proteiny, strukturální (cytoskeletální), transportní protein jako je hemoglobin, efektorní proteiny, mezi které patří například inzulin, defenzivní proteiny (imunoglobuliny), proteiny přenášející elektrony, receptory, nutriční proteiny atd. (Whitford 2005: 5-6) Tato klasifikace je uvedena jako příklad jedné z mnoha možných. Tato práce využívá dělení L.J. Jiana popsané v práci *Fundamentals of Biochemistry* z roku 2005 a pro větší přehlednost uvádí grafické schéma vyhotovené na základě klasifikace proteinů uvedené ve výše zmíněné publikaci.

2.1 Klasifikace podle zdroje molekuly proteinu. Živočišné a rostlinné proteiny.

Na základě zdroje molekuly se proteiny se tradičně dělí na své skupiny, živočišné protein a rostlinné proteiny. Proteiny živočišného původu obsažené například v mléce, vejcích, masu apod. jsou nazývány proteiny vyšší kvality, protože obsahují adekvátní množství základních aminokyselin. Rostlinné proteiny, tedy proteiny nižší kvality[18], obsahují malé množství jedné nebo více základních aminokyselin.[19] Rostlinné proteiny se mimo jiné vyskytují například v rýži a sojových bobech (methionin), dalších luštěninách, jako jsou hrách a fazole (tryptofan), ve slunenčicových semenech (lysin) a v neposlední řadě v obilných zrnech (threonin). (Jain 2005: 204)

2.2 Klasifikace podle tvaru molekuly. Globulární a fibrilární proteiny.

Na základě tvaru molekuly se proteiny dělí na dvě skupiny, globulární a fibrilární proteiny. Globulární proteiny mají kulovitý nebo vejcovitý tvar a jsou rozpustné ve vodě

[18] Ačkoli rostlinné proteiny mají limitovaný obsah některých aminokyselin, neměly by být považovány za nedostatečný zdroj proteinů. (Jain 2005: 204)

[19] Čtyři nejběžnější základní aminokyseliny jsou methionin, lysin, treonin a tryptofan. (Jain 2005: 204)

nebo vodných roztocích, které obsahují kyseliny, zásady, soli nebo alkohol. Jako skupina mají globulární proteiny komplexnější strukturu než fibrilární, a také mají větší množství biologických funkcí. (Jain 2005: 205) Jejich funkce jsou typicky dynamické, na příklad téměř všechny enzymy mají globulární tvar. (McKee 2012: kap. 5) Kromě enzymů se jedná například o proteinové hormony, proteiny uskladňující živiny a proteiny přenášejících krev nebo protilátky. Globulární proteiny lze dále dělit jen s obtížemi, protože na jednu stranu plní škálu odlišných funkcí, na stranu druhou mnoho značně rozdílných globulárních proteinů plní funkce velmi podobné. (Jain 2005: 205) Jako příklad uveďme jednu z možných klasifikací těchto proteinů. Conn a Stumpf klasifikovali globulární proteiny následujícím způsobem. Jedná se o cytochrom C, krevní proteiny (albumin, glykoprotein, imunoglobuliny, hemoglobin, hormony), enzymy a nutriční proteiny. (Conn, Stumpf in Jain 2005: 205)

Fibrilární proteiny vzhledem připomínají vlákna, jsou většinou živočišného původu a jsou nerozpustné v běžných rozpouštědlech/kapalinách jako jsou voda, zředěná kyseliny, zásady, soli a také organická rozpouštědla. Proteiny tohoto typu často zaujímají strukturální nebo ochrannou roli. (Jain 2005: 205) Fibrilární proteiny jsou silné a mají dvě důležité vlastnosti charakteristické pro elastomery. Těmito vlastnostmi jsou schopnost fibrilárního proteinu natáhnout se a poté se smrštit do své původní délky. Pokud jsou tyto proteiny nataženy dostatečně dlouho dobu, jejich původní délka se zvětší natolik, aby se vyrovnala délce natažení. Jestliže jsou oba konce fibrilárního proteinu uvolněné, protein se postupně zkracuje. (Jain 2005: 205) Proteiny fibrilárního tvaru tvoří například spojovací tkáně, kosti, cévy, kůži , vlasy, nehty, rohy, kopyta, vlnu a hedvábí. Mezi fibrilární proteiny patří colagen, elastin, fibroin a keratin. (Jain 2005: 206) Právě keratin se vyskytuje v kůži, vlasech, nehtech atd., a má zejména strukturální a ochrannou funkci. (McKee 2015: Chapter 5)

2.3 Klasifikace na základě složení a rozpustnosti

Na základě složení a rozpustnosti tento systém rozděluje proteiny na základě jejich složení do tří hlavních skupin. Jedná se o jednoduché, konjugované a derivované proteiny. Jednoduché proteiny jsou globulárního typu s výjimkou skleroproteinů, které patří mezi fibrilární. Tato skupina zahrnuje proteiny skládající se pouze z aminokyselin. Dále se dělí na základě své rozpustnosti na protaminy a histony, albuminy, globuliny, gluteliny, prolaminy, skleroproteiny nebo albuminoidy. (Jain 2005: 206)

Protaminy a histony jsou základními proteiny, které tvoří především spermie živočichů. Mají nejjednodušší strukturu a nejnižšší váhu ze všech proteinů, přibližně 5,000. Jsou rozpustné ve vodě a na rozdíl od většiny proteinů se horkem nesráží. Protaminy a histony se skládají zejména ze základních aminokyselin, lysinu a argininu. Protaminy neobsahují sulfur ani aromatické aminokyseliny a proto jsou rozpustitelné v NH4OH (dále čpavek). Histony jsou naproti tomu založeny na slabší bázi, a proto jsou ve nerozpustitelné ve čpavkových rozpouštědlech. (Jain 2005: 206)

Albuminy jsou v přírodě široce rozšířené a hojně se vyskytují v semenech. Lze je rozpustit ve vodě i zředěných roztocích kyselin, solí a zásad, sráží se se saturovanými roztoky kyselých solí jako $(NH_4)_2SO_4$ (síran amonný) nebo neutrálních solí, například NH_2SO_4. Mezi příklady albuminů patří například leukosin vyskytující se v obilovinách, ovalbumin ve vaječném bílku a serum albumin obsažený v krevní plasmě. Mezi další příklady patří myosin obsažený ve svalech a laktabumin v mléčné syrovátce. (Jain 2005: 206)

Globuliny se dále dělí na pseudoglobuliny a euglobuliny, podle rozpustnosti. Zatímco pseudoglobin je rozpustný v roztocích s nízkou ionickou silou, euglobuliny jsou rozpustné, dokud se ionická síla nezvýší. Euglobuliny jsou rozpustné až do izoelektrického bodu. Euglobuliny v přírodě se hojně vyskytují hojněji než pseudoglobuliny. Mezi příklady euglobulinů patří serum globulin vyskytující se v krevní plasmě, ovoglobulin ve vaječném bílku, globuliny obsažené v různých semenech rostlin, například glycinin, edestin amandin, tuberin apod. Příkladem pseudoglobulinu je stejnojmenný protein obsažený v mléčné syrovátce. (Jain 2005: 206-207)

Gluteliny a prolaminy[20] je možné extrahovat pouze ze semen rostlin. Nejsou rozpustné ve vodě, ve zředěných roztocích soli, lze je však rozpouštět ve zředěných kyselinách a zásadách. Prolaminy jsou navíc rozpustné v 60 – 80 % roztocích alkoholu a nleze je srážet horkem. Někteří vědci jsou toho názoru, že gluteliny a prolaminy by neměly být vydělovány jako podtypy jednoduchých proteinů, protože se jedná o malou skupinu rostlinných proteinů, vyskytujích se v obilných zrnech. (Jain 2005: 207)

Skleroproteiny a albuminoidy se vyskytují téměř výhradně u živočichů, proto jsou běžně nazývány „animal skeleton proteiny“. Nejsou rozpustné ve vodě, zředěných roztocích

[20] Podle některých biochemiků tvoří gluteliny a prolaminy pouze malou skupinu rostlinných proteinů a proto nemusí být nutně vydělovány zvlášť. (Karlson in Jian 2005: 207)

kyselin, zásad a solí, ani v 60 – 80 % alkoholových roztocích. Jedná se například o kolagen obsažený v kostech, elastin ve vazech, keratin ve vlasech a fibroin v hedvábí. (Jain 2005: 207)

Konjugované proteiny, nazývány také komplexními proteiny, patří rovněž mezi globulární[21]. Konjugované proteiny se skládají z jednoduchého proteinu a neproteinového komponentu. Neproteinový komponent se nazývá prostetická skupina a proteinu bez jeho prostetické skupiny se říká apoprotein. Proteinová molekula kombinovaná s prostetickou skupinou je pak holoprotein. Prostetické skupiny mají důležitou úlohu ve funkcích proteinů. (McKee 2012: kap. 5) Prostetická skupina může být tvořena ionty kovu, nebo sloučeninou. Následné rozdělení se provádí podle povahy prostetické skupiny. Jedná se o metaloproteiny, chromoproteiny, glykoproteiny, fosfoproteiny, lipoproteiny a nukleoproteiny. (Jain 2005: 207)

Metaloproteiny jsou proteiny pojící se s různými kovy a právě na základě své reaktivity s kovy dělí na tři skupiny. Kovy silně vázány proteiny, kovy slabě vázány proteiny a kovy, které se nepojí s proteiny. Některé těžké kovy, na příklad rtuť, stříbro, měď a zinek jsou silně vázány proteiny jako kolagen, albumin, kasein apod. Mezi kovy, které jsou proteiny vázány slabě patří vápník. Vazba vzniká pomocí radikálu s elektronovým nábojem. Kovy které se na proteiny neváží, se slučují se s nukleovými kyselinami, ve kterých se vyskytují elektrostatickými vazbami. Jedná se např. o sodík a draslík. (Jain 2005: 208)

Chromoproteiny jsou protein, které se vážou s barevnými pigmenty. Tyto pigmenty byly také nalezeny u enzymů jako je kataláza, peroxidáza a flavoenzymy. Podobně je chlorofyl přítomen v buňkách listů ve formě proteinu, chloroplastinu. Chloroplastin se rozpouští ve vodě jako koloid a je snadno denaturovatelný. Mezi další příklady chromoproteinů patří myoglobin, hemoglobin, hemocyanin, hemoerythin, cytochrom, flavoproteiny, kataláza apod. (Jain 2005: 208)

Glycoproteiny a mukoproteiny obsahují karbohydráty jako prostetickou skupinu. Glykoproteiny obsahují malé množství karbohydrátů (méně než 4 %), zatímco v mukoproteinech se vyskytuje více než 4 % karbohydrátů. Mezi glykoproteiny patří například vaječný albumin, mezi mukoproteiny pak ovomukoid z vaječného bílku, mucin ze slin a tendomukoid ze šlach. (Jain 2005: 208) Další podskupinou konjugovaných proteinů

[21] Jedinou výjimku tvoří pigment slepičích per, který je pravděpodobně fibrilární povahy. (Jain 2005: 207)

jsou fosfoproteiny. Jedná se o proteiny spojeny s fosforem, z velké části acidické, jako je například kasein z mléka a ovovitelin z vaječného žloutku. (Jain 2005: 208)

Lipoproteiny jsou proteiny pojící se s lipidy, jako příklady v tomto případě poslouží kefalin a lecitin, které jsou rozpustné ve vodě, ale nerozpustné v organických rozpouštědlech. Dalšími příklady jsou lipovitelin a lipovitelenin z vaječných žloutků. Lipoproteiny jsou ve skutečnosti přechodným meziproduktem při procesu transformace lipidů prostřednictvím absorpce[22] až do stavu jejich plného využití (utilizace). Klasifikace lipoproteinů je často založena na jejich hustotě a na tomto základě se lipoproteiny dělí na 4 skupiny, lipoproteiny s velmi vysokou hustotou (Very high density lipoproteins, zkr.VHDLs), s vysokou hustotou (High density lipoproteins, zkr. HDLs), nízkou hustotou (Low density lipoproteins, zkr. LDLs) a velmi nízkou hustotou (Ver low density lipoproteins, zkr. VLDLs). (Jain 2005: 208)

Nukleoproteiny se skládají z nukleových kyselin a proteinu, resp. protaminů a histonů. Jedná se o sloučeniny na bázi soli složené z proteinů, a díky k opačnému náboji jsou obě složky k sobě vázány elektrostatickými silami. Ty se vyskytují v nukleárních substancích stejně jako v cytoplasmě. Dají se považovat za prostor pro syntézu proteinů a enzymů. Příkladem mohou být nukleoproteiny z kvasnic a brzlíku jsou také viry, které mohou být považovány za velké molekuly nukleoproteinů a nukleohistony z materiálů bohatých na jádro jako jsou žlázové tkáně, nuclein. (Jain 2005: 209)

Derivované proteiny jsou proteiny vznikají nejčastěji působením tepla, enzymů, nebo chemických činidel. Tato skupina obsahuje také uměle vytvořené polypetidy a Jian je rozděluje na primárně derivované proteiny a sekundárně derivované proteiny. (Jain 2005: 209) Primárně derivované proteiny jsou derivovány z proteinů, ve kterých není velikost modifikována jiným materiálem. Jedná se o proteany, metaproteiny nebo infraproteiny a koagulační proteiny. Tato skupina se dále dělí na proteany, metaproteiny nebo infraproteiny a koagulační proteiny. Proteany jsou nerozpustné ve vodě a objevují se jako první produkt při působení kyselin, enzymů nebo vody na proteiny. Jedná se například o edestan, který je derivován z edestinu, a myosan derivován z myosinu. Metaproteiny nebo infraproteiny rovněž nejsou rozpustné ve vodě, ale rozpouští je zředěné kyseliny nebo alkaloidy. Jsou produkovány působením kyseliny nebo alkaloidu na protein v rozmezí 30 – 60 °C . Např.

[22] Absorpce, v tomto případě lipidů, je definován jako přenos vody a substancí rozpuštěných ve vodě do buňky, tkáně nebo organismu. (Jian 2005: 1093)

acid and alkalické metaproteiny. Koagulační proteiny jsou nerozpustné ve vodě, a jsou produkovány působením tepla nebo alkoholu na protein, na příklad koagulační vaječný bílek. (Jain 2005: 209)

Sekundárně derivované proteiny jsou derivovány z proteinů, ve kterých proběhla hydrolýza. Výsledné molekuly jsou zpravidla menší, než původní proteiny. Patří mezi ně protáza, peptony a polypeptidy. Proteáza je rozpustná ve vodě, lze ji koagulovat teplem. Vzniká, když hydrolýza postupuje za **level** metaproteinu. Primární proteáza je koagulována částečnou saturací s (NH4)2SO4 a sráží se při působení HNO3 a kyseliny pikrové. Sekundární proteázu lze koagulovat úplnou saturací pomocí (NH4)2SO4, ale nesráží se za působení HNO3 nebo kyseliny pikrové. Peptony jsou rozpustné ve vodě, nakoagulují působením tepla. Vznikají působením zředěných kyselin nebo enzymů ve chvíli, kdy hydrolýza překročí proteázu. Nelze je koagulovat pomocí (NH4)2SO4, ani se nesráží při působení HNO3 nebo kyseliny pikrové. Poslední podskupinou vyčleněnou v této klasifikaci jsou polypeptidy, které jak bylo výše zmíněno vznikají kombinací dvou nebo více jednotek amino kyselin. Proteiny je možné považovat za polypeptidy, kterým je vlastní velmi dlouhý řetězec. (Jain 2005: 209-210)

2.4 Klasifikace podle biologické funkce

Různé proteiny naplňují odlišné funkce v závislosti na své fyzické a chemické struktuře a umístění uvnitř buňky. Na základě těchto diferencí mohou být proteiny klasifikovány podle následujících kategorií, které jsou založeny na základě jejich odlišných metabolických funkcí. Jedná se o enzymy, strukturální proteiny, transportní proteiny, nutriční proteiny, kontrakční a pohyblivé proteiny, defenzivní proteiny, regulační proteiny a v neposlední řadě toxické proteiny. (Jain 2005: 210) Vzhledem k tomu že enzymy jsou tou nejrozmanitějších skupinou z výše uvedených proteinů, bude jim v práci věnována větší pozornost.

2.4.1 Enzymy

Albert Lehninger v publikaci *Princinples of Biochemistry* (2012) uvádí, že pro existenci života jsou nutné dvě podmínky. První z nich je schopnost sebereplikace živé entity, druhou pak schopnost organismu efektivně a selektivně katalyzovat chemické reakce, což umožňují právě enzymy. Díky tomu jsou enzymy klíčové pro všechny biochemické procesy. (Lehningher 2012: 190) Enzym tedy můžeme definovat *jako „[...] protein s katalytickými funkcemi vzhledem ke své specifické aktivaci“*. (Dixon, Webb 1964: 5)

V organizovaných sekvencích katalyzují postupné reakce, které rozkládají molekuly, konzervují a transformují chemickou energii a tvoří makromolekuly z jednoduchých prekurzorů. Je nutné zmínit, že téměř všechny enzymy patří mezi proteiny, s výjimkou malé skupiny katalytických RNA molekul. Pro účely této práce se však budeme dále zabývat pouze těmi enzymy, které mezi proteiny patří. (Lehningher 2012: 190)

Enzymy jsou skupinou, ve které se vyskytuje rozmanité spektrum proteinů, přičemž nejspecializovanějšími jsou proteiny s katalytickou aktivitou. Všechny chemické reakce organických biomolekul jsou katalyzovány právě pomocí enzymů. Většina enzymů patří mezi globulární proteiny. Z chemického hlediska jsou některé enzymy jednoduchými proteiny, obsahující pouze pozůstatky aminokyselin. Jiné enzymy jsou komplexními proteiny, které jsou tvořeny z velké části proteiny (apoenzyme) a zbylými nebílkovinnými částmi (prostetická skupina), které jsou spojené s proteinovou skupinou. Enzymy katalyzují spoustu rozličných reakcí. Ureáza, amyláza, kataláza, cytochrom C a alkoholdehydrogenáza jsou jen některými příklady enzymových proteinů. (Jain 2005: 211- 212)

Enzymy, stejně jako ostatní proteiny, mají molekulární váhu rozptýlenou v rozmezí od dvanácti tisíc do jednoho milionu. Některé enzymy nevyžadují pro svou aktivitu žádné jiné chemické skupiny než svá aminokyselinová rezidua. (Lehningher 2012: 191-192) Enzymové aktivity jsou závislé na tom, že neproteinové prostetické skupiny jsou blízce sdružené s proteinovými apoenzymy. V některých případech je prostetická skupina vázána na proteinovou jednotku volně a lze ji separovat dialýzou a přitom být stále nepostradatelnou pro aktivitu enzymu. V takovém to případě se dialyzovatelná prostetická skupina nazývá koenzym, nebo kofaktor. (Jian 2005: 350) V případě kofaktorů se jedná o jeden nebo více neorganické iontů, jako mohou být například Fe2, Mg2, Mn2, nebo Zn2, nebo komplexní organické nebo metalorganické molekuly nazývající se koenzymy. Některé enzymy vyžadují pro svou aktivitu koenzym a ještě jeden další iont kovu. Koenzym nebo iont kovu, který je velmi úzce vázán, případně kovalentně vázán k enzymu se nazývá prostetická skupin. Kompletní katalyticky aktivní enzym spolu se svou koenzymovou vazbou a/nebo iontem kovu se nazývá holoenzym. Takováto část proteinu se nazývá apoenzym nebo apoprotein. Koenzymy se chovají jako nositelé přechodových jevů specifických funkčních skupin. Většina z nich je derivována z vitamínů, organických živin potřebných v malém množství ve stravě. (Lehningher 2012: 191-192)

Většina enzymů pracuje uvnitř buněk, kterými jsou produkovány, proto se nazývají vnitrobuněčné enzymy, tedy endozymy. Jedná se z velké části o rostlinné enzymy. Endoenzymy katalyzují metabolické reakce buňky, proto jsou také nazývány metabolickými enzymy. Proti tomu se některé enzymy vyskytují vně buňky, kde katalyzují reakce mimo prostředí buňky. Proto jsou nazývány vně buněčnými enzymy, exoenzymy. Jedná se nejčastěji o trávicí enzymy, které katalyzují rozklad komplexních substancí na jednodušší, které buňky mohou snadněji absorbovat. (Jian 2005: 334)

Chemicky se enzymy dělí do dvou kategorií. Jedná se o enzymy skládající se z jednoduchých proteinů a komplexních proteinů. Enzymy skládající se z jednoduchých proteinů obsahují pouze jednoduché proteiny jako ureáza, amyláza, papain atd. Enzymy složené z komplexních proteinů obsahují konjugované proteiny. Mají tedy proteinovou část nazývající se apoenzym a neproteinovou část, prostetickou skupinu. Tyto dvě části dohromady tvoří holoenzymy jako jsou kataláza, cytochrom c atd. (Jian 2005: 350) Enzymy jsou dále klasifikovány podle toho jakým způsobem jsou schopny katalyticky reagovat. Podle těchto kritérií se dále dělí na oxidoreduktázy, transferázy, hydrolázy, lyázy, izomerázy a ligázy. Oxidoreduktázy přenášejí elektrony (iontové hydridy nebo atomy vodíku), transferázy katalyzují skupinové přenosové reakce, hydrolázy pak způsobují hydrolýzu (přenos funkčních skupin do vody). Lyázy umožňují přidání skupin ke dvojitým vazbám, případně formování dvojité vazby odebráním skupin. Izomerázy pak přenášejí skupiny uvnitř molekul tak, aby vytěžovaly izomerické formy. V neposlední řadě pak ligázy katalyzují formace C – C, C – S, C – O, a C – N vazby. (Lehninger 2012: 192)

Mnohé enzymy se vyskytují ve více než jedné molekulární formě stejného druhu organismu, stejné tkáni, dokonce ve stejné buňce. V takových případech katalyzují rozličné formy enzymů stejné reakce, ale vzhledem k tomu, že mají rozdílné kinetické vlastnosti a různou kompozici aminokyselin, mohou být separovány vhodnými technikami, jako je například elektroforéza[23]. Tyto rozdílné formy enzymů se nazývají izoenzymy nebo také izozymy. Izoenzymy jsou v přírodě hojně rozšířené. Přes sto enzymů je známých svou izomickou povahou a vyskytují se tudíž ve dvou nebo více molekulárních formách. Na

[23] Elektroforéza, někdy také nazývána kataforéza, je pohyb elektricky nabitých částic v tekutině, který je ovlivňován elektrickým polem. Elektroforéza je používána k analýze a separaci proteinů, prptože pozitivně a negativně nabité boční řetězce proteinů způsobují, že v elektrickém poli se chovají jako aminokyseliny. (Encyklopaedia Britannica)

příklad laktátdehydrogenáza (LDH) je enzymem, který existuje v pěti rozdílných formách v různých orgánech většiny obratlovců. (Jian 2005: 343)

Co se týká biologické role enzymů, ty sehrávají mnohé role a mají nejrůznější aplikace. Pro názornost jsou dále uvedeny příklady ilustrující rozmanitost aplikací enzymů. Jedná se například o využití ve výrobě vína, sýru, chleba a dalších pochutin. Dále jsou enzymy používán ve výrobě tkanin a látek. V těle živočichů slouží enzymy zejména při zažívání (např. pepsin, papain, amyláza, pankreatické enzymy apod.). K hojení ran pak slouží proteolytické enzymy extrahované z prasečí slinivky a využívají se k léčbě kožních nemocí, proleženin a mokvajících ran. Jedná se o enzymy jako streptodornáza, ficin a trypsin. (Jian 2005: 344 -345)

Specifické enzymy se užívají také při rozboru biochemikálií při klinické analýze. Na příklad urikáza a ureáza jsou využívány při determinaci kyseliny močové a močovin v krvi. Dalším využitím může být rozklad krevních sraženin. Pomocí urokinázy jsou léčeny krevní sraženiny v mozku a tepnách, streptodekináza pak umožňuje rozpouštění krevních sraženin v cévách. Zajímavou aplikací ve zdravotnictví je pak využití enzymů při změnách krevních skupin. V Japonsku došlo v 90. letech k úspěšnému použití specifických enzymů v experimentech se změnami krevních skupin. Dalším využitím enzymů v medicíně je také diagnostika hypertenze, vysokého tlaku, probíhá pomocí metody nazývané radio immunoassay procedura, která využívá reninu, proteolytického enzymu vylučovaného ledvinami, přičemž renin se chová jako součást komplexního zpětného mechanismu pro regulaci krevního tlaku. (Jian 2005: 346) V následné analýze popsané ve čtvrté kapitole používá tato práce 15 vzorků enzymů, jedná se o zástupce proteinů α-Amyláza, Celuláza a Lysozom.

2.4.2 Strukturní proteiny

Strukturní proteiny většinou biochemickým reakcím nepodléhají. Zachovávají svou přirozenou formu a pozici orgánů. Buněčná stěna a primární vláknité konstituenty buňky obsahují strukturální proteiny. Kolagen, který má velkou tažnou sílu, je jedním z nejhojnějších živočišných proteinů. Nalézá se ve spojovacích tkáních jako jsou šlachy, chrupavky, základní materiál kosti, oční rohovka apod. Kůže je tvořena téměř výhradně kolagenem. (Jain 2005: 212) Konkrétně je jsou to tři čtvrtiny sušiny (dry weight) kůže, kolagen tvoří až třetinu všech proteinů v lidském těle. V tělech obratlovců je možné najít 28 rozdílných typů kolagenu, složených nejméně ze 46 odlišných polypeptidových řetězců, a

mnohé ostatní proteiny obsahují kolagenové domény. Kolagen je také dodnes nejstarším proteinem, který se prozatím podařilo detekovat[24]. (Shoulders, Raines 2009: 1) Dalším strukturním proteinem je elastin, který je možné nalézt napříč druhy obratlovců, s výjimkou některých primitivních ryb. Elastin má neobvyklou chemickou kompozici, která ovlivňuje jeho charakteristické vlastnosti[25]. Vyskytuje se v elastických vláknech tkání spolu s kolagenem a také v kůži. (Rosenbloom 1984 :1) Jiným příkladem strukturního proteinu je keratin. Termín keratin byl v minulosti používán pro všechny proteiny, extrahované z modifikací kůže, jako jsou rohy, drápy, kopyta apod. Dnes termín keratin zahrnuje všechny proteiny, které formují vlákna se specifickými fyzickými a chemickými vlastnostmi, a jsou produkovány v epitelu obratlovců. (Bragulla, Homberg 2009: 517) Alfa keratin tvoří téměř celou sušinu (dry wight) vlasů, chlupů, vlny, peří, nehtů, drápů, brků, šupin, rohů, kopyt, želvích krunýřů a také velkou část vnější vrstvy kůže. Majoritním komponentem vláken přírodního hedvábí a pavoučích vláken je fibroin. Křídlové závěsy některých druhů hmyzu jsou tvořeny resilinem, který má téměř dokonalé elastické vlastnosti. (Jain 2005: 212) V následné analýze popsané ve čtvrté kapitole používá tato práce 10 vzorků strukturních proteinů, jedná se o zástupce proteinu kolagen a dalších proteinů se strukturní funkcí.

2.4.3 Transportní proteiny

Mnohé proteiny jsou zapojeny do přenosu molekul nebo iontů přes membrány nebo mezi buňkami. Příkladem proteinů schopných přenosu skrz membrány zahrnují enzymy Na^+-K^+ ATPázu a také přenašeč glukózy. (McKee 2015: Chapter 5) Některé živočišné proteiny jsou zapojeny do přenosu základních biologických faktorů různým částem organismu. Hemoglobin erythrocytů na sebe váže kyslík při průchodu krve plícemi, nese ho periferním tkáním, a tam kyslík uvolňuje, aby se mohl účastnit oxidace živin. (Jain 2005: 212) Hemoglobin formuje nestabilní reverzibilní vazby s kyslíkem, přičemž v okysličeném stavu se nazývá oxyhemoglobin. (Encyclopaedia Britannica) Krevní plasma obsahuje lipoproteiny, které nesou lipidy z jater do ostatních orgánů. Ostatní druhy transportních proteinů se vyskytují v plazmových membránách a mezibuněčných membránách všech organismů. Měď je přenášena krví pomocí celuroplasminu. (Jain 2005: 212) V následné analýze popsané ve čtvrté kapitole používá tato práce 15 vzorků transportních proteinů,

[24] Nporušený kolagen byl nalezen v kostech 68 milionů let staré fosilie Tyranosaura rexe. (Shoulders, Raines 2009: 1)

[25] Elastin se skládá z glycinu, prolinu a dalších hydrofobních reziduí a obsahuje několik křížových vazeb, které spojují jednotlivé polypeptidové řetězce do sítě podobné pryži.

jedná se o zástupce proteinů Hemoglobin, Lipoprotein a dalších proteinů s transportní funkcí.

2.4.4 Nutriční proteiny

Tyto proteiny slouží k ukládání živin. Jedním z nejvýznamnějších nutričních proteinů je ovalbumin, hlavní proteine obsaženým ve vaječném bílku. Mléčný protein kasein ukládá aminokyseliny. Semena mnoha rostlin ukládají nutriční proteiny, které jsou potřebné pro růst klíčivých semínek. Feritin, který se nalézá v některých bakteriích a rostlinných i živočišných tkáních, ukládá železo. (Jain 2005: 212) V následné analýze popsané ve čtvrté kapitole používá tato práce 15 vzorků nutričních proteinů, jedná se o zástupce proteinu Kasein a dalších proteinů ukládajících živiny.

2.4.5 Kontrakční a pohyblivé proteiny

Existují některé proteiny dodávají buňce a organismu schopnost měnit tvar nebo se pohybovat. Tubulin je protien, který je základní stavevní jednotkou mikrotubul. Aktin a myosin působí v kontrakčním systému kosterních svalů a také v mnoha nesvalových buňkách. (Jain 2005: 212) Myosin interaguje s aktinem, přičemž proteiny aktinu jsou organizovány do vláken tak, aby tvořily cytoskeleton, který dává sturkturu buňkám a dále může sloužit jako stopa pro pohyb myosinu. Některé myosinové proteiny na sebe vážou jiné proteiny a transportují je mezi buňkami po stopě vytvořené aktinem. Protein myosin je zapojen do buněčného pohybu. Jeho schopnost transportovat materiál a vytvářet sílu prostřednictvím kontrakcí, je činí důležitým v procesu buněčného dělení. (U.S. National Library of Medicine) Kontrakční a pohyblivé proteiny jsou zapojeny také v buněčných pohybech. Například aktin, tubulin a jiné proteiny obsahují cytoskeleton. Cytoskeletální proteiny jsou aktivní v buněčném dělení, endocytóze, exocytóze a zejména v améboidním pohybu bílých krvinek. (McKee 2015: Chapter 5) V následné analýze popsané ve čtvrté kapitole používá tato práce 15 vzorků kontrakčních a pohyblivých proteinů, jedná se o zástupce proteinů Aktin, Myozin a Troponin.

2.4.6 Defenzivní proteiny

Mnoho proteinů brání organismus před invazí jiných druhů, nebo je chrání před zraněními. Imunoglobuliny jsou speciálními proteiny složené z lymfocytů obratlovců, mohou precipitovat nebo neutralizovat invazivní bakterie a viry cizorodých proteinů jiných organismů. Fibrinogen a thrombin, ačkoli se jedná o enzymy, jsou proteiny umožňující srážení krve. Předcházejí ztrátě krve při zranění vaskulárního systému. (Jain 2005: 212)

V následné analýze popsané ve čtvrté kapitole používá tato práce 10 vzorků defenzivních, jedná se o zástupce proteinu Imunoglobulin a dalších proteinů s defenzivní funkcí.

2.4.7 Regulační proteiny

Tyto proteiny napomáhají regulovat buněčnou nebo fyziologickou aktivitu. Vyskytuje se mezi nimi mnoho hormonů, jako je inzulin, který reguluje metabolické cukry, a růstový hormon potřebný pro růst kostí. Buněčná reakce na hormonální děje je často zprostředkována třídou GTP proteinů, G proteiny. Ostatní regulační proteiny se váží na DNA a regulují biosyntézu enzymů a RNA molekul zapojených v buněčném dělení. (Jain 2005: 212) V následné analýze popsané ve čtvrté kapitole používá tato práce 15 vzorků regulačních proteinů.

2.4.8 Toxické proteiny

Existují bílkoviny chovající se jako toxické substance. Jedná se například o hadí jed, bakteriální toxiny a toxické rostlinné proteiny jako je například ricin. Toxické proteiny mají také obranné funkce. (Jain 2005: 212) V následné analýze popsané ve čtvrté kapitole používá tato práce 15 vzorků toxických proteinů, jedná se o zástupce toxinů.

3 Struktura proteinu

Aminokyseliny propojené a zobrazené v ploché dvojdimenzionální reprezentaci polypeptidovéo řetězce nedostatečně zobrazují prostorovou, trojdimenzionální, strukturu proteinu. Z tohoto důvodu je nutné blíže představit struktury, kterých protein může nabývat, protože právě struktura proteinu určuje jeho další charakteristické vlastnosti a funkce. (Whitford 2005: 39) *„K popisu makromolekul proteinů slouží vydělení základních strukturálních stupňů jejich organizace. Konfigurace proteinů rozpoznává čtyři stupně, jejich primární, sekundární, terciární a kvartérní strukturu.“* (Jain 2005: 155) Tři z těchto struktur (primární, sekundární, terciární) mohou existovat v molekulách složených z jednoho polypeptidového řetězce, zatímco kvartérní struktura zahrnuje interakce polypeptidů v rámci více než jednoho řetězce proteinové molekuly. (Jain 2005: 155)

3.1 Primární struktura

Primární struktura proteinů je sekvencí aminokyselin[26], která je specifikována genetickou informací. V průběhu skládání polypeptidového řetězce se formují lokalizované uspořádání vedlejších, avšak ne nutně spolu sousedících aminokyselin, které tvoří sekundární strukturu. (McKee 2012: Chapter 5) Ta se následně může skládat do specifických vzorců (patterns) tak, aby vytvořila trojdimenzionální, tedy terciární strukturu. Další určité proteiny jsou tvořeny dílčími jednotkami podobných nebo odlišných typů polypeptidových řetězců. Tyto dílčí jednotky spolu interagují tak, aby umožnily vznik kvartérní struktury, což následně definuje stupeň polymerizace[27] proteinu. (Jain 2005: 155-156)

Primární struktura proteinu referuje k počtu a sekvenci aminokyselin, konstituujících jednotky polypeptidového řetězce, přičemž proteiny se mohou skládat z jednoho nebo více peptidových řetězců. (Jain 2005: 156) *„Primární struktura proteinu je lineárním pořadím jeho aminokyselinových reziduí v polypeptidovém řetězci.“* (Whitford 2005: 39) Tato struktura vzniká kovalentním spojením jednotlivých aminokyselin pomocí petidových vazeb. (Whitford 2005: 39) Peptidová vazba se vytvoří tak, že dojde ke spojení α – karboxylovou skupinu jednoho amino kyselinového rezidua s α – amino skupinou druhé amino kyseliny. (Jain 2005: 156) To, že jsou polypeptidy kovalentně vázány peptidovými vazbami, hraje roli také v popisování struktury proteinů. Jak bylo výše řečeno první

[26] Je nutné podotknout, že základní primární struktura proteinu se skládá z jednoho nebo více lineárních řetězců aminokyselinových jednotek. (Jain 2005:155)
[27] Jedná se o proces, ve kterém se mnoho malých identických dílčích jednotek (monomerů) chemicky spojí tak, aby vytvořily dlouhý řetězec molekul polymeru. (Jain 2005:1140)

reziduum v polypeptidovém řetězci je navázáno na aminokyselinu (-NH_2), zatímco poslední reziduum v řetězci na sebe váže karboxylovou kyselinu (-COOH). Konce polypeptidového řetězce jsou proto pojmenovány N- terminus a C-terminus. Primární sekvence proteinu je tedy aminokyselinovou sekvencí, kterou konvenčně čteme ve směru od N-terminu k C-terminu[28]. (Twyman 1999: 288)

Jak již bylo zmíněno, každý polypeptid má specifickou sekvenci aminokyselin a interakce mezi aminokyselinovými rezidui determinují trojdimenzionální strukturu proteinu a jeho funkci a vztah k ostatním proteinům. Polypeptidy které mají podobné sekvence aminokyselin a pocházejí ze stejného dědičného genu se pak nazývají homologické. Komparace sekvencí mezi homologickými polypeptidy jsou využívány při hledání genetických vztahů různých druhů.[29] Rezidua aminokyselin, jejichž homolické proteiny (invarianty) jsou identické, jsou považovány za esenciální pro funkci proteinu. (McKee 2012: Chapter 5)

Každý protein je tedy definován svou unikátní sekvencí aminokyselinových reziduí a všechny subsekventní úrovně jeho organizace (sekundární, super sekundární, terciární a kvarterální) jsou založeny právě na primární struktuře. (Whitford 2005: 39) Primární struktura polypeptidu do jisté míry determinuje jeho strukturu terciární. Teplo a určité chemické látky, jako jsou kyseliny a močovina (NH_2CONH_2), způsobují ztrátu biologické aktivity proteinu, protože narušují slabé nekovalentní vazby, které stabilizují sekundární a terciární strukturu. Předtím, než jsou narušeny nekovalentní vazby, je protein ve své přirozené konformaci. Po narušení vazeb je protein ve stavu denaturovaném, jeho konformace jsou náhodné. Hydrofobické segmenty, které jsou za normálních okolností v jádru vodou rozpustných proteinů stálé, se po denaturaci vážou na hydrofobické segmenty ostatních denaturovaných proteinů tak, aby vytvořily celky, které jsou ve vodě nerozpustné. Denaturovat například enzymy je ovšem velmi obtížné, protože obsahují disulfidické můstky, které udržují pohromadě terciární strukturu. (Tropp 2008: 59)

[28] Abychom mohli primární strukturu číst, přeložíme trojpísmenný nebo jednopísmenným kód ve směru z leva do prava, od amino terminálu (N-terminal) ke karboxylovému terminálu(C-terminal). (Whitford 2005:39)

[29] Například sekvence homologů mitochondriálního redox proteinu, cytochromu c, jsou využívány k při studiích evoluce druhů. Sekvenční komparace cytochromu c, molekuly zásadní v produkci energie, mezi četnými druhy ukzuje významné množství zachovaných sekvencí. (McKee 2012: Chapter 5)

Některé proteiny jsou příbuzné právě kvůli podobnosti v primární struktuře[30]. Pokud je dostatečná část sekvence identická, je snazší najít mezi nimi rozdíly a pokud nastane změna v primární sekvenci proteinu, velmi často se týká dvou blízce příbuzných reziduí[31]. (Whitford 2005: 39)

3.2 Sekundární struktura

Na základě primární struktury proteinu vzniká jeho struktura sekundární. *„Sekundární struktura je lokální struktura polypeptidového řetězce nebo prostorový vztah aminokyselinových reziduí vzájemně blízkých v primární sekvenci.“* (Whitford 2005: 39) Sekundární struktura tedy referuje k prostorovému vztahu mezi aminokyselinami, které jsou v sekvenci ve vzájemné blízkosti. Na základě povahy hydrogenových vazeb (intramolekulárních nebo intermolekulárních) jsou rozlišovány dva typy sekundárních struktur α-šroubovice (dále α-helix) a β-skládaný list (dále β-pleated sheet). (Jain 2005: 160) Obě formy uspořádání, α-helix a β-pleated sheet, jsou stabilizované lokalizovanými vodíkovými vazbami mezi karbonylem a N-H skupinou v páteřním polypeptidu. Protože jsou peptidové vazby rigidní, α-karbony jsou otočnými body polypeptidového řetězce. (McKee 2012: Chapter 5) Uvnitř sousedících úseků polypeptidového řetězce dochází k obvyklým interakcím jako jsou hydrogenové vazby, které umožňují vznik α-helixů β-pleated sheets, které tvoří sekundární strukturu proteinů. Určité kombinace α-šroubovic β-pleated sheets tvoří svazky tak, aby vytvořily kompaktně skládané globulární jednotky, které se nazývají proteinové domény. Domény jsou obvykle konstruovány částí polypeptidového řetězce, který obsahuje 50 až 350 aminokyselin, a zdají se být modulárními jednotkami, které konstruují proteiny. Zatímco malé proteiny mohou obsahovat pouze jedinou doménu, ve větších proteinech můžeme nalézt několik domén, které jsou často spojeny otevřenými délkami polypeptidového řetězce. (Alberts 2007: 205)

α-helix je rigidní tyčovitá struktura, která vzniká když polypeptidový řetězec nabude tvaru pravotočivé šroubovice. Vodíkové vazby se tvoří mezi N-H skupinou každé aminokyseliny a karbonylovou skupinou aminokyselin se čtyřmi rezidui. Aminokyselinové R skupiny přesahují vně helixu. (McKee 2012: Chapter 5) α-helixová struktura se ze všech

[30] Například myoglobin**,** protein ukládající kyslík, který se nachází v různých organismech, ukazuje podobnost v sekvenci 153 reziduí u člověka a velryby. (Whitford 2005:39)

[31] Například na 118. Pozici má člověk reziduum lysinu, zatímco velrybí myoglobin má na této pozici argininové reziduum. Arginin i lysin jsou aminokyseliny, které obsahují pozitivně nabitý boční řetězec a tato změna se nazývá konzervativní transpozice. (Whitford 2005:39)

ostatních možných formuje nejčastěji. Částečně je tomu proto, že α-helixy umožňuje optimálnější použití hydrogenových vazeb. Tato struktura je stabilizována hydrogenovou vazbou mezi atomem vodíku připojeném k elektronegativnímu atomu dusíku peptidové vazby a elektronegativnímu atomu karbonylové skupiny čtvrté aminokyseliny na boku terminálního amino-řetězce téže peptidové vazby. (Lehninger 2012: 120) Kvůli strukturálním omezením nejsou některé aminokyseliny schopny nabýt α-helixové struktury.[32] Sekvence aminokyselin s velkým počtem nabytých aminokyselin (například aminokyseliny glutamová a asparagová) a objemných R skupin (tryptofan) jsou rovněž nekompatibilní s helixovou strukturou. (McKee 2012: Chapter 5)

β-pleated sheet vzniká, když se vedle sebe seřadí dva nebo více segmentů polypeptidových řetězců. Každý jednotlivý segment se v tomto případě nazývá β-vláknem. Ty mají tendenci se napínat, spíše něž se stáčet. β-pleated sheets jsou stabilizovány hydrogenovými vazbami, které se formují mezi polypeptidovými páteřními N-H skupinami a karbonylovými skupinami a přilehlými řetězci. β-pleated sheets jsou paralelní, nebo neparalelní. V neparalelních strukturách jsou hydrogenové vazby v polypeptidovém řetězci uspořádány ve stejném směru, v neparalelním ve směru opačném. Občasně se mohou vyskytovat β-sheets, které jsou zároveň paralelní i neparalelní. Mnoho globulárních proteinů obsahuje kombinaci α-helix a β-pleated sheet sekundární struktury, tzv. supersekundární struktury[33] (motivy). (McKee 2012: Chapter 5)

3.3 Terciární struktura

Terciární stuktura reprezentuje složený polypeptidový řetězec. Je definována jako *„ […] prostorové uspořádání aminokyselinových reziduí, které jsou v primární struktuře oddělené nebo více zhuštěné jako celková topologie formována polypeptidem.“* (Whitford 2005: 52) Pro malé globulární proteiny složené ze 150 nebo méně reziduí zahrnuje složená struktura sférickou kompaktní molekulu složenou z jednotek sekundární struktury a vyskytuje se v ní také stuktura částečně nepravidelná.[34] Terciární struktura proteinu

[32] Například glycin obsahuje R skupinu (atom vodíku) je tak malý, že polypeptidový řetězce je až příliš flexibilní. Naproti tomu aminokyselina prolin obsahuje rigidní kruh, který zabraňuje rotaci $N—C_\alpha$ vazby. Prolin také neobsahuje N—H skupinu, která by umožňovala vytvořit hydrogenovou vazbu uvnitř řetězce, která je nezbytná pro vznik α-helixové structury. (McKee 2012: Chapter 5)

[33] Jednotky βαβ jsou tvořeny dvěma paralelními β-pleated sheets, které jsou spojeny α-helix segmentem. Tato struktura je stabilizována hydrofobní interakcí mezi nepolárními postranními řetězci, které vyčnívají z interakčních ploch β-strands a α-helix. (McKee 2012: Chapter 5)

[34] Narušená nebo nepravidelná struktura v proteinech je běžně omezena na N-terminal nebo C-terminal, méně často pak na smyčkové (loop) oblasti uvnitř proteinu, nebo spojnic jedné a více domén. (McKee 2012: Chapter 5)

odpovídá jeho skládání (folding). Skládání proteinu vzniká propojováním sekundárních struktur, které formují kompaktní globulární molekulu. Elementy sekundární struktury na sebe vzájemně působí prostřednictvím hydrogenových vazeb (uvnitř β-pleated sheets), dále jsou vztahy závislé na disulfidických můstcích, elektorstatických interakcích, hydrogenových vazbách a van der Waalsových silách[35]. (McKee 2012: Chapter 5)

Ačkoli globulární proteiny často obsahují signifikantní množství elementů sekundárních struktur, existují další faktory, které přispívají k jejich struktuře. *„Termín terciární struktura označuje unikátní trojdimenzionální strukturu, které globulární proteiny nabývají při skládání do své přirozené, biologicky aktivní, konfirmace, pokud se k nim připojuje i prostetická skupina.“* (McKee 2012: Chapter 5) Skládání proteinu je proces, při kterém neorganizovaná nově se syntetizující molekula nabývá vysoce organizované struktury, nastává jako důsledek interakcí mezi postranními řetězci v jejich primární struktuře. Terciární struktura má několik důležitých charakteristických rysů. Například mnoho polypeptidů se skládá tak, že rezidua aminokyselin, které jsou v primární struktuře daleko od sebe, se k sobě v terciární struktuře přiblíží. Dále terciární struktura umožňuje kompaktnost globulárních proteinů a to díky svému efektivnímu skládaní polypeptidů. V průběhu procesu skládání je většina molekul vody vyloučena ze vnitř proteinu, čímž umožňuje interakci s polárními i nepolárními skupinami. Dalším specifikem terciární struktury je to, velké globulární proteiny obsahující více než 200 aminokyselinových reziduí často obsahují několik kompaktních jednotek, domén. Domény jsou typicky strukturálně nezávislé segmenty se specifickými funkcemi (vázání iontů nebo malých molekul). Hlavní jádro (core) trojdimenzionální struktury domény se nazývá sklad (dále FOLD**)**. Domény jsou klasifikovány na základě jádrového (core) motivu jejich struktury. Jedná se o α –domény, β-domény, α/β-domény a α + β-domény (α, β, α/β, α + β). α -domény se skládají výlučně z α-helixů, β-domény zase z antiparalelních β-strands (βαβ motivy). α/β-domény obsahují různé kombinace α-helixů a β-strands (βαβ motivů). α + β domény jsou primárně β-sheets s jedním vně ležícím α-helixem. Většina proteinů obsahuje dvě nebo více domén**.** Velká část proteinů tvořící eukaryotické buňky obsahuje mnohé duplikace nebo nedokonalé kopie jedné nebo více domén, které jsou spojeny. Doménové modely jsou kódovány kvůli genetické sekvenci vytvořené duplikací genů, jedná se tedy o kopie genů, které vznikají z chyb v replikaci DNA.

[35] Van der Waalsovy síly jsou výsledkem indukovaných elektrických interakcí mezi blízkými atomy nebo molekulami, jejichž záporně nabité elektronové mraky mohou kolísat v okamžitém čase. Tyto fluktuace umožňují vzájemnou přitažlivost pozitivně nabitých jader a elektronů blízkých atomů. **(Garret, Grisham** 2016: 12)

Takové sekvence jsou využívány žijícími organismy k vytvoření nových proteinů. Na příklad strukturální domény imunoglobulinu se vyskytují nejen v protilátkách, ale také v různých proteinech buněčného povrchu. (McKee 2012: Chapter 5) Dalším proteinem, který má terciární strukturu, je myoglobin. Tento protein umožňuje přenos kyslíku ve svalových tkáních, rovněž je obsažen v některých enzymech, jako jsou lisozym, pepsin, trypsin a jiné. (Gavriliuc 2011: 11) Existuje několik interakcí, které stabilizují terciární strukturu. Jsou to například hydrofobní interakce, elektrostatické interakce, vodíkové vazby, kovalentní vazby, hydratace apod. (McKee 2012: Chapter 5)

3.4 Kvartérní struktura

Mnoho proteinů, zvláště ty s vysokou molekulovou vahou, obsahuje více než jeden polypeptidový řetězec. Interakce mezi těmito řetězci jsou základem kvartérní struktury proteinů. Tyto interakce jsou totožné s těmi, které napomáhají tvorbě terciární struktury. Jedná se o disulfidické můstky, hydrofobní interakce a vodíkové můstky. Jedinou výjimkou oproti terciární struktuře je ta, že výše zmíněné interakce působí mezi jedním a více polypeptidovými řetězci. (Whitford 2005: 62) Jak bylo výše zmíněno, každý polypeptidový komponent se nazývá dílčí jednotkou. Dílčí jednotky v proteinovém komplexu mohou být identické, nebo rozdílné. Proteiny s několika dílčími jednotkami, z nichž některé nebo všechny jednotky jsou identické, se nazývají oligomery. Oligomery se skládají z protomerů, které mohou obsahovat jednu nebo dvě dílčí jednotky. Velké množství oligomerických proteinů obsahuje dvě nebo čtyři dílčí jednotky protomerů, nazývané dimery (dimers) nebo tetramery (tetramers). Existuje několik důvodů pro běžný výskyt proteinů s několika dílčími jednotkami. Jedním z nich je například to, že syntéza separačních dílčích jednotek může být efektivnější, než značné navyšování délky jednoho polypeptidového řetězce. Dalším důvodem je pak zvýšení efektivity při náhradě opotřebeného nebo poškozeného komponentu v supramolekulárních komplexech jako jsou kolagenová vlákna. V neposlední řadě komplexní interakce několika násobných jednotek napomáhá regulovat biologickou funkci proteinu. (McKee 2012: Chapter 5) Čtvrtý stupeň komplexity proteinových struktur tedy může vzniknout v proteinu s identickými nebo neidentickými dílčími jednotkami. Kvartérní struktura totiž umožňuje jak formaci prostoru pro katalýzu a tvorbu vazeb, tak propojení mezi dílčími jednotkami. (Whitford 2005: 62)

Příkladem proteinu s kvartérní strukturou je hemoglobin[36], protein který přenáší kyslík v erythocytech, a tvoří zhruba 90 % červených krvinek. Jedná se o tetramerický protein, tedy takový, který je tvořen čtyřmi polypeptidovými řetězci. Tyto čtyři řetězce, z nichž vždy dva a dva jsou stejného typu, jsou spojeny nekovalentními interakcemi. Každý řetězec obsahuje heme skupinu a jeden kyslík, který je váže. (Jian 2005: 180) Vzhledem k tomu že všechny čtyři dílčí jednotky jsou tak úzce propojeny, tetramer hemoglobinu je nazýván molekula, a to i přesto že se mezi peptidovými řetězci všech čtyř dílčích jednotek nevyskytují žádné kovalentní vazby. V ostatních proteinech s kvartérní strukturou jsou dílčí jednotka k sobě vázány kovalentními vazbami, dislufidickými můstky. (Encyklopeadia Britanicca)

[36] Hemoglobin A, základní hemoglobin, se skládá z dvou α řetězců a dvou β řetězců. (Jian 2005: 180)

4 Analytická část

Následující kapitola je věnována analýze vybraných vzorků proteinů klasifikovaných do skupin podle biologických funkcí, které mohou proteiny zastávat, jak bylo popsáno výše ve druhé kapitole této práce. Kvantitativní lingvistická analýza používá software Quantitative Text Index Analyzer[37] (QUITA), s jehož pomocí jsou sledovány n-gramy[38] (konkrétně trigramy[39]) stringů a jejich projevy v modelu Bag of Words. Jednotlivé typy stringů jsou analyzovány ve skupinách s podobnou délkou pohybující se v rámci desítek až stovek aminokyselinových reziduí tak, aby byl zohledněn vliv délky na jednotlivé lingvistické metriky, přičemž ze stejného důvodu je použita kosinova vzdálenost. Analýza se zaměřuje zejména na indexy Tokens (tokeny), Types (typy), TTR (type token ratio), Entropy (entropie), Giniho Koeficient a RR (repeat rate). Jednotlivé lingvistické metriky dále slouží jako vlastnosti textu v data miningové analýze (hierarchické shlukování, MDS, PCA). Zvolené indexy a pojmy data miningové analýzy jsou popsány, vysvětleny a specifikovány níže.

Tato kapitola je dále členěna na podkapitoly, které se věnují výběru a popisu vzorků proteinů použitých v analýze a jejich popisu na základě modelu Bag of Words a vybraných indexů užívaných ke kvantitativní lingvistické analýze. Takto jsou popsány jednotlivé skupiny proteinů, tedy enzymy, strukturní proteiny, transportní proteiny, nutriční proteiny, kontrakční a pohybové proteiny, defenzivní proteiny, regulační proteiny a toxické proteiny. Následně je provedena celková analýza pro všechny výše zmíněné skupiny proteinů najednou spolu s celkovou interpretací. Celá tato kapitola se snaží odpovědět na výzkumnou otázku vytyčenou v úvodu, tedy zda a jakým způsobem je možné provést kvantitativní lingvistickou analýzu proteinů klasifikovaných podle jejich biologické funkce a popsat tak charakteristické znaky jejich genetického textu pomocí programu QUITA (Quantitative Text Index Analyzer), modelu Bag of Words a vybraných indexů kvantitativní lingvistiky. Popisu

[37] QUITA je software umožňující provádět kvantitativní lingvistickou analýzu, tedy vykonávat automatizované výpočty pro velké množství vstupních dat. (Matlach 2014:7)

[38] Encyklopedický slovník češtiny definuje n-gramy jako: *„Prosté zřetězení, posloupnost n jednotek stejného druhu (písmen, častěji však slov) v textu. Mezi členy n. se automaticky nepředpokládá lingvistický vztah. O n. se mluví většinou jako o typech, při jejich popisu tedy není zkoumána jejich konkrétní realizace, ale celková frekvence, příp. distribuce v textech atp. […] Studium n. je výrazem snahy objevovat v povrchově pouze syntagmatickém toku textu opakovaně se vyskytující shluky jednotek, které mají languovou povahu.“* (CzechEncy - Nový encyklopedický slovník češtiny)

[39] Trigramy vznikají zřetězením více slov, nebo jednotek. Jedná se o n-gram, kde n je rovno 3. (CzechEncy - Nový encyklopedický slovník češtiny)

vzorků, samotná analýza i její interpretace využívá teoretického základu, který byl vytyčen v předchozích kapitolách.

Popis provedeného experimentu v programu QUITA

Pečlivě označené vzorky ve formátu fasta byly nahrány do programu QUITA. Pro kalkulaci byl zvolen oddíl Kvantitativní lingvistika, model Bag of Words, do něhož byly přidány zvolené soubory ve formátu textového souboru (txt). V nastavení Processing Designer byla navolena tokenizace, v tomto případě byl zvolen tokenizér určený pro aminokyseliny. Z tokenizovaných souborů byly následně vytvořeny n-gramy, konkrétně trigramy (3-gramy). Původní defaultní nastavení bylo zachováno, byla zvolena kosinova vzdálenost, a dále byl zachován formát hierarchického klastrování, tedy ward.D2. Kvůli velikosti a viditelnosti byl v nastavení zvolen formát 800x600px.[40] Výsledkem tohoto nastavení pro model Bag of Words jsou pak grafy HClust (hierarchické klastrování) a MDS.

Pečlivě označené vzorky byly nahrány do programu QUITA. Pro kalkulaci byl zvolen oddíl Kvantitativní lingvistika, výpočet Indexů, do něhož byly přidány zvolené soubory ve formátu textového souboru (txt). V nastavení Processing Designeru byla navolena tokenizace, v tomto případě byl zvolen tokenizér určený pro aminokyseliny. Z tokenizovaných souborů byly následně vytvořeny n-gramy, konkrétně trigramy (3-gramy). Původní defaultní nastavení bylo zachováno. Výsledkem byla tabulka indexů. V oddílu Data Exploration byla v nastavení Calculation Settings zvolena kosinova vzdálenost. V nastavení Data Settings pak byly zvoleny indexy Tokeny, Typy, TTR, Entropie, Giniho Koeficient a RR. Výsledkem pak bylo několik grafů, HClust, MDS a PCA, opět ve formátu 800x600px.

Tato analýza využívá při pozorování stringů trigramy. Jedním z faktorů pro volbu trigramů je velikost abecedy, která se skládá z 20 písmen označujících základní aminokyseliny, které se řadí do sekvence proteinu. Vzhledem k možnosti kombinací utvořených 20 písmeny v sekvenci o délce desítek až stovek dílčích jednotek, trigramy oproti bigramům ve vyšší míře diverzifikují jednotlivé vzorky. Zároveň je výskyt trigramů dostatečně častý na to, aby reflektoval společné znaky vzorků. Trigramy byly zvoleny jako vhodné na základě pozorování při provádění experimentu v softwaru QUITA.

[40] V případě grafů pro celkovou analýzu všech vzorků byl zvolen formát A2 z důvodu čitelnosti.

V experimentu provedeném v softwaru QUITA byla navolena kosinova[41] vzdálenost vzhledem k nestejné délce vložených textů. Kvantitativní analýza pak využívá šesti indexů. Jedná se o tokeny[42], typy[43], TTR[44], entropie[45], Giniho Koeficient[46] a RR.[47] Výsledkem experimentu jsou grafy hierarchického shlukování[48] (Hclust), multidimenzionálního škálování[49] (MDS) a analýzy hlavních komponent[50] (PCA).

4.1 Výběr vzorků a jejich popis

4.1.1 Vzorky enzymů a jejich popis na základě modelu BoW a vybraných indexů

Pro analýzu enzymů využívá tato práce patnácti vzorků, které jsou dále řazeny do skupin po pěti, podle vymezených podskupin, které slouží jako příklady zástupců proteinů s katalytickou funkcí. Jednotlivé dále vydělené podskupiny jsou proteiny α- amyláza (vzorky označené zkratkou AA), celuláza (vzorky označené zkratkou CL) a lysozym (vzorky označené zkratkou LS), přičemž všechny tyto enzymy patří mezi globulární proteiny, mají tedy kulovitý nebo vejcovitý tvar. α-amylázy lze podle zdroje molekuly zařadit mezi živočišné proteiny, protože se nacházejí v zažívacím traktu živočichů, například ve slinách a pankreatických šťávách. (Jain 2005: 117) Celuláza je skupina enzymů, které umožňují

[41] Kosinova vzdálenost je standartní způsob kvantifikace podobností mezi dvěma texty. Jedná se o reprezentaci jejich vektorové vzdálenosti. (Ali 2011: 19)
[42] Encyklopedický slovník češtiny definuje token jako*: „Nejmenší jednotka textu, většinou grafické slovo, resp. jedna jeho realizace.“* (Nový encyklopedický slovník češtiny)
[43] Encyklopedický slovník češtiny definuje type jako jednotku abstrakci, jehož realizací je právě token. Proti tomu type je dekontextualizovaná jednotka, která je schopna nabývat takových vlastností, jako je například frekvence. (Nový encyklopedický slovník češtiny)
[44] Encyklopedický slovník češtiny definuje TTR následujícím způsobem: „*TTR (token-type ratio) je poměr počtu různých slov (typů) k celkovému počtu všech slov v korpusu (někdy vyjádřený v procentech). Pokud je poměr vysoký, můžeme mluvit o textu s velkou lexikální bohatostí (užívá mnoho různých jednotek), malý poměr značí velkou míru opakování.“* (Nový encyklopedický slovník češtiny)
[45] Entropie se používá v kvantitativní lingvistice při analýzách textu diverzifikovanosti sledovaných jednotek. (Nový encyklopedický slovník češtiny)
[46] Giniho koeficient je jedním z indexů, které používá kvantitativní lingvistika. Jedná se o index slovního bohatství, jehož výpočet získáme pomocí Lorenzovy křivky, která graficky vyjadřuje kumulativní distribuční funkci slov v textu. (Čech 2014: 41)
[47] RR (repeat rate) je index opakování slov. Pomocí tohoto indexu je možné vypočítat bohatství textu.
[48] Hierarchické klastrování je technika oddělování objektů do optimálně homogenních skupin na základě empirického měření jejich podobností. (Johnson 1967: 241)
[49] Multidimenzionální škálování je technika měření jevů, které je zachycováno sadou znaků a formuje měřící stupnice, škály. Multidimenzionální škálování slouží k identifikaci latentních kontinuí v daném znakovém prostoru. (Sociologická encyklopedie)
[50] Analýza hlavních komponent je jednou z metod mnohorozměrné statistické analýzy. Analýza hlavních komponent hledá skryté veličiny, komponenty. Základní charakteristikou každé veličiny je její míra variability, tedy její rozptyl. První hlavní komponenta obsahuje informace o variabilitě původních dat a druhá hlavní komponenta obsahuje informace o rozptylu původních dat, které nejsou obsaženy v první komponentě. Tato metoda umožňuje analyzovat malý počet nekorelovaných hlavních komponent. (Tonhauserová 2013: 16-17)

hydrolýzu celulózy, která je nejbohatším organickým zdrojem potravy, paliva a chemikálií. Nejefektivnější metodou degradace celulózy na užitečné komponenty je enzymatická degradace pomocí celulázy. (Washington Biochemical Corporation) Lysozym je enzym vyskytující se v slzách, nosním mazu, žaludečních šťávách, mléku a vaječném bílku. Jeho katalytická funkce umožňuje rozkládat buněčné stěny bakterií, čímž poskytuje ochranu před invazivními bakteriemi v kůži, v membránách mazu a v mnoha tělesných tekutinách. (Jain 2005: 1128)

Mezi α- amylázy jsou zařazeny následující vzorky:

AA1PIF Pig Alpha-Amylase, pochází z organismu Sus scrofa, dále je klasifikován jako glykosyltransferáza, vzorek má 496 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: metabolické procesy karbohydrátů, metabolické procesy, katabolické procesy karbohydrátů.

AA4X0N Porcine pancreatic alpha-amylase in complex with helianthamide, a novel proteinaceous inhibitor, pochází z organismů Sus strofa a Stichodactyla helianthus, dále je klasifikován jako inhibitor hydrolázy, vzorek má 540 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: metabolické procesy karbohydrátů, metabolické procesy, katabolické procesy karbohydrátů.

AA5E0F Human pancreatic alpha-amylase in complex with mini-montbretin A, pochází z organismu Homo sapiens, dále je klasifikován jako inhibitor hydrolázy, vzorek má 496 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: metabolické procesy karbohydrátů, metabolické procesy, katabolické procesy karbohydrátů, trávení polysacharidů.

AA5KEZ Selective and potent inhibition of the glycosidase human amylase by the short and extremely compact peptide piHA from mRNA display, pochází z organismu Homo sapiens, dále je klasifikován jako inhibitor hydrolázy, vzorek má 507 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: metabolické procesy karbohydrátů, metabolické procesy, katabolické procesy karbohydrátů, trávení polysacharidů.

AA5U3A Ultra High Resolution Crystal Structure of Human Pancreatic Alpha Amylase, pochází z organismu Homo sapiens, dále je klasifikován jako hydroláza, vzorek má 496 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se

následujících biologických procesů: metabolické procesy karbohydrátů, metabolické procesy, katabolické procesy karbohydrátů, trávení polysacharidů.

Mezi celulázy jsou zařazeny následující vzorky:

CL5UHX Structure of cellulase Cel5C_1, pochází z nekultivovaného organismu, dále je klasifikován jako hydroláza, vzorek má 323 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: metabolické procesy karbohydrátů.

CL5ECU The unliganded structure of Caldicellulosiruptor saccharolyticus GH5, pochází z organismu Caldicellulosiruptor saccharolyticus, dále je klasifikován jako hydroláza, vzorek má 555 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: metabolické procesy karbohydrátů, katabolické procesy celulózy.

CL5XBU Crystal structure of GH45 endoglucanase EG27II in apo-form, pochází z organismu Ampullaria crossean, dále je klasifikován jako hydroláza, vzorek má 190 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: metabolické procesy.

CL5LJF Crystal structure of the endo-1,4-glucanase RBcel1 E135A with cellotriose, pochází z organismu nekultivované bakterie, dále je klasifikován jako hydroláza, vzorek má 642 aminokyselinových reziduí, v sekvenci se nachází 1 mutace, účastní se následujících biologických procesů: metabolické procesy karbohydrátů, metabolické procesy.

CL5L6Z Crystal structure of D62A mutant of Thermotoga maritima TmPEP1050 aminopeptidase, pochází z organismu Thermotoga maritima, dále je klasifikován jako hydroláza, vzorek má 662 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: proteolýza, metabolické procesy.

Mezi lysozymy jsou zařazeny následující vzorky:

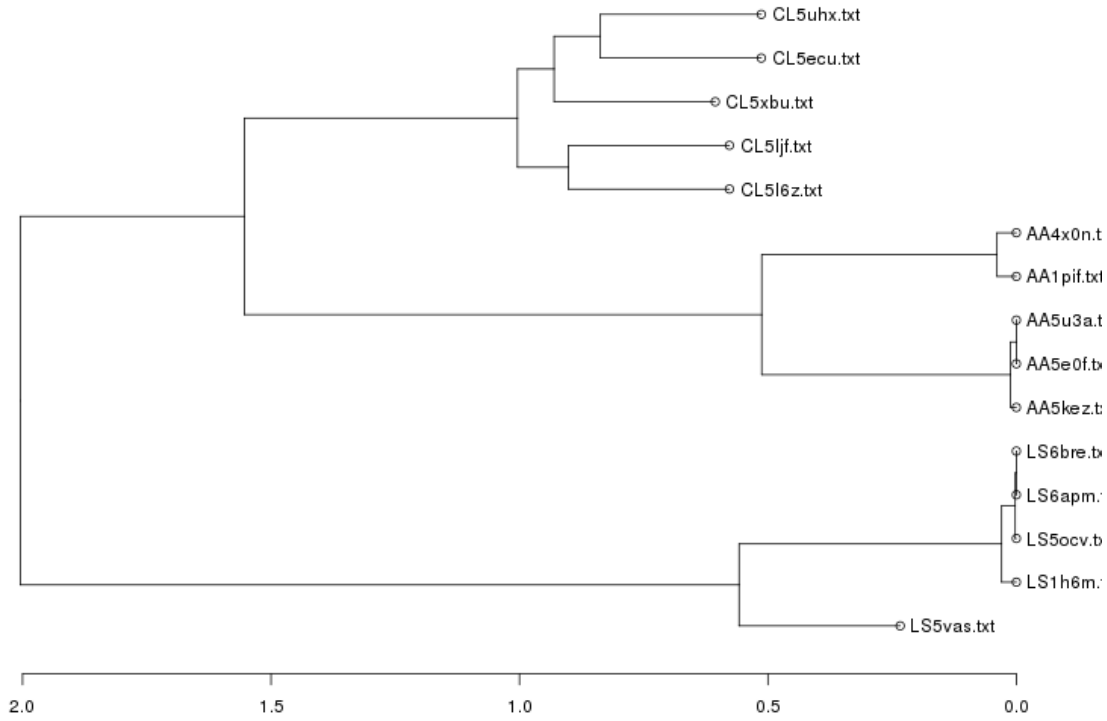
LS1H6M Covalent glycosyl-enzyme intermediate of hen egg white lysozyme, pochází z organismu Gallus Gallus, dále je klasifikován jako hydroláza, vzorek má 129 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: metabolické procesy, katabolické procesy buněčných stěn makromolekul, cytolýza, hubení buněk ostatních organismů, defenzivní reakce na bakterie.

LS5OCV A Rare Lysozyme Crystal Form Solved Using High-Redundancy 3D Electron Diffraction Data from Micron-Sized Needle Shaped Crystals, pochází z organismu Gallus Gallus, dále je klasifikován jako hydroláza, vzorek má 258 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: metabolické procesy, katabolické procesy buněčných stěn makromolekul, hubení buněk ostatních organismů, defenzivní reakce na bakterie.

LS6BRE Hen Egg-White Lysozyme cocrystallized with Cadmium sulphate using CuKalpha source, pochází z organismu Gallus Gallus, dále je klasifikován jako hydroláza, vzorek má 129 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: metabolické procesy, katabolické procesy buněčných stěn makromolekul, hubení buněk ostatních organismů, defenzivní reakce na bakterie.

LS6APM Hen egg-white lysozyme (WT), solved with serial millisecond crystallography using synchrotron radiation, pochází z organismu Gallus Gallus, dále je klasifikován jako hydroláza, vzorek má 129 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: metabolické procesy, katabolické procesy buněčných stěn makromolekul, hubení buněk ostatních organismů, defenzivní reakce na bakterie.

LS5VAS Pekin duck egg lysozyme isoform III (DEL-III), orthorhombic form, pochází z organismu Anas platyrhynchos, dále je klasifikován jako hydroláza, vzorek má 258 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: cytolýza, hubení buněk ostatních organismů, defenzivní reakce na bakterie.

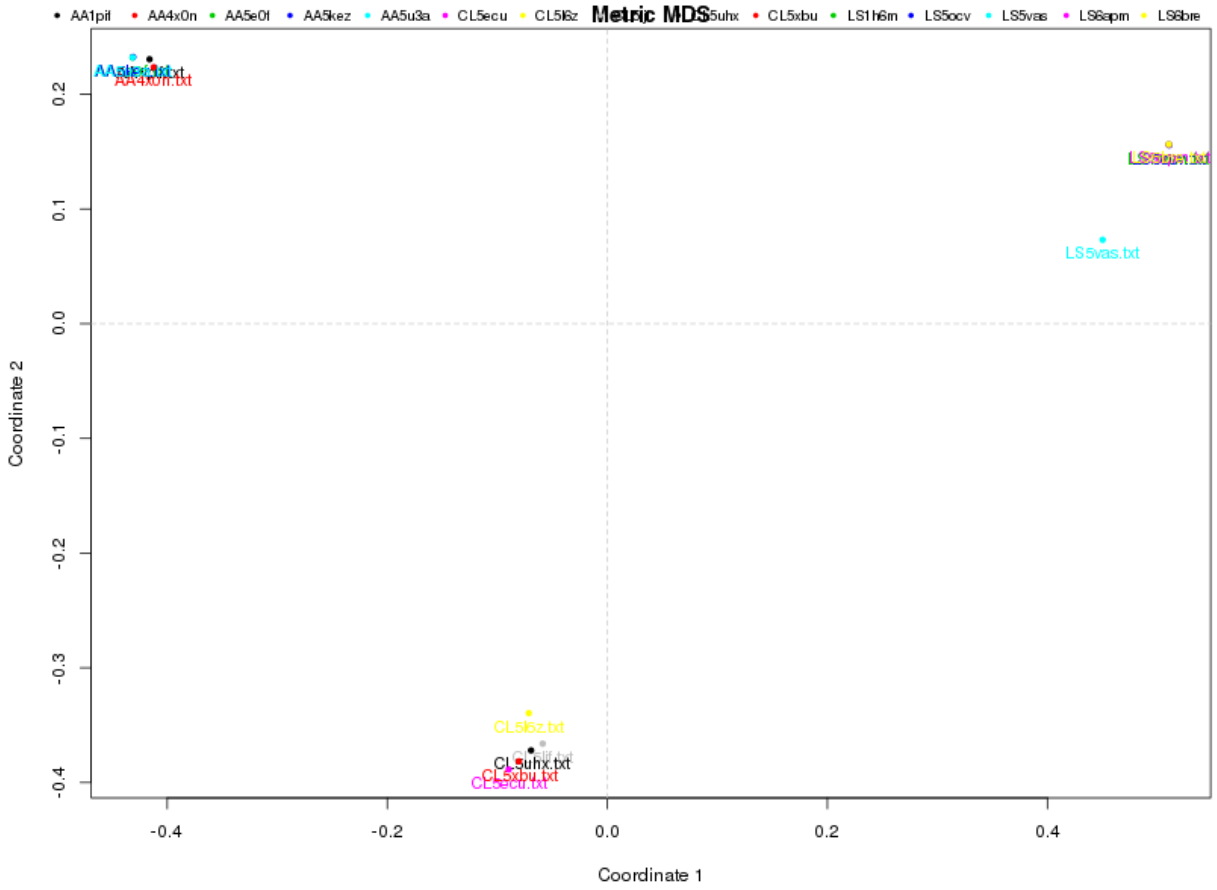


Obr. 2 Dendrogram (hierarchické shlukování)

Z výše uvedeného dendrogramu vytvořeného pomocí metody hierarchického shlukování je možné vyvodit několik závěrů. Tento graf má dvě větve, přičemž na jedné z nich se vyskytují vzorky enzymu lysozym a na druhé větvi vzorky dalších analyzovaných enzymů, α-amylázy a celulázy. Z toho lze usuzovat, že lysozymy mají mezi zkoumanými enzymy specifickou funkci, neboť jejich katalýza zajišťuje zároveň ochranu i obranu. Mezi lysozymy je pak výrazný vzorek LS5VAS. To je možné vysvětlit například tím, že jako jediný přísluší k jinému živočišnému druhu, než ostatní vzorky lysozymu. Enzym LS5VAS se účastní cytolýzy, stejně jako vzorek LS1H6M, který se nachází v jeho největší blízkosti. Ostatní použité vzorky lysozymu se účastní podobných biologických procesů, hubí buňky cizorodých organismů, defenzivně reagují na bakterie apod. Vzorky, které se nachází nejblíže další větvi, potažmo skupině enzymů, jsou LSAPM a LS6BRE. Oba tyto proteiny se jako jediné dva lysozymy účastní katabolických biologických procesů stejně jako α-

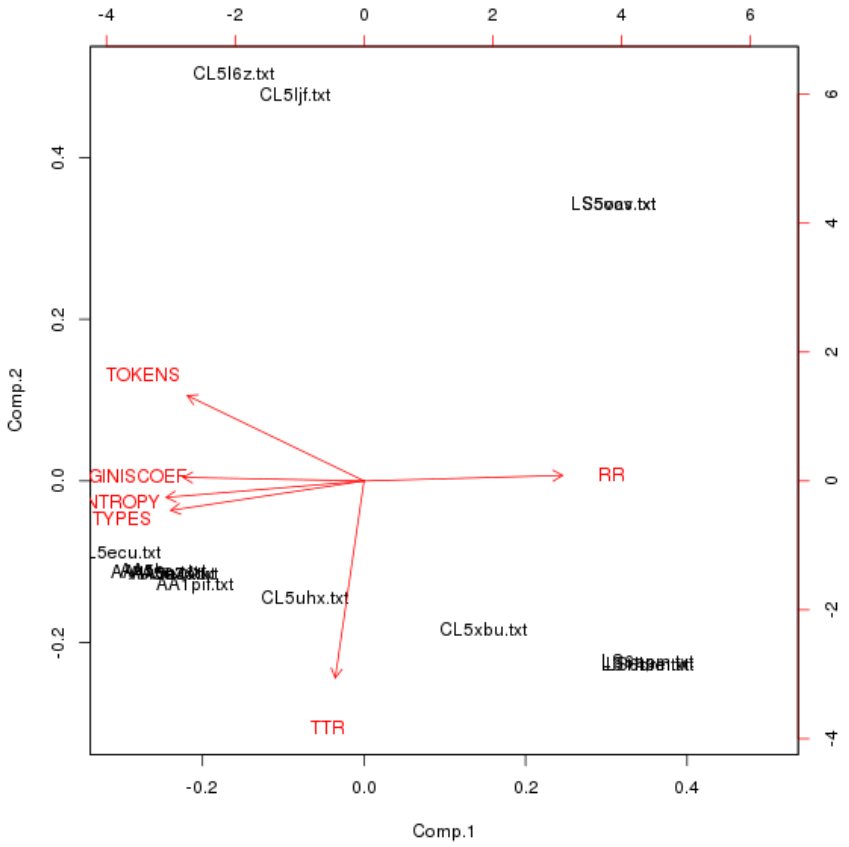
amylázy, v čemž je možné spatřovat podobnost mezi katalytickými funkcemi těchto odlišných enzymů.

Skupina vzorků α-amylázy se v grafu nachází na společné větvi se vzorky celulázy. Obě tyto skupiny vzorků se účastní podobných biologických procesů, zejména metabolických a katabolických. α-amylázy se v grafu dále dělí na dvě skupiny. Graf shlukuje skupinu tří vzorků, které se jako jediné ze skupiny účastní trávení polysacharidů, a skupinu dalších dvou vzorků, které zasahují do stejných metabolických a katabolických procesů. To lze interpretovat tak, že tento graf je schopen reflektovat podobnost funkcí těchto enzymů. Toto tvrzení lze podložit stejným případem, kdy se na společné větvi se vyskytují vzorky CL5UHX a CL5ECU, které se oba účastní metabolických procesů karbohydrátů.



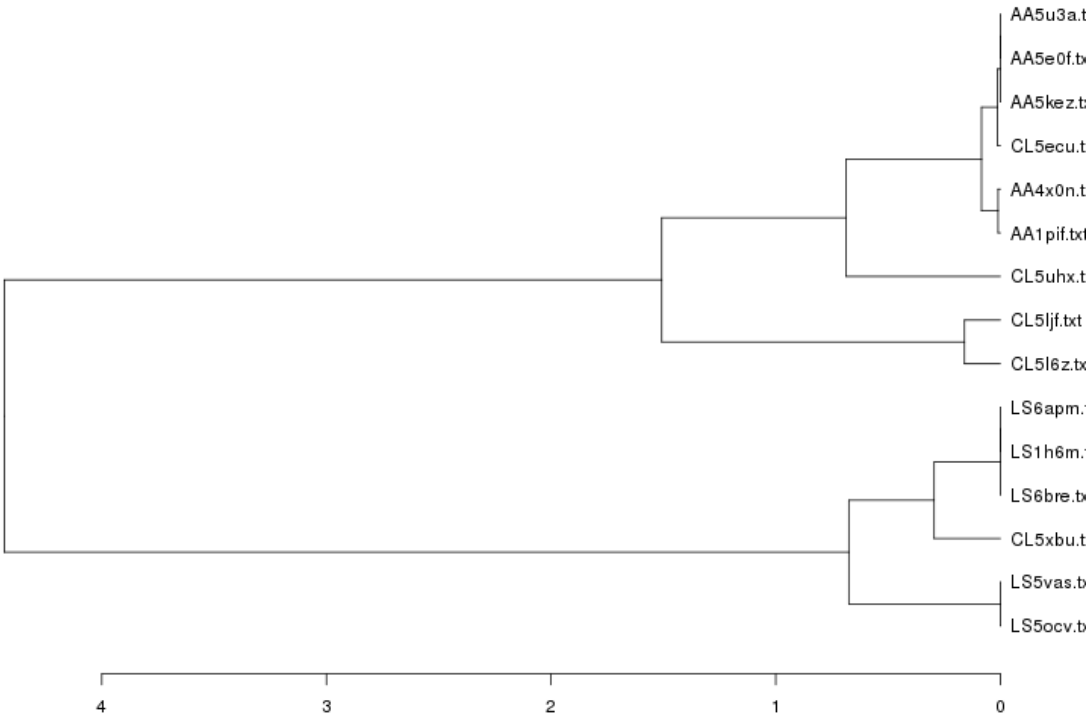
Obr. 3 MDS (vícerozměrové škálování)

Graf vytvořený pomocí metody vícerozměrového škálování podporuje interpretaci výše uvedeného grafu. Vzorky enzymů se v tomto případě shlukují do tří samostatných skupin, což odpovídá rozmístění vzorků na větvích předchozího grafu.



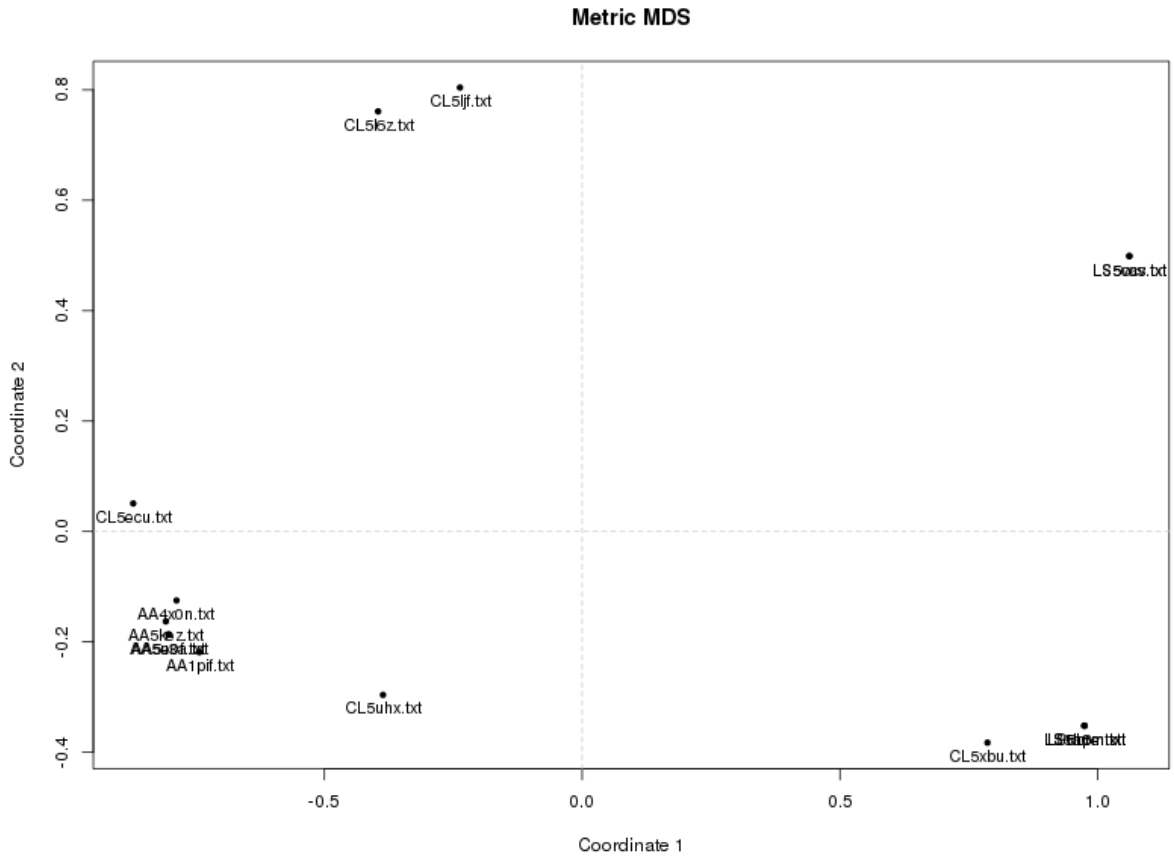
Obr. 4 PCA (analýza hlavních komponent)

Graf vytvořený pomocí metody Analýzy hlavních komponent (PCA) využívá zvolených indexů, které byly popsány výše v této kapitole, a na jejich základě zobrazuje odlišnosti mezi vzorky enzymů. Z grafu je patrné, že vzorek celulázy CL5XBU se nachází v blízkosti vzorků lysozymů, a to na základě jejich podobné hodnoty RR a TTR. Protein CL5UHX se vyskytuje v blízkosti α-amyláz vzhledem k podobným hodnotám typů. V grafu jsou rovněž v samostatné skupině vzorky CL5L6Z a CL5LJF, které se odlišují od ostatních vzorků stejné skupiny i dalších analyzovaných enzymů zejména počty tokenů a vyskytují se vedle sebe i v ostatních grafech.



Obr. 5 Dendrogram (h ierarchické shlukování)

Tento graf zobrazuje hierarchické shlukování vzorků enzymů na základě vybraných indexů. Shluky v tomto grafu jsou, až na malé výjimky pozorované u vzorků celulázy, podobné, jako tomu bylo u grafu hierarchického klastrování v modelu Bag of Words. Lysozymy i α-amylázy se nacházejí na zvláštních větvích. V tomto případě do těchto shluků pronikají proteiny celulázy, konkrétně vzorek CL5BUX se nachází na společné větvi s lysozymy, a vzorek CL5EUC se vyskytuje mezi α-amylázami, jak bylo ukázáno na předchozím grafu Analýzy hlavních komponent.



Obr. 6 MDS (vícerozměrové škálování)

V grafu vícerozměrového škálování na základě použitých indexů můžeme pozorovat několik shluků a pak proteiny, které se nacházejí mimo skupiny. Jak můžeme vidět, vzorky lysozymů i α-amyláz se nacházejí ve shlucích, z čehož je patrné, že v rámci obou skupin těchto enzymů mají vzorky mezi sebou podobnosti. Proteiny nacházející se mimo hlavní shluky jsou vzorky celulázy, které se vyskytují v různých částech grafu. Stejně jako u grafu hierarchického shlukování je viditelné, že vzorky celulázy zasahují jak do skupiny α-amyláz, tak mezi lysozymy.

4.1.2 Vzorky strukturních proteinů a jejich popis na základě modelu BoW a vybraných indexů

Pro analýzu strukturních proteinů využívá tato práce deseti vzorků, které jsou dále řazeny do skupin podle vymezených podskupin, které slouží jako příklady zástupců proteinů s výše zmíněnou funkcí. Jednotlivé dále vydělené podskupiny jsou proteiny kolagen a skupina dalších strukturních proteinů.

Mezi kolagen jsou zařazeny následující vzorky:

CG1BKV Collagen, databáze RCS PDB neuvádí, z jakého organismu vzorek pochází, dále je klasifikován jako strukturální protein, vzorek má 90 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, databáze RCS PDB neuvádí, jakých biologických procesů se vzorek účastní.

CG1CGD Hydration Structure of a Collagen Peptide, databáze RCS PDB neuvádí, z jakého organismu vzorek pochází, dále je klasifikován jako kolagen, vzorek má 90 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, databáze RCS PDB neuvádí, jakých biologických procesů se vzorek účastní.

CG1G9W Structural Basis of Collagen Stabilization Induced by Proline Hydroxylation, databáze RCS PDB neuvádí, z jakého organismu vzorek pochází, dále je klasifikován jako strukturní protein, vzorek má 21 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, databáze RCS PDB neuvádí, jakých biologických procesů se vzorek účastní.

CG1GR3 Structure of the human collagen X NC1 trimer, pochází z organismu Homo sapiens, dále je klasifikován jako kolagen, vzorek má 160 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, databáze RCS PDB neuvádí, jakých biologických procesů se vzorek účastní.

CG2CUO Collagen model peptide (PRO-PRO-GLY)9, pochází z organismu Saimiriine herpesvirus, dále je klasifikován jako strukturní protein, vzorek má 162 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, databáze RCS PDB neuvádí, jakých biologických procesů se vzorek účastní.

Mezi strukturní proteiny jsou zařazeny následující vzorky:

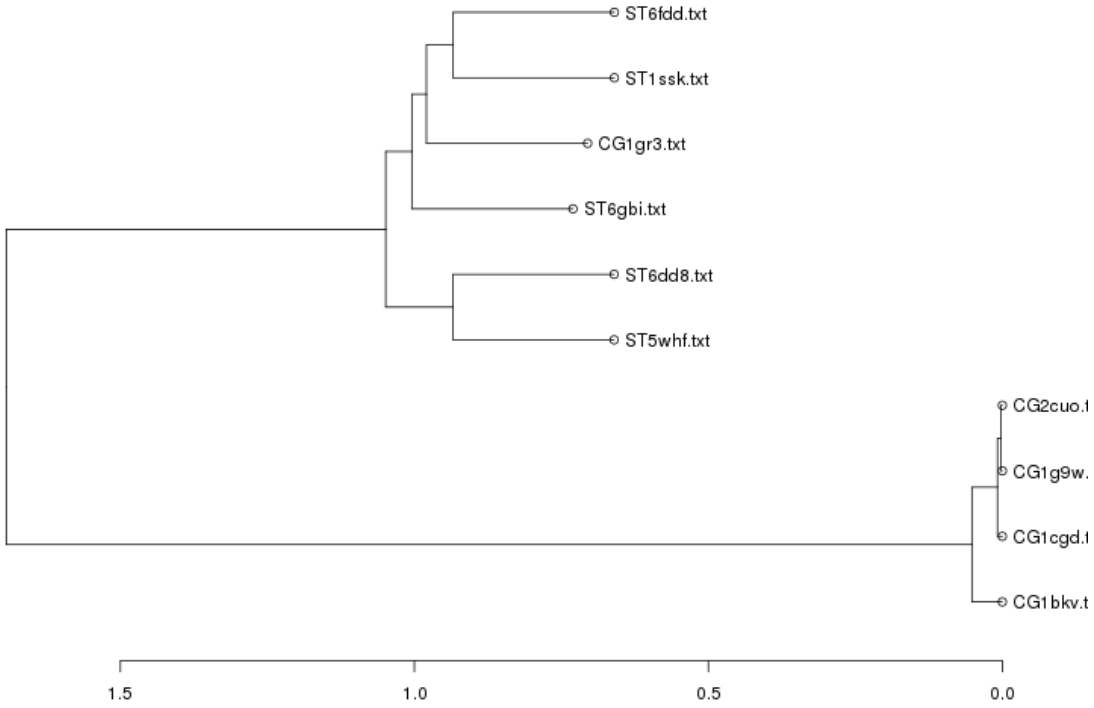
ST1SSK Structure of the N-terminal RNA-binding Domain of the SARS CoV Nucleocapsid Protein, pochází z organismu Human SARS coronavirus, dále je klasifikován jako strukturní protein, vzorek má 158 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, vzorek se neúčastní žádných biologických procesů.

ST6GBI Wnt signalling, pochází z organismu Homo sapiens, dále je klasifikován jako strukturní protein, vzorek má 140 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, databáze RCS PDB neuvádí, jakých biologických procesů se vzorek účastní.

ST6FDD Crystal Structure of the HHD2 Domain of Whirlin, pochází z organismu Mus mutulus, dále je klasifikován jako strukturní protein, vzorek má 510 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, databáze RCS PDB neuvádí, jakých biologických procesů se vzorek účastní.

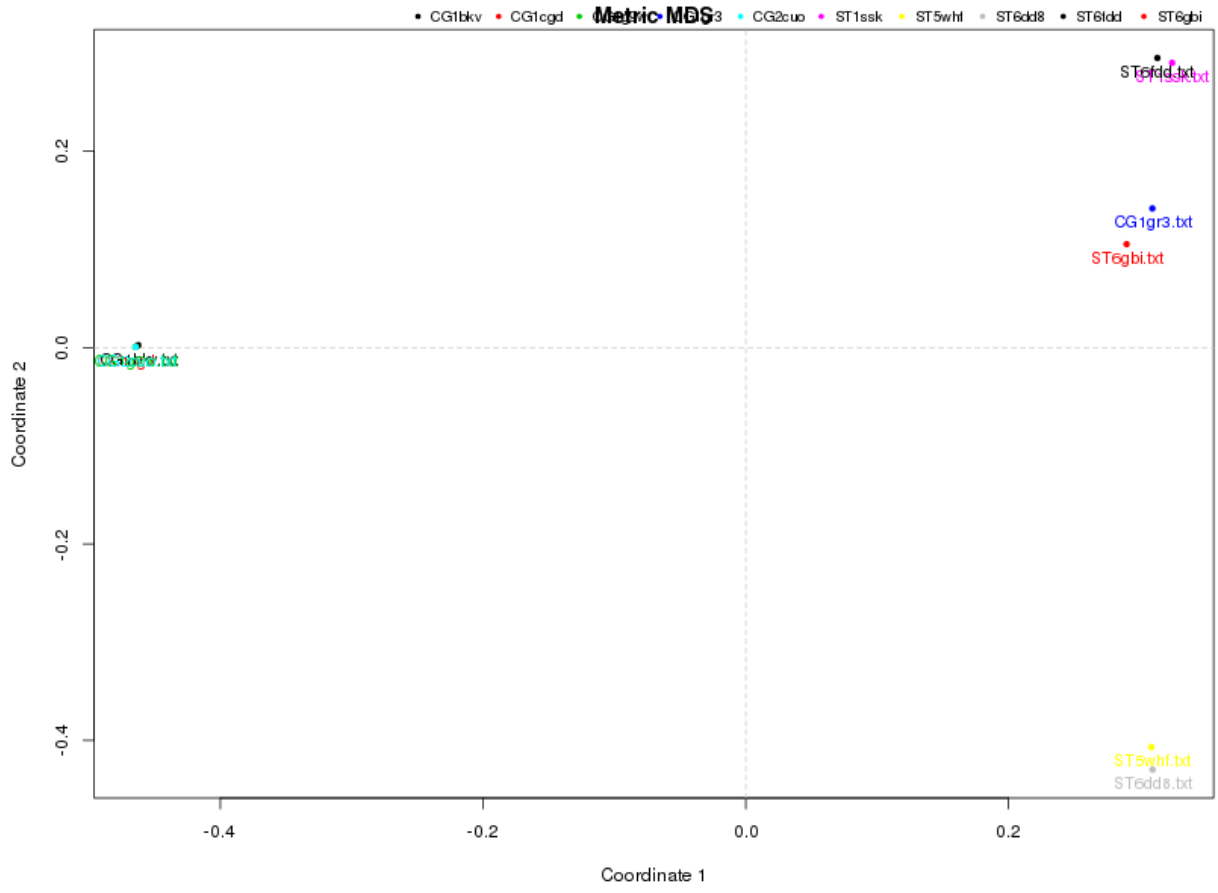
ST6DD8 Structure of mouse SYCP3, P21 form, pochází z organismu Mus mutulus, dále je klasifikován jako strukturní protein, vzorek má 576 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, databáze RCS PDB neuvádí, jakých biologických procesů se vzorek účastní.

ST5WHF Crystal structure of vimentin coil 1B packed in a high-order filamentous form, pochází z organismu Homo sapiens, dále je klasifikován jako strukturní protein, vzorek má 720 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, databáze RCS PDB neuvádí, jakých biologických procesů se vzorek účastní.



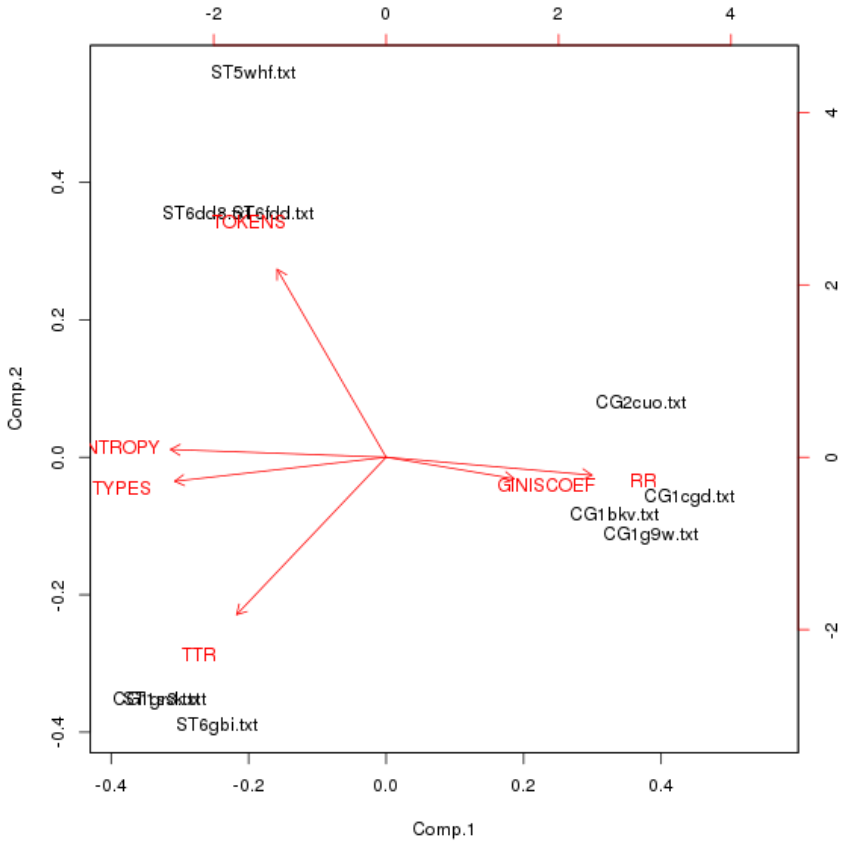
Obr. 7 Dendrogram (hierarchické shlukování)

Výše uvedený dendrogram je možné interpretovat následujícím způsobem. Strukturní proteiny jsou v grafu umístěny ve dvou větvích, na jedné z nich se vyskytují čtyři z pěti vzorků kolagenu, na druhé pak zbývající kolagen CG1GR3, v čemž můžeme spatřovat výraznou podobnost mezi proteiny se strukturní funkcí. Zároveň jsou analyzované vzorky kolagenu dostatečně specifické na to, aby většina z nich tvořila samostatnou větev grafu.



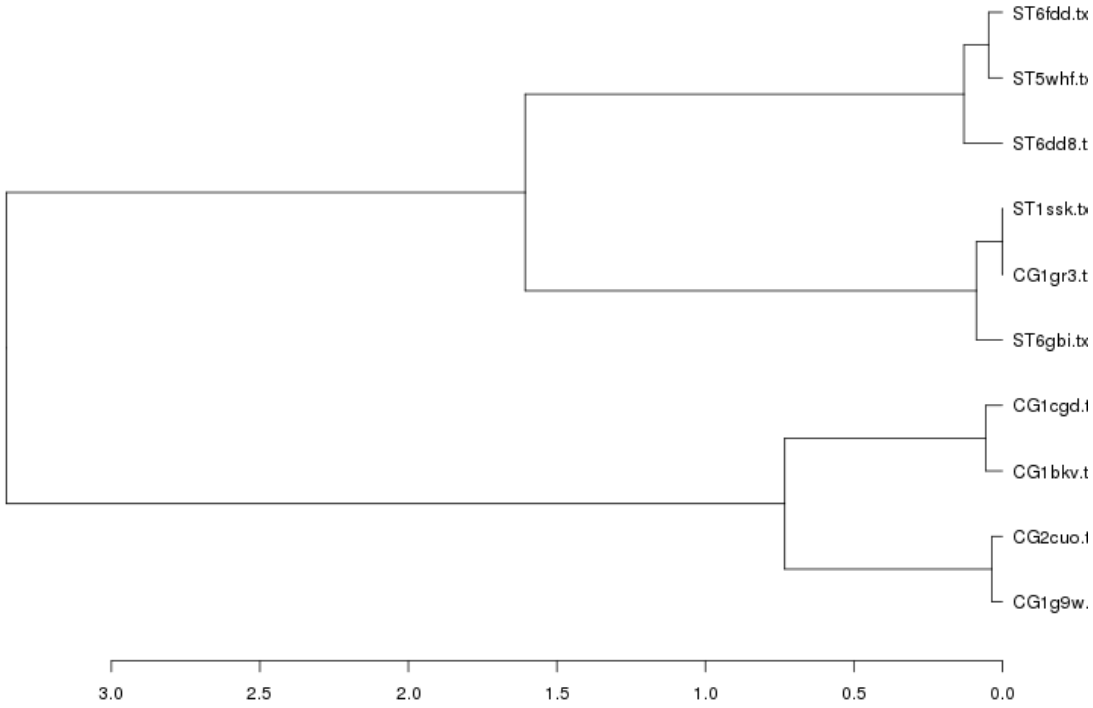
Obr. 8 MDS (vícerozměrové škálování)

Graf víceúrovňového škálování zobrazuje tři shluky. V jednom z nich můžeme pozorovat vzorky kolagenu, což podporuje interpretaci předchozího grafu, tedy že kolagen je v rámci skupiny strukturních proteinů specifický. V dalším shluku se nachází vzorky strukturních proteinů a jednoho ze vzorků kolagenu, jak tomu bylo u předchozího grafu. Třetí shluk pak tvoří vzorky ST5WHF a ST6DD8, z čehož lze usuzovat, že mezi nimi existuje výrazná podobnost. Toto tvrzení podkládá také blízkost těchto vzorků v grafu hierarchického shlukování, kde se nachází na společné větvi v bezprostřední blízkosti.



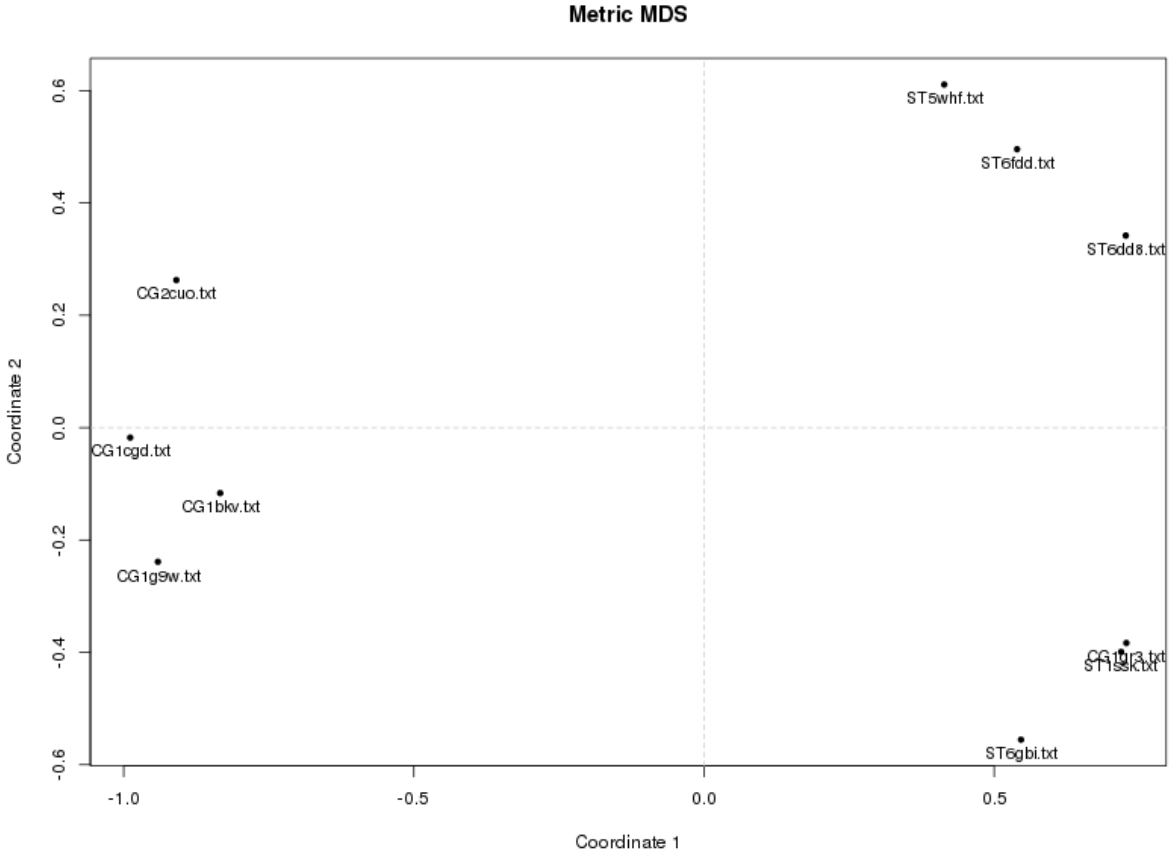
Obr. 9 PCA (analýza hlavních komponent)

V tomto grafu analýzy hlavních komponent je rovněž možné identifikovat několik shluků vytvořených na základě podobnosti hodnot zvolených indexů. Čtyři z pěti vzorků kolagenu se znovu vyskytují blízko sebe, stejně jako vzorky strukturních proteinů. I v tomto grafu je možné spatřovat odlišnost kolagenu CG1GR3 od ostatních vzorků z této skupiny. Jeho podobnost s ostatními strukturními proteiny lze vysvětlit podobnými hodnotami TTR.



Obr. 10 Dendrogram (hierarchické shlukování)

Na základě zvolených indexů byl vytvořen graf hierarchického shlukování. Z tohoto grafu je rovněž patrná odlišnost jednoho ze vzorků kolagenu, který se nachází ve shluku dalších strukturních proteinů. Jedná se opět o vzorek s kódovým označením CG1GR3, což podporuje interpretaci předchozího grafu, tedy jeho podobnost s ostatními strukturními proteiny na základě podobných hodnot indexu TTR.



Obr. 11 MDS (vícerozměrové škálování)

Z grafu víceúrovňového škálování je opět patrná odlišnost vzorku kolagenu CG1GR3, který se vyskytuje v blízkosti strukturních proteinů ST6GBI a ST1SSK, což zdůrazňuje jejich výše popsanou podobnost. Zbylé vzorky kolagenu se v grafu nacházejí ve vzájemné blízkosti, což podporuje interpretaci předchozích grafů.

4.1.3 Vzorky transportních proteinů a jejich popis na základě modelu BoW a vybraných indexů

Pro analýzu transportních proteinů využívá tato práce patnácti vzorků, které jsou dále řazeny do skupin po pěti podle vymezených podskupin, které slouží jako příklady zástupců proteinů s transportní funkcí. Jednotlivé dále vydělené podskupiny jsou proteiny hemoglobiny (vzorky označené zkratkou HG), lipoproteiny (vzorky označené zkratkou LP) a skupina dalších proteinů s transportní funkcí (vzorky označené zkratkou TS). Hemoglobiny jsou globulární heme proteiny v červených krvinkách obratlovců a v plazmě mnoha bezobratlých. Hemoglobin je nositelem a transportérem kyslíku a karbon dioxidu. Heme skupina na sebe váže kyslík a karbon dioxid a způsobuje, že krev je červená. (Jain

2005: 1120) Lipoproteiny jsou konjugované proteiny, které přenášejí vodou nerozpustné lipidy v krvi. Samotný proteinový komponent se nazývá apolipoprotein. (Jain 2005: 1127)

Mezi hemoglobiny jsou zařazeny následující vzorky:

HG1A3N Deoxy Human Hemoglobin, dále je klasifikován jako transportér kyslíku, pochází z organismu Homo sapiens, vzorek má 574 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: endocytóza zprostředkovaná receptory, pozitivní regulace buněčné smrti, transport kyslíku, transporty bikarbonátů, reakce na peroxid vodíku, katabolický proces peroxidu vodíku, heterooligomerizace proteinu, detoxikace buněčného oxidu.

HG1IT2 Hagfish deoxy hemoglobin, je klasifikován jako ukladatel a transportér kyslíku, pochází z organismu Eptatretus burgeri, vzorek má 292 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: transport kyslíku.

HG1OUT Trout Hemoglobin I, dále je klasifikován jako transportér kyslíku, pochází z organismu Oncorhynchus mykiss, vzorek má 289 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: transport kyslíku.

HG2R80 Pigeon Hemoglobin (OXY form), dále je klasifikován jako transportér kyslíku, pochází z organismu Columba livia, vzorek má 574 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: transport kyslíku.

HG6BB5 Human Oxy-Hemoglobin, dále je klasifikován jako transportér kyslíku, pochází z organismu Homo sapiens, vzorek má 284 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: endocytóza zprostředkovaná receptory, pozitivní regulace buněčné smrti, transport kyslíku, transporty bikarbonátů, reakce na peroxid vodíku, katabolický proces peroxidu vodíku, heterooligomerizace proteinu, detoxikace buněčného oxidu.

Mezi lipoproteiny jsou zařazeny následující vzorky:

LP3HOE Crystal Structure of Surface Lipoprotein, dále je klasifikován jako transportní protein, pochází z organismu Actinobacillus pleuropneumoniae, vzorek má 510

aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, v anotaci RCSB PDB nejsou uvedeny biologické procesy, kterých se vzorek účastní.

LP3KSN Crystal structure of the lipoprotein localization factor, LolA, dále je klasifikován jako transportní protein, pochází z organismu Escherichia coli, vzorek má 183 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: transport proteinů, transport lipoproteinů, lokalizace lipoproteinů ve vnější membráně.

LP2ZPC Crystal structure of the R43L mutant of LolA in the closed form, dále je klasifikován jako transportní protein, pochází z organismu Escherichia coli, vzorek má 190 aminokyselinových reziduí, v sekvenci se nachází 1 mutace, se následujících biologických procesů: transport proteinů, transport lipoproteinů, lokalizace lipoproteinů ve vnější membráně.

LP2W7Q Structure of Pseudomonas aeruginosa LolA, dále je klasifikován jako transportní protein, pochází z organismu Pseudomonas aeruginosa, vzorek má 408 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: transport proteinů, transport lipoproteinů, lokalizace lipoproteinů ve vnější membráně.

LP1SOH The structure of human apolipoprotein C-II in dodecyl phosphocholine, je klasifikován jako transportní protein, pochází z organismu Homo sapiens, vzorek má 79 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: metabolické procesy retinolů, metabolické procesy, metabolické procesy lipidů, transport lipidů.

Mezi ostatní transportní proteiny jsou zařazeny následující vzorky:

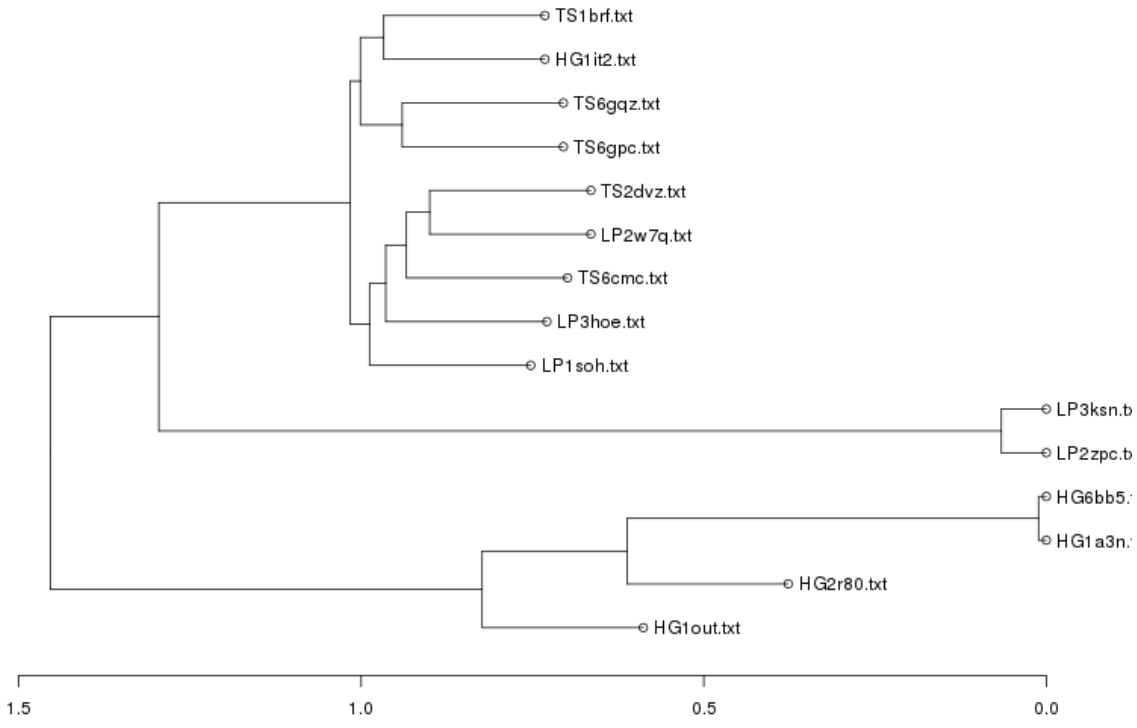
TS1BRF Rubredoxin (Wild Type) from Pyrococcus Furiosus, dále je klasifikován jako transportér elektronů, pochází z organismu Pyrococcus furiosus, má 53 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: transport elektronů, proces redukce oxidace.

TS2DVZ Structure of a periplasmic transporter, dále je klasifikován jako transportní protein, pochází z organismu Bordetella pertusis, má 314 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, neúčastní se žádného biologického procesu.

TSP6GPC Crystal structure of the arginine-bound form of domain 1 from TmArgBP, dále je klasifikován jako transportní protein, pochází z organismu Thermotoga maritima, má 252 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, v anotaci RCSB PDB nejsou uvedeny biologické procesy, kterých se vzorek účastní.

TS6GQZ Petrobactin-binding engineered lipocalin without ligand, dále je klasifikován jako transportní protein, pochází z organismu Homo sapiens, má 348 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, v anotaci RCSB PDB nejsou uvedeny biologické procesy, kterých se vzorek účastní.

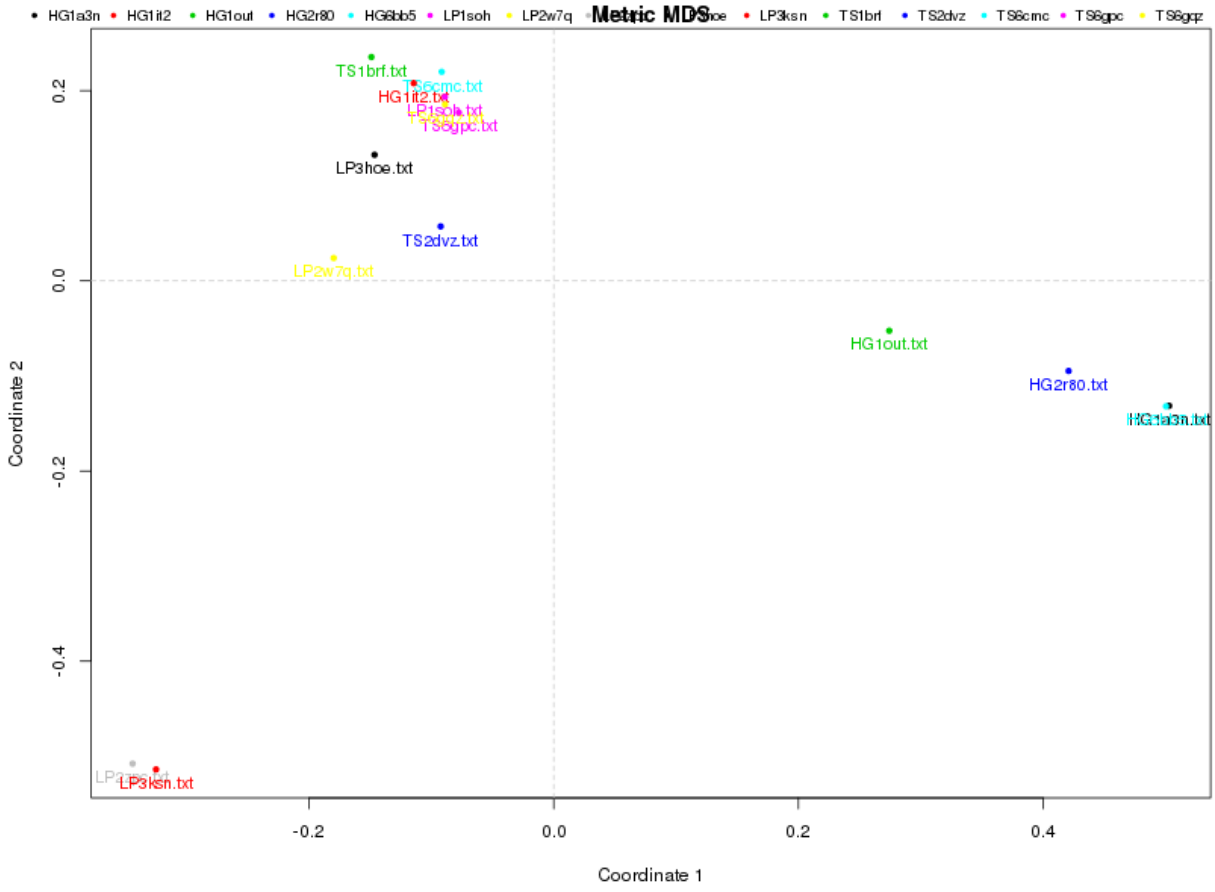
TS6CMC Barium sites in the structure of a desensitized acid sensing ion channel, dále je klasifikován jako transportní protein, pochází z organismu Gallus gallus, má 450 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, v anotaci RCSB PDB nejsou uvedeny biologické procesy, kterých se vzorek účastní.



Obr. 12 Dendrogram (hierarchické shlukování)

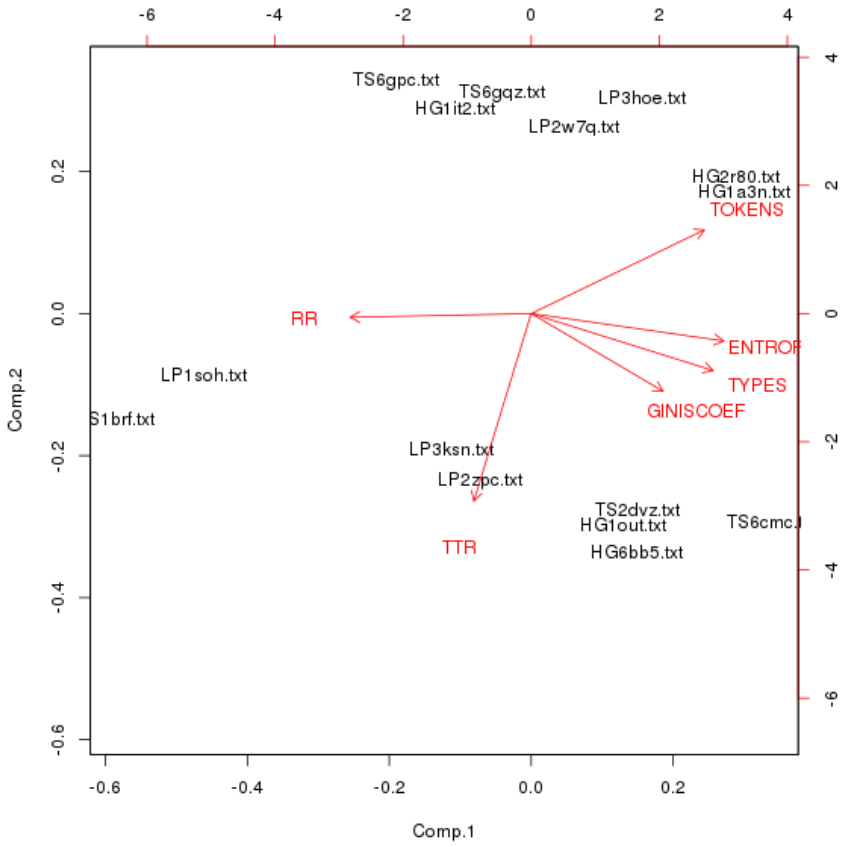
V grafu zobrazujícím shlukovou analýzu transportních proteinů můžeme pozorovat několik větví. Na jedné z nich se samostatně vyskytují pouze vzorky hemoglobinu, z čehož lze usoudit, že hemoglobin je mezi transportními proteiny výrazně specifický. Tuto skutečnost lze vysvětlit jeho ojedinělou funkcí vázat na sebe a přenášet právě kyslík, zatímco ostatní analyzované transportní proteiny přenášejí jiné proteiny a lipidy (lipoproteiny), elektrony (vzorek TS1BRF) apod. Jediný vzorek hemoglobinu, který se nachází na jiné větvi než zbytek jeho skupiny, HG1IT2, se vyskytuje v blízkém shluku právě transportního proteinu TS1BRF. Je nutné zmínit, že vzorek hemoglobinu HG1IT2 vykazuje anomálie i v ostatních níže uvedených grafech. Výrazná odlišnost hemoglobinů může být způsobena právě jejich schopností nabýt ojedinělou kvartérní strukturu.

V blízkosti hemoglobinu se na samostatné větvi nachází dvojice lipoproteinů, vzorky LP3KSN a LP2ZPC. Tyto lipoproteiny se oba účastní stejných biologických procesů, stejně jako jejich druhá dvojice, která se vyskytuje v jednom shluku, vzorky LP1SOH a LP3HOE. Zbylé transportní proteiny se nachází na další větvi ve skupině se zmíněným samostatným vzorkem hemoglobinu. Skupiny transportních proteinů se účastní přenosu elektronů a iontů, graf tedy reflektuje jejich podobnost mezi sebou a odlišnost od ostatních vzorků.



Obr. 13 MDS (vícerozměrové škálování)

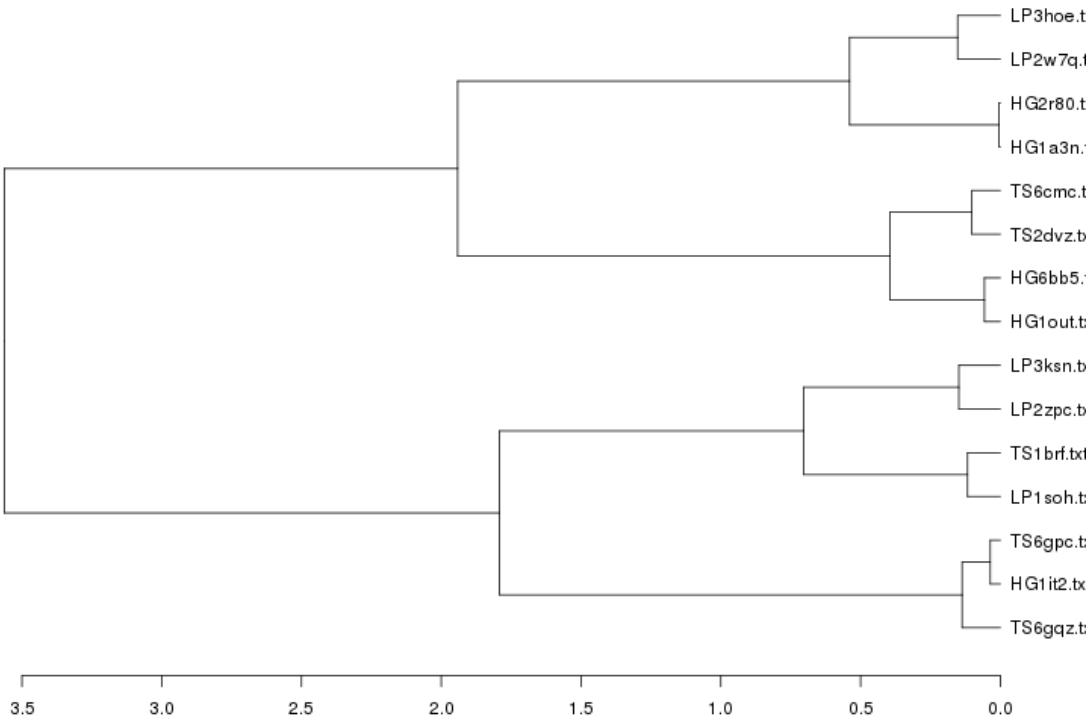
Tento graf víceúrovňového škálování zobrazuje tři shluky analyzovaných vzorků, které korespondují s třemi větvemi vytvořenými hierarchickým shlukováním. Opět můžeme vidět shluk vzorků hemoglobinu, což podporuje předchozí interpretaci o specifičnosti tohoto proteinu. Stejně jako v předchozím grafu se jeden ze vzorků hemoglobinu, protein HG1IT2, nachází v bezprostřední blízkosti transportního proteinu TS1BRF a dalších vzorků transportních proteinů, které v předchozím grafu tvořily samostatnou větev. Posledním shlukem je dvojice lipoproteinů LP3KSN a LP2ZPC, které se účastní stejných biologických procesů.



Obr. 14 PCA (analýza hlavních komponent)

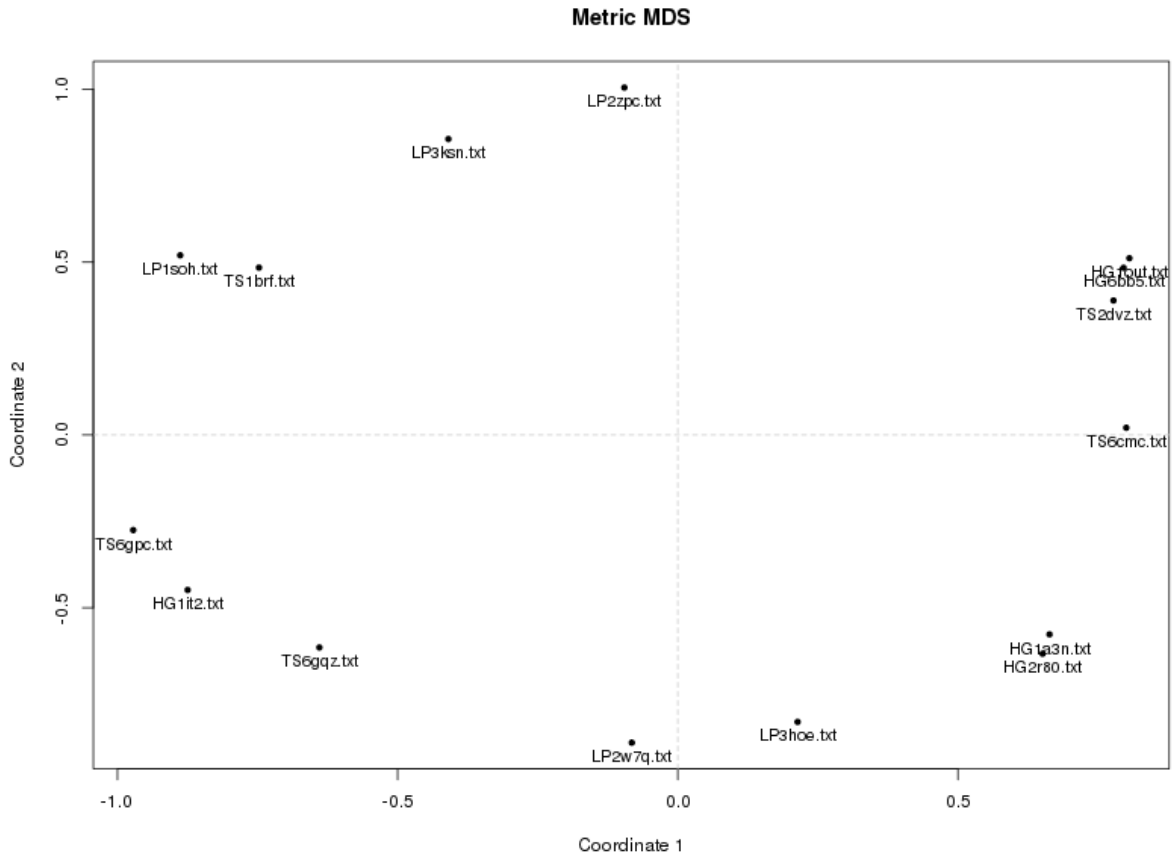
Pomocí grafu zobrazujícího analýzu hlavních komponent můžeme interpretovat podobnosti a odlišnosti vzorků na základě zvolených indexů. Z tohoto grafu je patrná například podobnost hodnot TTR u dvojice lipoproteinů LP3KSN a LP2ZPC, které se i v předchozích grafech vyskytovaly v bezprostřední blízkosti. Dále můžeme vidět podobnost

dvojice proteinů vzorků hemoglobinu, HG2R80 a HG1A3N, a to na základě podobné hodnoty tokenů. Rovněž si můžeme povšimnout shluku vzorků, které se vyskytují v horní části grafu. Tato skupina se ve výše uvedeném grafu hierarchického škálování vyskytovala na jedné větvi, což lze nyní interpretovat pomocí jejich podobných hodnot tokenů a indexu RR.



Obr. 15 Dendrogram (hierarchické shlukování)

Graf hierarchického shlukování zobrazuje uspořádání vzorků transportních proteinů, které vykazuje výrazné změny oproti předchozímu grafu stejného typu (Hclust). V tomto případě se vzorky proteinů uspořádaly do dvou hlavních větví namísto tří, a dále se vydělily na dvojice podle podobných hodnot zvolených indexů. Opět můžeme vidět větev, která obsahuje skupinu transportních proteinů, vzorku hemoglobinu HG1IT2 a lipoproteinů LP3KSN a LP2ZPC, které i v předchozím grafu byly na jedné větvi. Další větev pak tvoří zbylé vzorky hemoglobinů, transportních proteinů a lipoproteinu, které tvoří dvojice. Tento graf tak podporuje interpretaci grafu analýzy hlavních komponent, ze kterého je jasně patrná jejich podobnost na základě tokenů a indexu RR.



Obr. 16 MDS (vícerozměrové škálování)

Tento graf víceúrovňového škálování podporuje vizualizaci výše uvedených grafů. I zde je patrné, že vzorky tvoří dvojice totožné s těmi, které se vyskytují na společných koncových větvích předchozího grafu.

4.1.4 Vzorky nutričních proteinů a jejich popis na základě modelu BoW a vybraných indexů

Pro analýzu nutričních proteinů využívá tato práce patnácti vzorků, které jsou dále řazeny do skupin po pěti podle vymezených podskupin, které slouží jako příklady zástupců proteinů s nutriční funkcí. Jednotlivé dále vydělené podskupiny jsou proteiny kasein (vzorky označené zkratkou KS) a dvě skupiny proteinů ukládajících živiny (storage) (vzorky označené zkratkou SG).

Mezi kaseiny jsou zařazeny následující vzorky:

KS3SV0 Crystal structure of casein kinase-1 like protein in plant, dále je klasifikován jako transferáza, pochází z organismu Oryza sativa subsp. japonica, vzorek má 483

aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: proteinová fosforylace**,** endocytóza, regulace tvaru buňky, fosforylace, peptidyl-serine fosforylace.

KS4JJR A P21 crystal form of mammalian casein kinase 1 delta, dále je klasifikován jako transferáza, pochází z organismu Mus mutulus, vzorek má 598 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: fosforylace proteinová fosforylace.

KS5IH4 Human Casein Kinase 1 isoform delta apo (kinase domain), dále je klasifikován jako transferáza, pochází z organismu Homo sapiens, má 249 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: proteinová fosforylace.

KS2CHL Structure of casein kinase 1 gamma 3, dále je klasifikován jako transferáza, pochází z organismu Homo sapiens, má 351 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: proteinová fosforylace.

KS2IZT Structure of casein kinase gamma 3 in complex with inhibitor, dále je klasifikován jako transferáza, pochází z organismu Homo sapiens, má 330 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: proteinová fosforylace.

Mezi proteiny ukládající živiny jsou zařazeny následující vzorky:

SG4FYP Crystal Structure of Plant Vegetative Storage Protein, dále je klasifikován jako rostlinný protein, pochází z organismu Arabidopsis thaliana, má 526 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, neúčastní se žádných biologických procesů.

SG4NDO Crystal structure Molybdenum Storage Protein with fully Mo-loaded cavity, dále je klasifikován jako protein vázající kovy, pochází z organismu Azotobacter vinelandii, má 546 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, neúčastní se žádných biologických procesů.

SG5ARM APO-CSP3 (Copper Storage Protein 3) from Methylosinus, dále je klasifikován jako protein vázající měď, RCSB PDB neuvádí organismus, ze kterého vzorek

pochází, má 133 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, neúčastní se žádných biologických procesů.

SG1PHS The Three-Dimensinoal Structure of the Seed Storage Protein Phaseolin at 3 Angstroms Resolution, dále je klasifikován jako rostlinný protein ukládající živiny (vicilin), pochází z organismu Phaseolus vulgaris, má 397 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, neúčastní se žádných biologických procesů.

SG1Y00 Solution structure of the Carbon Storage Regulator protein CsrA, dále je klasifikován jako protein vázající RNA, pochází z organismu Escherichia coli, má 122 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: regulace metabolického procesu karbohydrátů, MRNA katabolické procesy, regulace translace, negativní regulace glykogenového biosyntetického procesu, negativní a pozitivní regulace iniciace translace.

SG1WBA WIinged Bean Albumin 1, dále je klasifikován jako protein ukládající živiny, pochází z organismu Psophocarpus tetragonolobus, má 175 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: negativní regulace endopeptidázové aktivity.

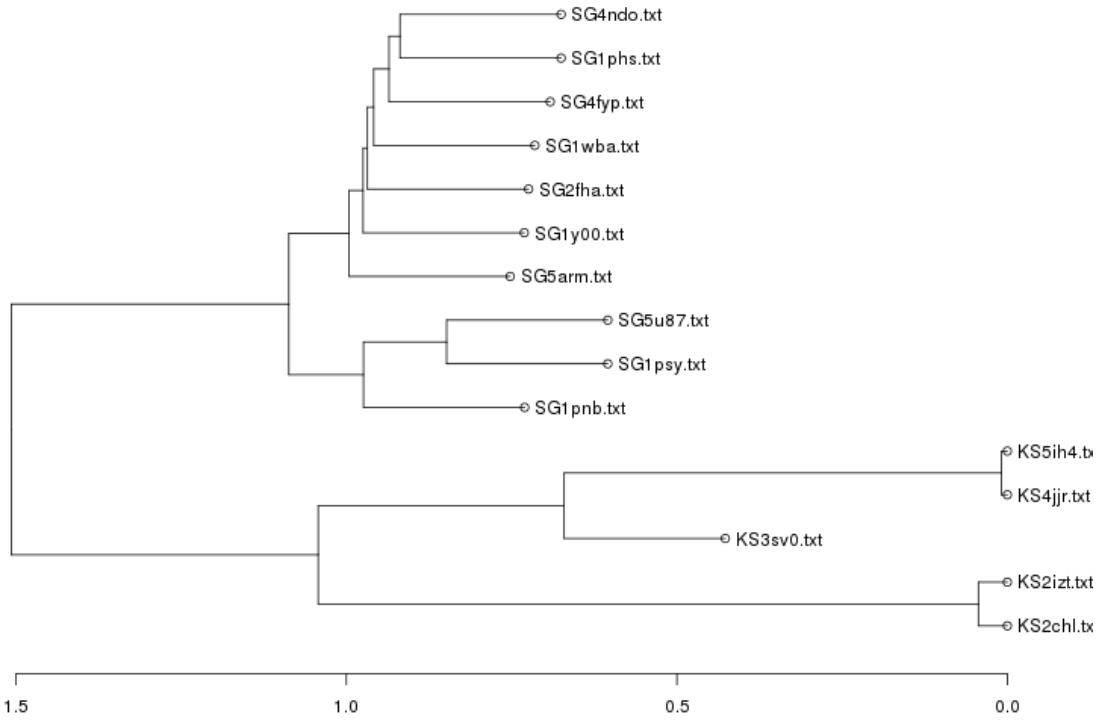
SGb2FHA Human H Chain Ferritin, dále je klasifikován jako protein ukládající železo, pochází z organismu Homo sapiens, má 183 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: transport iontů železa, buněčná homeostáza iontů železa, negativní regulace buněčné proliferace, neutrofilická degranulace, negativní regulace proliferace fibroblastů, oxidační redukční procesy.

SG1PNB Structure of Napin BNIB, NMR, 10 Structures, dále je klasifikován jako protein ukládající živiny, pochází z organismu Brassica napus, má 106 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, neúčastní se žádných biologických procesů.

SG1PSY Structure OF RicC3, NMR, 20 Structures, dále je klasifikován jako rostlinný protein, pochází z organismu Ricinus communis, má 125 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, neúčastní se žádných biologických procesů.

SG5U87 NMR structure of the precursor protein PawS1 comprising SFTI-1 and a seed storage albumin, dále je klasifikován jako rostlinný protein, pochází z organismu

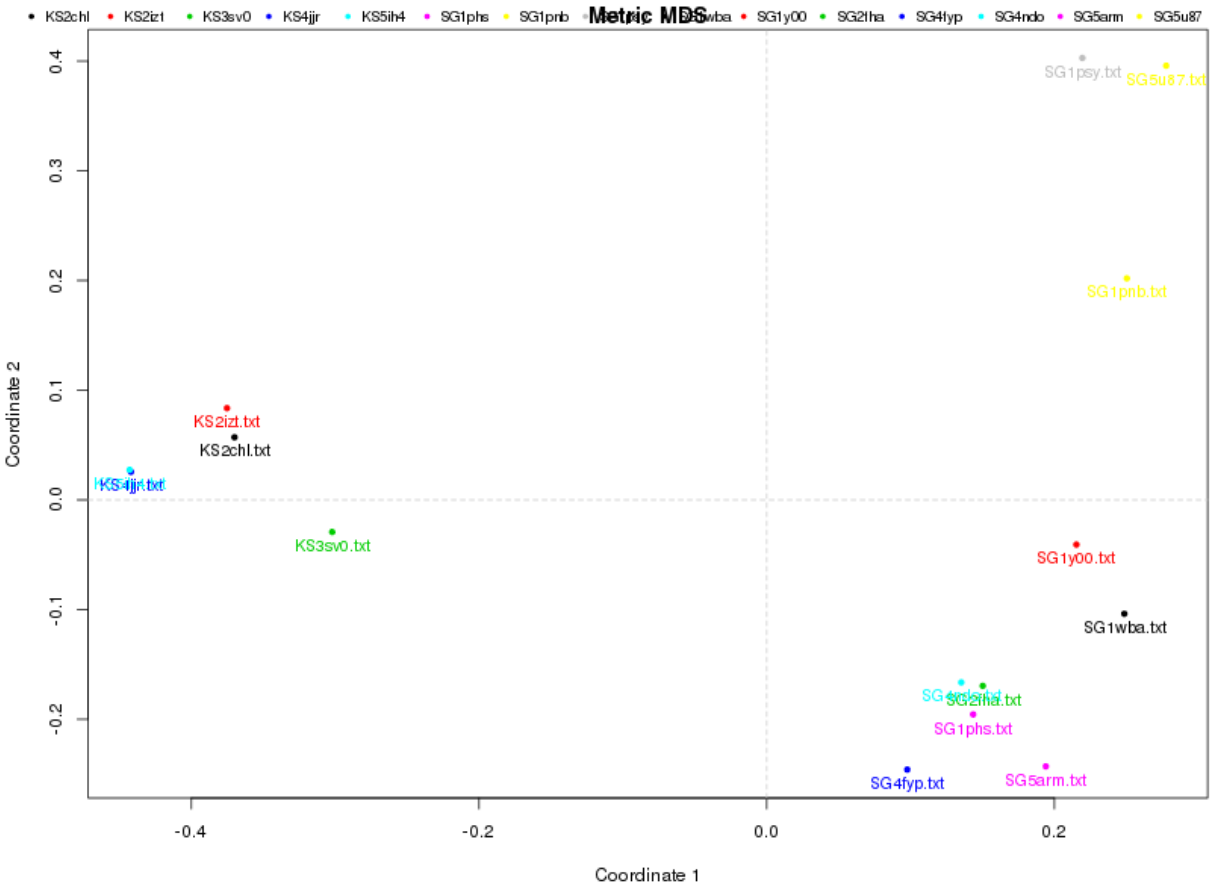
Helianthus annuus, má 116 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, neúčastní se žádných biologických procesů.



Obr. 17 Dendrogram (hierarchické shlukování)

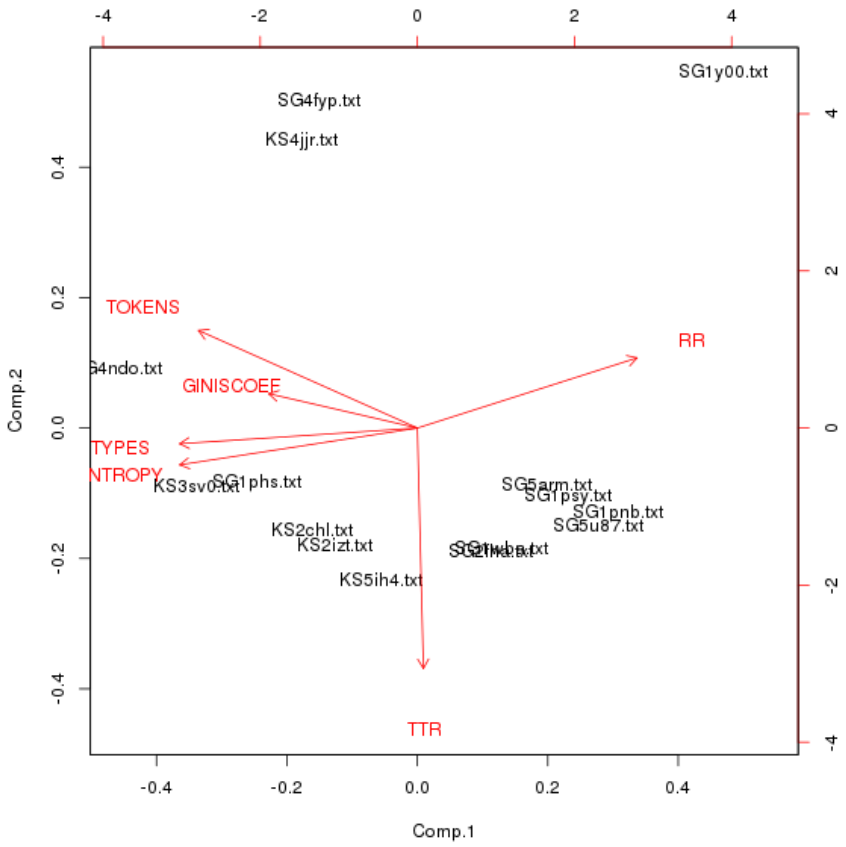
V tomto grafu hierarchického shlukování můžeme pozorovat výraznou vzájemnou podobnost mezi vzorky kaseinu, které se vyskytují na společné samostatné větvi. Čtyři z pěti vzorků kaseinu tvoří dvojice. Tyto čtyři proteiny se účastní stejného biologického procesu, v tomto případě proteinové fosforylace. Pátý vzorek kaseinu se vyskytuje uprostřed mezi oběma dvojicemi, přičemž vzorek KS3SV0 se kromě proteinové fosforylace účastní ještě dalších procesů jako je například endocytóza. Z takto vykreslené části grafu pak lze vyvodit, že zvolené vzorky proteinu kasein jsou si dostatečně podobné, protože se vyskytují na jedné větvi. Zároveň mezi nimi existují odlišnosti, které je možné vysvětlit rozdílností biologických procesů, kterých se účastní. Druhá větev grafu zobrazuje skupinu proteinů, které se účastní ukládání živin. Ve střední části grafu je pozorována trojice nutričních proteinů SG1PNB, SG1PSY a SG5U87, které patří mezi rostlinné proteiny. Podobnost další

skupiny proteinů, SG5ARM, SG100Y a SG2FHA, můžeme spatřovat v jejich schopnosti vázat na sebe kovy. Podle klasifikace vytyčené výše v práci se jedná o metaloproteiny.



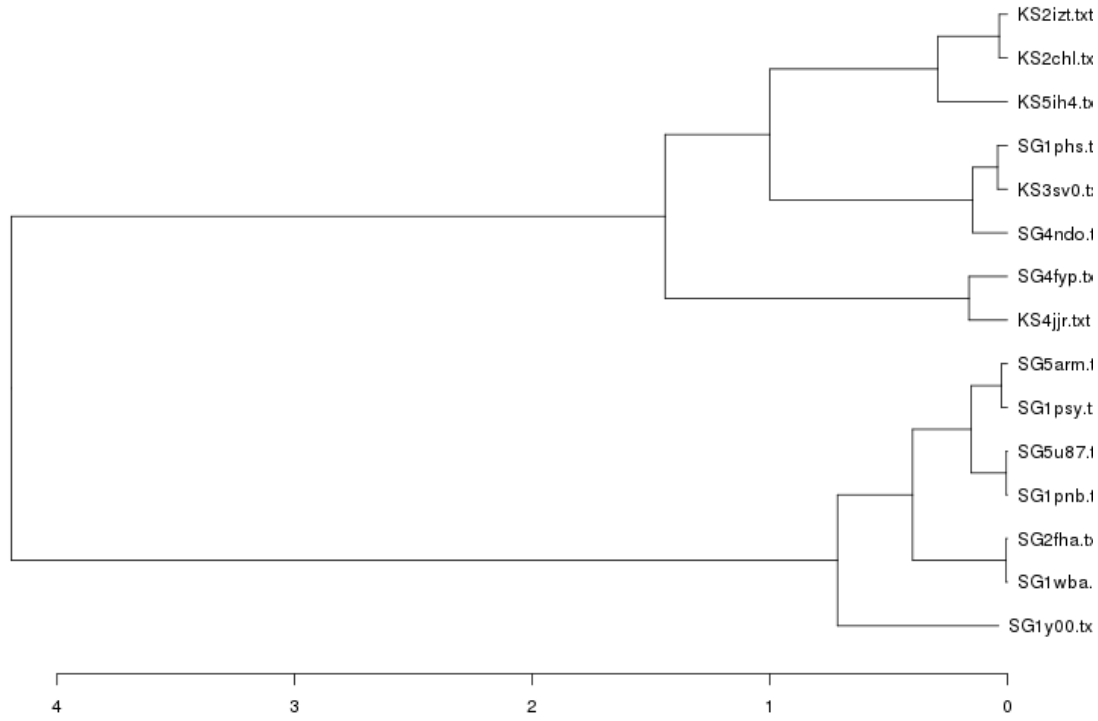
Obr. 18 MDS (vícerozměrové škálování)

I tento graf víceúrovňového škálování podporuje interpretaci předchozího grafu. V tomto případě graf rovněž zobrazuje vzorky kaseinu, které se nacházejí ve skupině. Stejně tak můžeme vidět blízkost nutričních proteinů SG1PNB, SG1PSY a SG5U87, což podporuje tvrzení o jejich podobnosti.



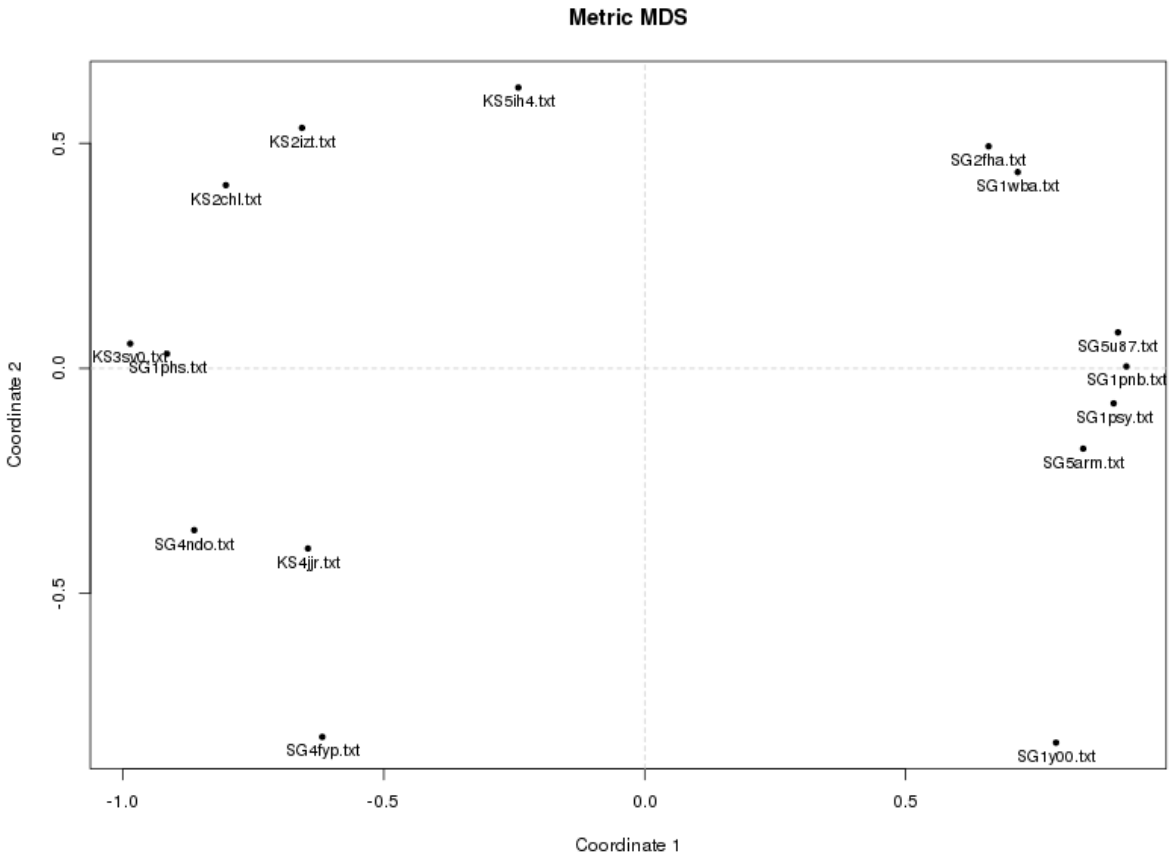
Obr. 19 PCA (analýza hlavních komponent)

Graf analýzy hlavních komponent rovněž podporuje předchozí závěry. V tomto grafu pozorujeme blízkost vzorků kaseinu a to na základě indexů entropie a TTR. Jediný ze vzorků kaseinu, který se vyskytuje mimo skupinu, je KS4JJR a to kvůli odlišné hodnotě tokenů. Další značně odlišný vzorek je nutriční protein SG1Y00, který se nachází v pravé horní části grafu a od zbylých proteinů se odlišuje hodnotou RR.



Obr. 20 Dendrogram (hierarchické shlukování)

Tento graf hierarchického shlukování se odlišuje od původního grafu tohoto typu. Oba dendrogramy mají dvě větve, přičemž vždy na jedné z nich se vyskytují všechny vzorky kaseinu. V tomto grafu však spolu s kaseinem pozorujeme i další nutriční proteiny, konkrétně SG1PHS, SG4NDO a SG4YP. Tyto vzorky tvoří s kaseiny dvojice na základě podobných hodnot tokenů, typů a indexu entropie.



Obr. 21 MDS (vícerozměrové shlukování)

Pomocí tohoto grafu víceúrovňového škálování můžeme dále vizualizovat podobnosti mezi vzorky, které byly patrné v předchozích grafech. V levé polovině grafu se vedle sebe vyskytují vzorky kaseinu, což podporuje tvrzení o jejich podobnosti. Zbylé nutriční proteiny tvoří stejné skupiny, jako tomu bylo v předchozích grafech.

4.1.5 Vzorky kontrakčních a pohybových proteinů a jejich popis na základě modelu BoW a vybraných indexů

Pro analýzu kontrakčních a pohybových proteinů využívá tato práce patnácti vzorků, které jsou dále řazeny do skupin po pěti podle vymezených podskupin, které slouží jako příklady zástupců proteinů s výše zmíněnou funkcí. Jednotlivé dále vydělené podskupiny jsou proteiny aktin, myozin a troponin.

Mezi aktiny jsou zařazeny následující vzorky:

AK1J6Z Uncomplexed Actin, dále je klasifikován jako kontrakční protein, pochází z organismu Oryctolagus cuniculus, vzorek má 375 aminokyselinových reziduí, v sekvenci

se nenachází žádná mutace, účastní se následujících biologických procesů: pozitivní regulace exprese genu, polymerizace aktinových vláken, sestava tenkých vláken kosterních svalů, vývoj kosterních svalových vláken, shromáždění sady aktinových vláken, migrace mesenchymu, pozitivní regulace aktivity ATPázy závislé na aktinu.

AK2ZWH Model for the F-actin structure, dále je klasifikován jako kontrakční protein, pochází z organismu Oryctolagus cuniculus, vzorek má 375 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: pozitivní regulace exprese genu, polymerizace aktinových vláken, sestava tenkých vláken kosterních svalů, vývoj kosterních svalových vláken, shromáždění sady aktinových vláken, migrace mesenchymu, pozitivní regulace aktivity ATPázy závislé na aktinu.

AK3HBT The structure of native G-actin, je klasifikován jako kontrakční protein, pochází z organismu Oryctolagus cuniculus, má 375 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: pozitivní regulace exprese genu, polymerizace aktinových vláken, sestava tenkých vláken kosterních svalů, vývoj kosterních svalových vláken, shromáždění sady aktinových vláken, migrace mesenchymu, pozitivní regulace aktivity ATPázy závislé na aktinu.

AK4PKG Complex of ATP-actin With the N-terminal Actin-Binding Domain of Tropomodulin, dále je klasifikován jako kontrakční protein a protein vázající aktin, pochází z organismu Homo sapiens, má 563 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: pozitivní regulace exprese genu, polymerizace aktinových vláken, sestava tenkých vláken kosterních svalů, vývoj kosterních svalových vláken, shromáždění sady aktinových vláken, migrace mesenchymu, pozitivní regulace aktivity ATPázy závislé na aktinu.

AK1YXQ Crystal structure of actin in complex with swinholide A, dále je klasifikován jako kontrakční protein, pochází z organismu Oryctolagus cuniculus, má 750 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: pozitivní regulace exprese genu, polymerizace aktinových vláken, sestava tenkých vláken kosterních svalů, vývoj kosterních svalových vláken, shromáždění sady aktinových vláken, migrace mesenchymu, pozitivní regulace aktivity ATPázy závislé na aktinu.

Mezi myoziny jsou zařazeny následující vzorky:

MY2DRK Acanthamoeba myosin I SH3 domain bound to Acan125, dále je klasifikován jako kontrakční protein, pochází z organismu Acanthamoeba castellanii, má 69 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, databáze RCSB PDB neuvádí, jakých biologických procesů se vzorek účastní.

MY2DRM Acanthamoeba myosin I SH3 domain bound to Acan125, dále je klasifikován jako kontrakční protein, pochází z organismu Acanthamoeba castellanii, má 268 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, databáze RCSB PDB neuvádí, jakých biologických procesů se vzorek účastní.

MY1YV3 The structural basis of blebbistatin inhibition and specificity for myosin II, dále je klasifikován jako kontrakční protein, pochází z organismu Dictyostelium discoideum, má 762 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, databáze RCSB PDB neuvádí, jakých biologických procesů se vzorek účastní.

MY2FXM Structure of the human beta-myosin S2 fragment, dále je klasifikován jako kontrakční protein, pochází z organismu Homo sapiens, má 762 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, databáze RCSB PDB neuvádí, jakých biologických procesů se vzorek účastní.

MY2FXO Structure of the human beta-myosin S2 fragment, dále je klasifikován jako kontrakční protein, pochází z organismu Homo sapiens, má 516 aminokyselinových reziduí, v sekvenci se nachází 1 mutace, databáze RCSB PDB neuvádí, jakých biologických procesů se vzorek účastní.

Mezi troponiny jsou zařazeny následující vzorky:

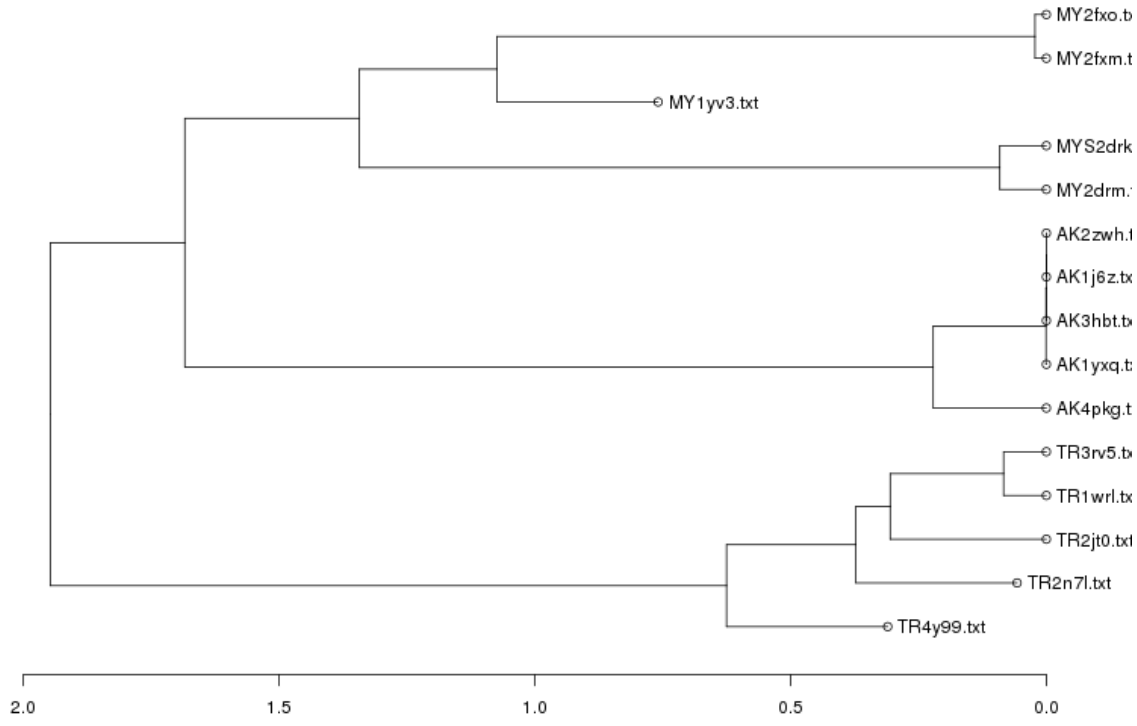
TR2N7L NMR structure of the N-domain of troponin C bound to the switch region of troponin I and the covalent levosimendan analog i9, dále je klasifikován jako kontrakční protein, pochází z organismu Homo sapiens, má 141 aminokyselinových reziduí, v sekvenci se nachází 1 mutace, databáze RCSB PDB neuvádí, jakých biologických procesů se vzorek účastní.

TR4Y99 Core domain of human cardiac troponin, dále je klasifikován jako kontrakční protein, pochází z organismu Homo sapiens, má 477 aminokyselinových reziduí, v sekvenci se nachází 4 mutace, účastní se následujících biologických procesů: kontrakce kosterních svalů, regulace svalové kontrakce, regulace srdečních svalů.

TR3RV5 Crystal structure of human cardiac troponin C regulatory domain in complex with cadmium and deoxycholic acid, dále je klasifikován jako kontrakční protein, pochází z organismu Homo sapiens, má 356 aminokyselinových reziduí, v sekvenci se nachází 4 mutace, nemá žádnou biologickou funkci.

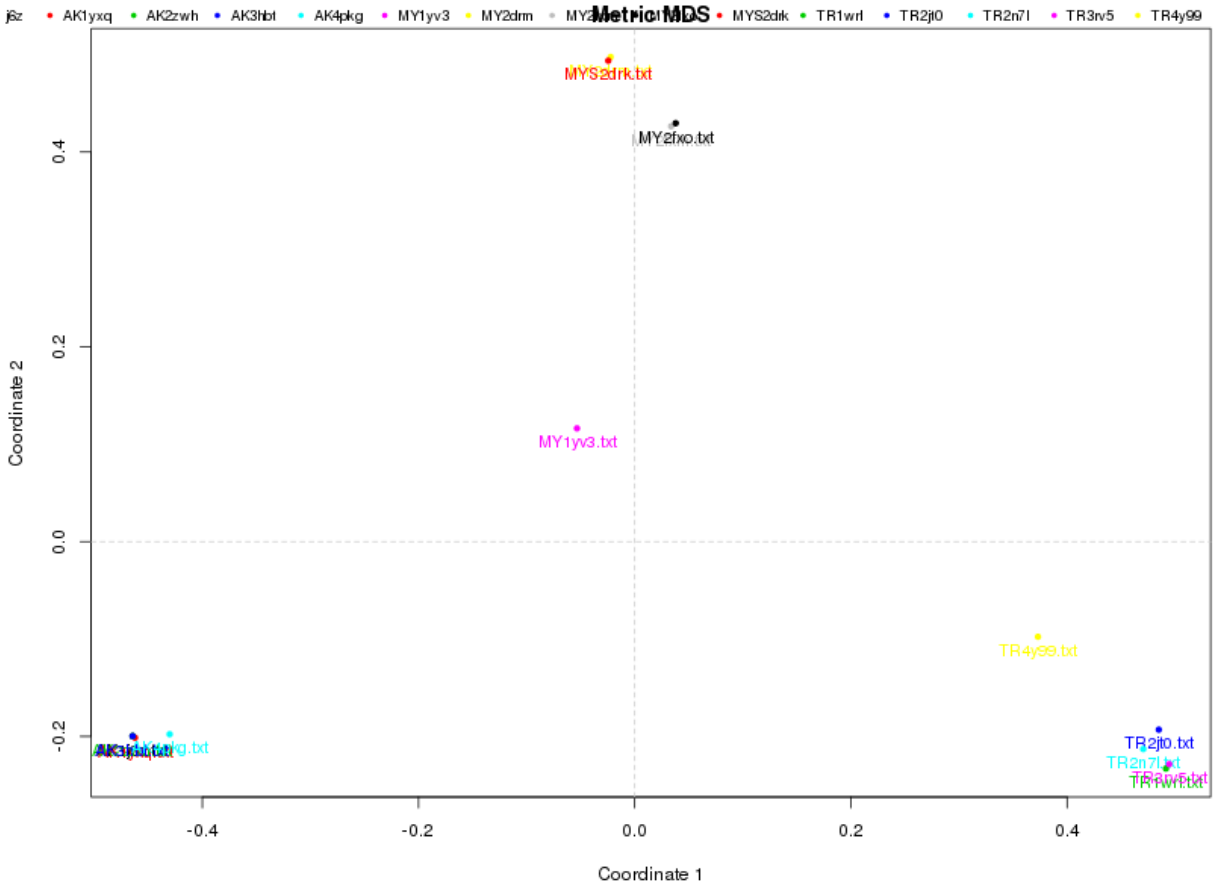
TR1WRL Crystal structure of the N-terminal domain of human cardiac troponin C in complex with trifluoperazine (monoclinic crystal form), dále je klasifikován jako kontrakční protein, pochází z organismu Homo sapiens, má 528 aminokyselinových reziduí, v sekvenci se nachází 2 mutace, nemá žádnou biologickou funkci.

TR2JT0 Solution structure of F104W cardiac troponin C, dále je klasifikován jako kontrakční protein, pochází z organismu Homo sapiens, má 161 aminokyselinových reziduí, v sekvenci se nachází 3 mutace, účastní se následujících biologických procesů: kontrakce kosterních svalů, regulace svalové kontrakce, regulace srdečních svalů.



Obr. 22 Dendrogram (hierarchické shlukování)

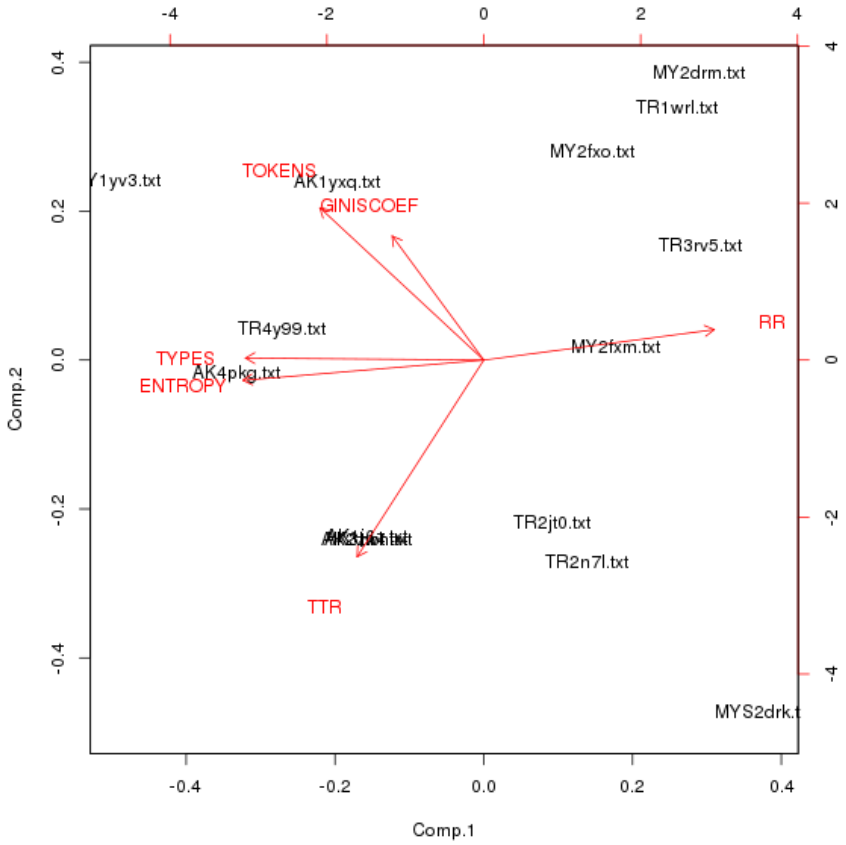
Graf hierarchického klastrování zobrazuje kontrakční a pohybové proteiny ve dvou větvích. Na spodní samostatné větvi se nachází vzorky troponinu, z čehož lze usuzovat, že tento protein je mezi ostatními kontrakčními a pohybovými proteiny výrazně specifický. Mezi vzorky troponinu můžeme nalézt podobnosti, které souvisí s biologickými procesy, kterých se vzorky účastní. Nejspecifičtější z nich je pak vzorek TR4Y99, který má výrazně více biologických funkcí, což odpovídá jeho poloze v grafu. Ostatní troponiny se biologických procesů neúčastní, kromě vzorku TR2N7L, který je činitelem v podobných procesech jako protein TR4Y99. Oba se účastní například regulace svalových kontrakcí. Na druhé větvi grafu se nachází v oddělených skupinách vzorky myozinu a aktinu, z čehož lze usuzovat, že obě tyto skupiny proteinů vykazují podobnost. Čtyři z pěti vzorků aktinu leží na stejné koncové větvi. Vzorek AK4PKG se od nich liší svou schopností vázat na sebe další aktin. Skupina myozinů je dále větvena na dvojici a trojici.



Obr. 23 MDS (vícerozměrové škálování)

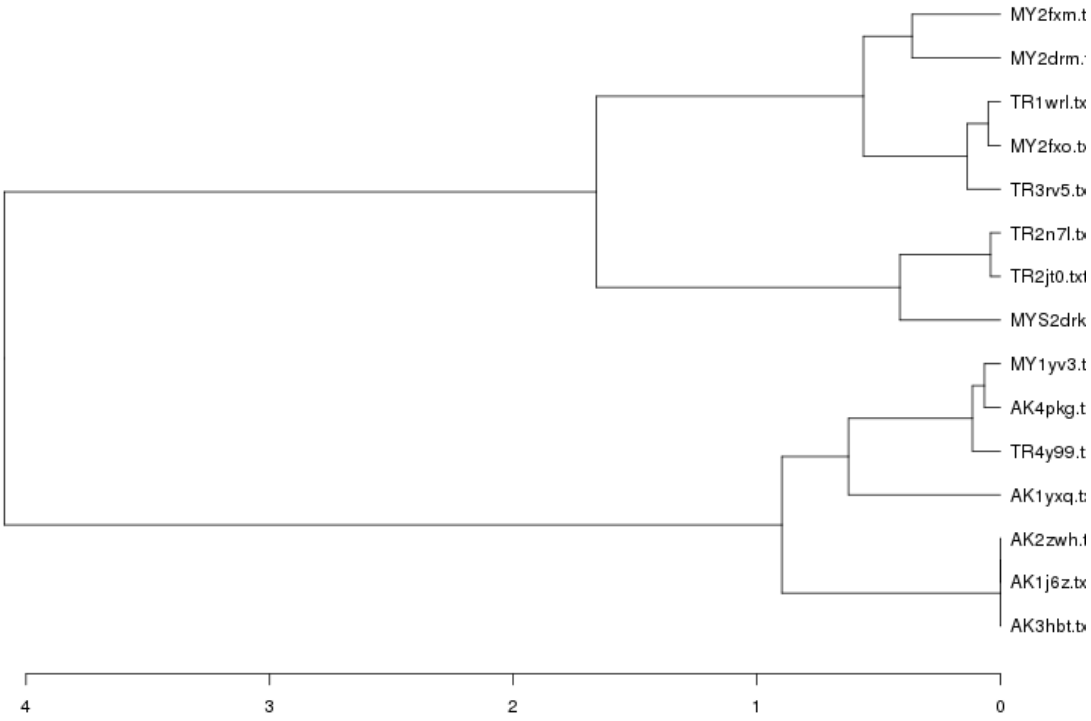
Graf multidimenzionálního škálování zobrazuje několik shluků, které podporují interpretaci předchozího grafu. Vzorky troponinu se v grafu vyskytují v jedné skupině a

stejně tak se shlukují i aktiny. Mimo skupinu dalších myozinů se nachází vzorek MY1YV3, který i v předchozím grafu leží na větvi dál od ostatních vzorků.



Obr. 24 PCA (analýza hlavních komponent)

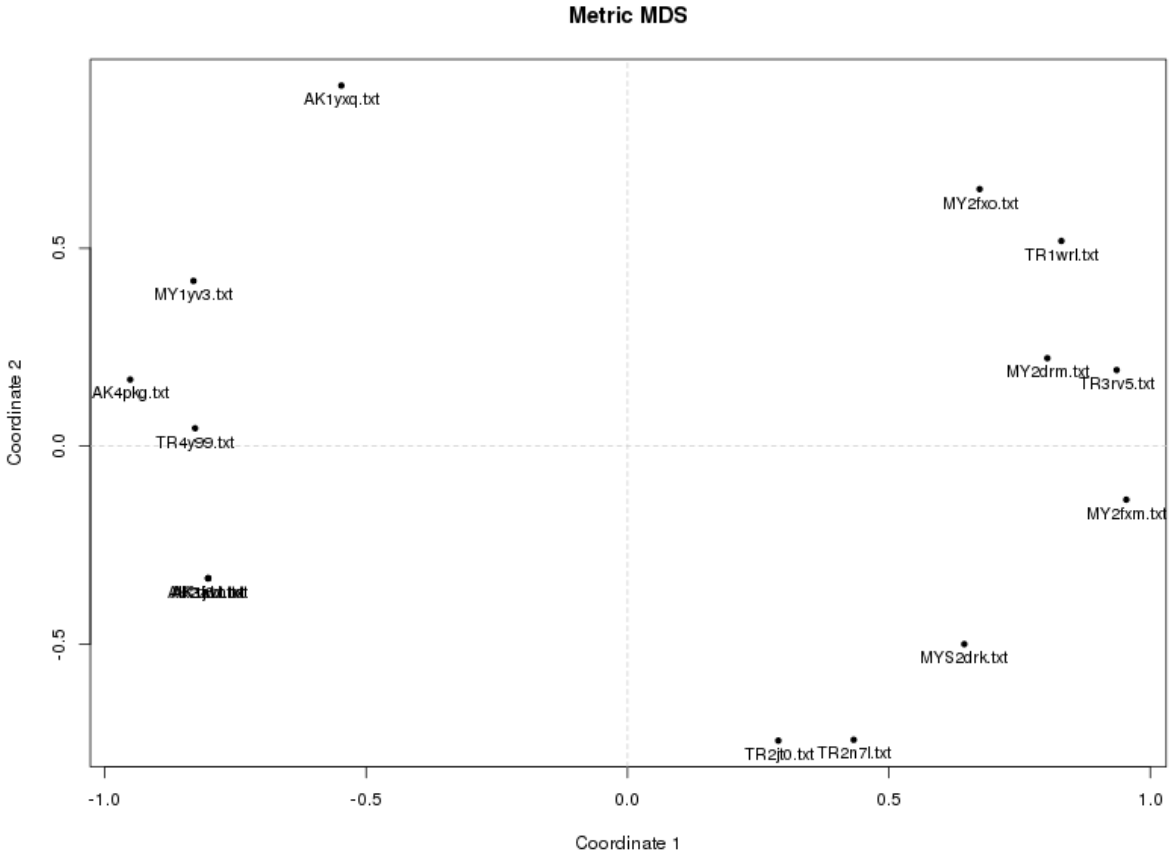
Podle grafu analýzy hlavních komponentů můžeme určit podobnost vzorků myozinu na základě hodnot indexů RR a Giniho koeficientu. Vzorek MY1YV3 se hodnotami indexů výrazně odlišuje nejen od ostatních vzorků myozinu, ale i od zbylých proteinů, a to hodnotami tokenů a typů. Dále můžeme pozorovat skupinu aktinů, které mají natolik podobnou hodnotu TTR, že se v grafu překrývají. Z grafu je dále pozorovatelné, že vzorek TR4Y99 se od ostatních troponinů odlišuje svou hodnotou typů a tokenů.



Obr. 25 Dendrogram (hierarchické shlukování)

Graf hierarchického škálování potvrzuje interpretaci předchozího grafu. I zde je možné pozorovat podobnost vzorků aktinu, které se nacházejí na jedné větvi. V jejich bezprostřední blízkosti se vyskytují vzorky TR4Y99 a MY1YV3, které byly výrazně odlišné od ostatních i v předchozích grafech. Oproti předchozímu dendrogramu se v tomto grafu sloučily vzorky myozinu a troponinu na společnou větev, což vypovídá o podobných

hodnotách naměřených indexů.



Obr. 26 MDS (vícerozměrové škálování)

I v tomto grafu multidimenzionálního škálování pozorujeme překrývání skupiny aktinů, stejně jako odlišnost vzorků TR4Y99 a MY1YV3 od ostatních vzorků patřících do stejných skupin. V tomto grafu se v jejich blízkosti vyskytuje také aktin AK4PKG, který má podobnou hodnotu tokenů a indexu entropie.

4.1.6 Vzorky defenzivních proteinů a jejich popis na základě modelu BoW a vybraných indexů

Pro analýzu defenzivních proteinů využívá tato práce deseti vzorků, které jsou dále řazeny do dvou skupin po pěti podle vymezených podskupin, které slouží jako příklady zástupců proteinů s výše zmíněnou funkcí. Jednotlivé dále vydělené podskupiny jsou proteiny imunoglobuliny a další proteiny s defenzivní funkcí.

Mezi imunoglobuliny jsou zařazeny následující vzorky:

IG1MRF Preparation, Characterization and Crystalization of an Antibody FAB Fragment that Recognizes RNA. Crystal Structures of Native FAB and three FAB-

Mononucleoid Complexes, dále je klasifikován jako imunoglobulin, organismus, ze kterého vzorek pochází, není uveden, má 434 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, neúčastní se žádných biologických procesů.

IG1OL0 Crystal structure of a camelised human VH, dále je klasifikován jako imunoglobulin, organismus, ze kterého vzorek pochází, není uveden, má 242 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, neúčastní se žádných biologických procesů.

IG1OHQ Crystal structure of HEL4, a soluble human VH antibody domain resistant to aggregation, dále je klasifikován jako imunoglobulin, pochází z organismu Homo sapiens, má 242 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: procesy imunitního systému, proteolýza.

IG1BRE Immunoglobulin Light Chain Protein, dále je klasifikován jako imunoglobulin, organismus, ze kterého pochází, není uveden, má 648 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, neúčastní se žádných biologických procesů.

IG4X99 Immunoglobulin Fc heterodimers variant, dále je klasifikován jako součást imunitního systému, pochází z organismu Homo sapiens, má 466 aminokyselinových reziduí, v sekvenci se nachází 7 mutací, neúčastní se žádných biologických procesů.

Mezi proteiny s defenzivní funkcí jsou zařazeny následující vzorky:

DF1IOA Arcelin-5, A Lecitin-Like Defense Protein From Phaseolus vulgaris, dále je klasifikován jako lecitin, pochází z organismu Phaseolus vulgaris, má 480 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: defenzivní reakce, patogeneze.

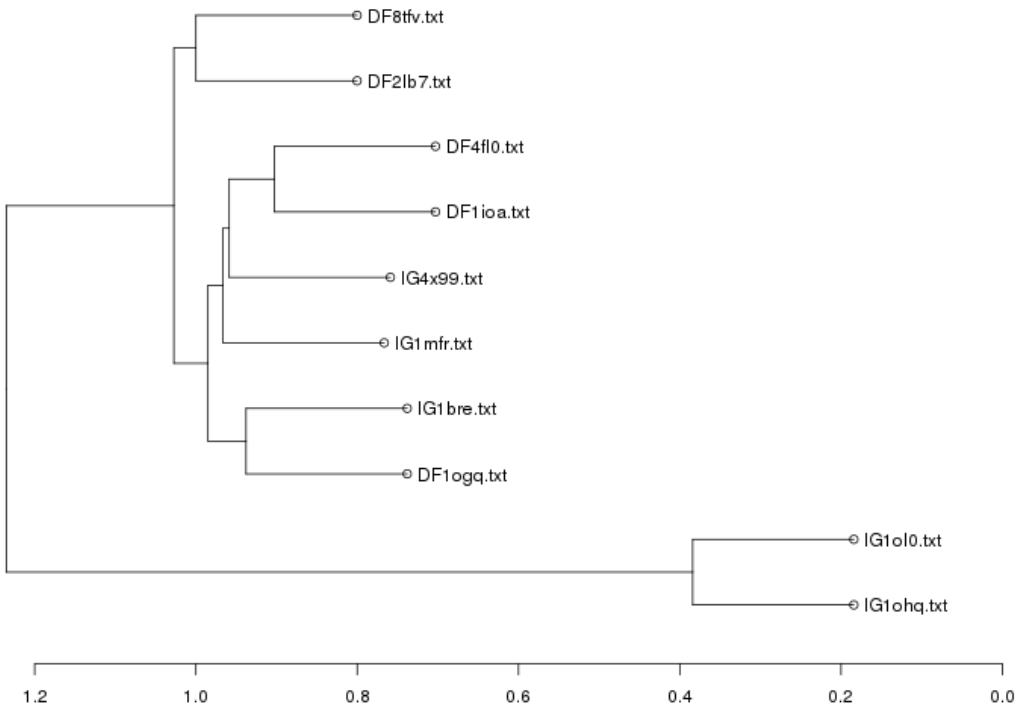
DF8TFV Insect Defense Peptide, dále je klasifikován jako antimikrobický, pochází z organismu Podisus maculiventris, má 21 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: procesy imunitního systému, hubení buněk cizích organismů, defenzivní reakce na bakterie, defenzivní reakce na houby.

DF1OGQ The crystal structure of PGIP (polygalacturonase inhibiting protein), a leucine rich repeat protein involved in plant defense, dále je klasifikován jako inhibitor,

pochází z organismu Phaseolus vulgaris, má 313 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: defenzivní reakce.

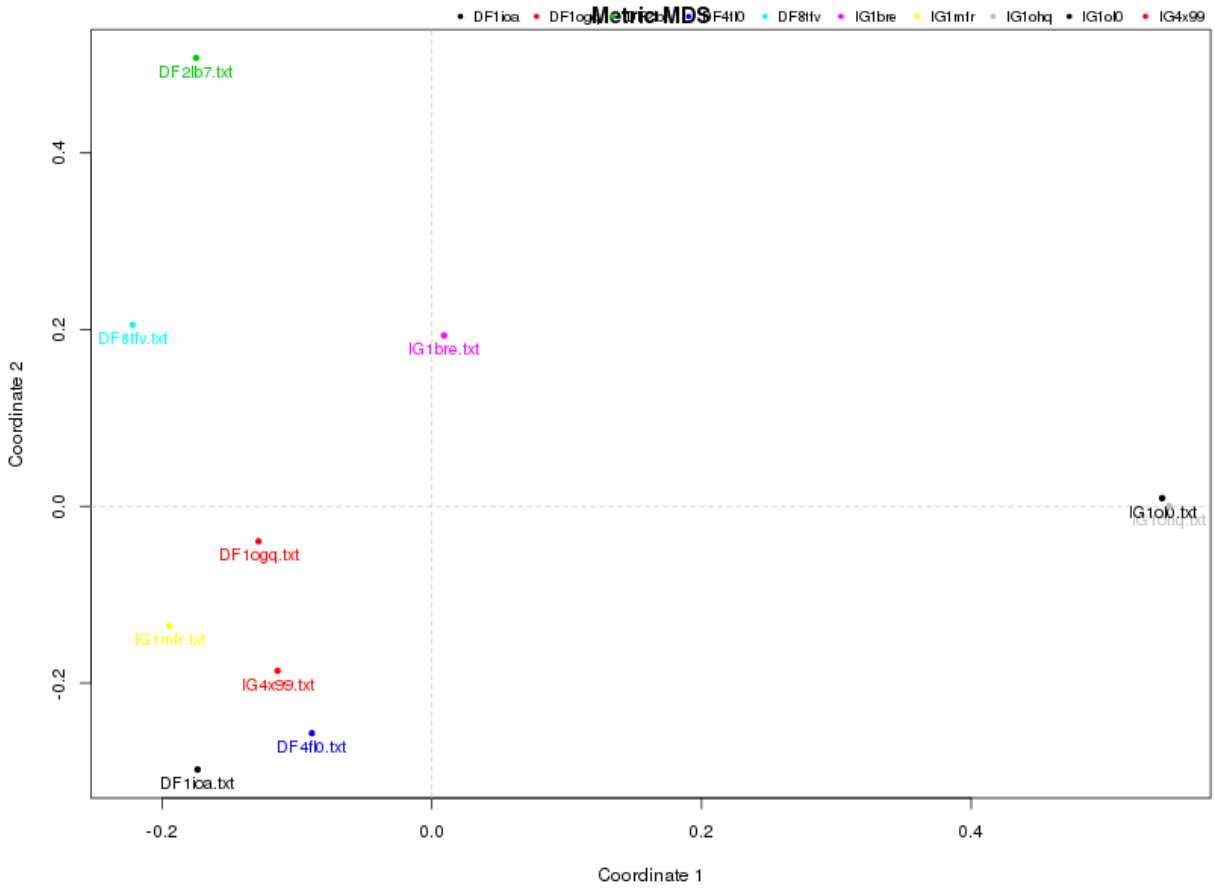
DF4FL0 Crystal structure of ALD1 from Arabidopsis thaliana, dále je klasifikován jako transferáza, pochází z organismu Arabidopsis thaliana, má 912 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: defenzivní reakce, biosyntetické procesy, defenzivní reakce na bakterie.

DF2LB7 Hevein-type Antifungal Peptide with a Unique 10-Cysteine Motif, Classification: dále je klasifikován jako antimikrobický, pochází z organismu Triticum kiharae, má 44 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: defenzivní reakce, hubení buněk cizích organismů, defenzivní reakce na bakterie, defenzivní reakce na houby.



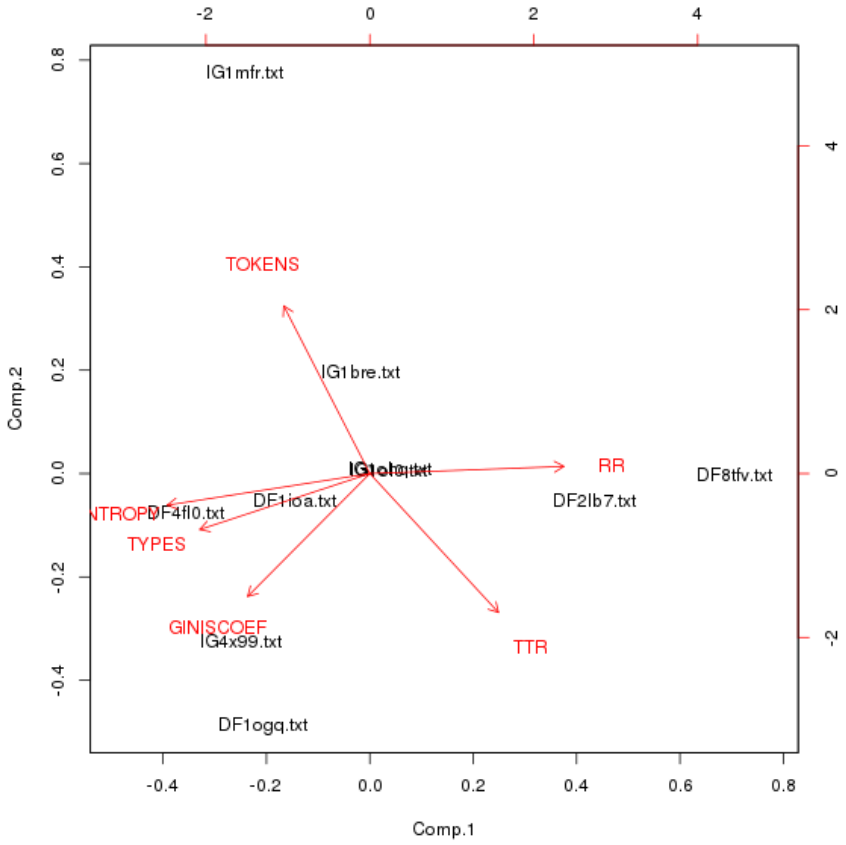
Obr. 27 Dendrogram (hierarchické shlukování)

Tento dendrogram člení skupinu defenzivních proteinů do dvou větví. Na první z nich se samostatně nacházejí dva vzorky imunoglobulinu, IG1OHQ a IG1O10. V obou případech se jedná o krystalovou strukturu proteinu, čímž můžeme interpretovat jejich vzájemnou podobnost a odlišnost od ostatních vzorků. Na druhé větvi se nalézají zbylé tři vzorky imunoglobulinu, které se neúčastí žádných biologických procesů, a další defenzivní proteiny. Největší podobnost pak vykazují defenzivní proteiny DF2L67 a DF8TFV, které jsou oba peptidy, což vysvětluje jejich blízkost v grafu.



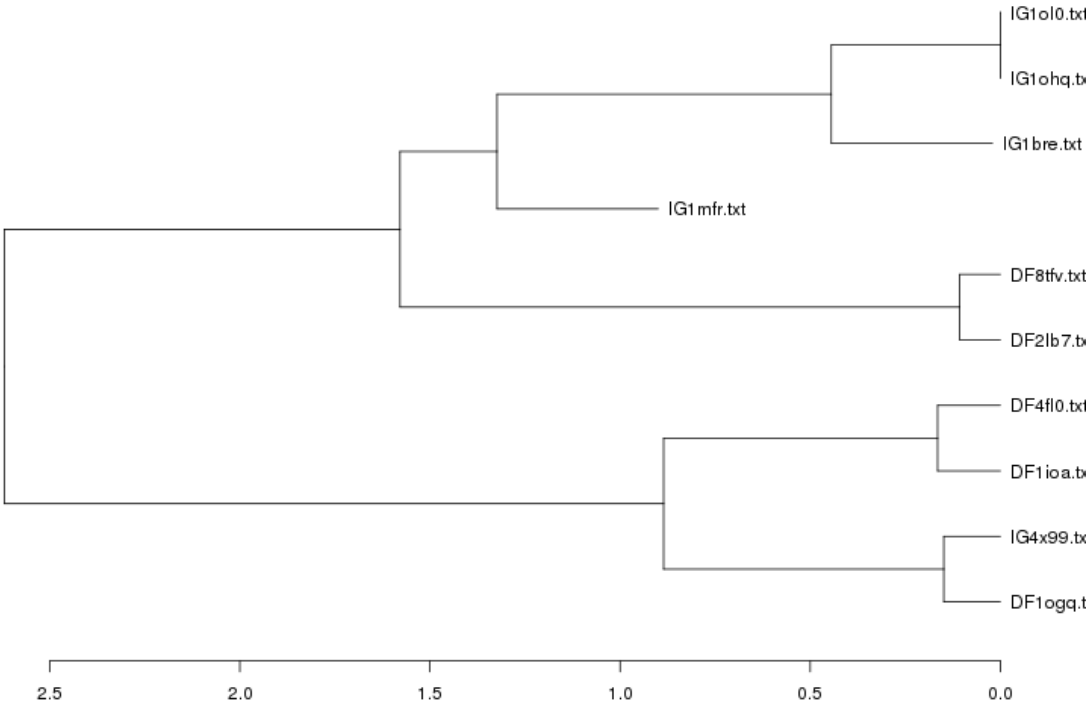
Obr. 28 MDS (vícerozměrové škálování)

V tomto grafu multidimenzionálního škálování pozorujeme jednu větší skupinu vzorků, která odpovídá druhé větvi předchozího dendrogramu, na které se nachází skupina imunoglobulinů a defenzivních proteinů. Z grafu je rovněž patrná podobnost imunoglobulinů IG1OHQ a IG1O10, které se překrývají, což podporuje předchozí interpretaci.



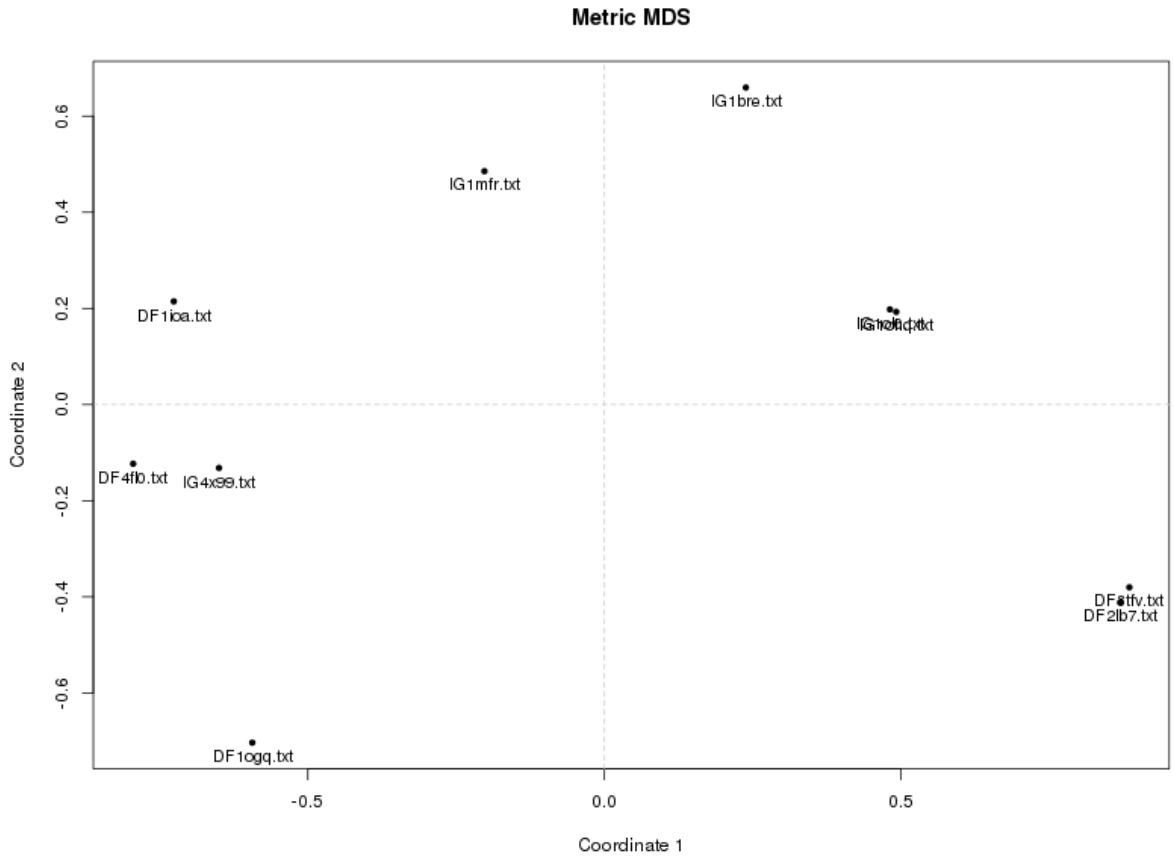
Obr. 29 PCA (analýza hlavních komponent)

Z grafu analýzy hlavních komponentů můžeme vypozorovat, že na základě naměřeného indexu vykazují podobnost vzorky DF2L67 a DF8TFV, což podporuje předchozí interpretaci jejich podobnosti. Ve středu grafu se pak nalézají imunoglobuliny IG1OHQ a IG1O10, které se překrývají, což potvrzuje předchozí tvrzení o jejich podobnosti. Odlišnost naopak vidíme u hemoglobinu IG1MRF, který se liší hodnotou tokenů od všech dalších použitých vzorků.



Obr. 30 Dendrogram (hierarchické shlukování)

V tomto dendrogramu můžeme pozorovat opět dvě větve. Tentokrát se na první větvi nachází defenzivní proteiny a vzorek imunoglobulinu IG4X99, který se od zbylých vzorků liší především hodnotou Giniho koeficientu. Další větev pak tvoří zbylé imunoglobuliny, což potvrzuje tezi o podobnosti mezi těmito vzorky.



Obr. 31 MDS (vícerozměrové škálovaní)

V grafu multidimenzionálního škálování opět můžeme vidět překrývající se vzorky dvou imunoglobulinů, což rovněž potvrzuje jejich podobnost. Dále pozorujeme blízkost dvou defenzivních proteinů, DF2L67 a DF8TFV, což prokazuje jejich podobnost. Odlišnost znovu vidíme u hemoglobinu IG1MRF, což potvrzuje jeho odlišnost od ostatních imunoglobulinů.

4.1.7 Vzorky regulačních proteinů a jejich popis na základě modelu BoW a vybraných indexů

Pro analýzu regulačních proteinů využívá tato práce patnácti vzorků, které slouží jako příklady zástupců proteinů s regulační funkcí. Analyzovány jsou vzorky regulačních proteinů, které se váží na DNA a další regulační proteiny.

Mezi proteiny s regulační funkcí, které se vážou na DNA, jsou zařazeny následující vzorky:

DN1A0A Phosphate System Positive Regulatory Protein PHO4/DNA Complex, pochází z organismu Saccharomyces cerevisce, dále je klasifikován jako protein účastnící se transkripce DNA, vzorek má 160 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, neúčastní se žádného biologického procesu.

DN1DBQ DNA-Binding Regulatory Protein, pochází z organismu Escherichia coli, dále je klasifikován jako DNA-BINDING REGULATORY PROTEIN, vzorek má 578 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, v databázi RCSB PDB není uvedeno, jakých biologických procesů se vzorek účastní.

DN2ARA APO Form of Escherichia coli Regulatory Protein ARAC, pochází z organismu Escherichia coli, dále je klasifikován jako regulátor transkripce, vzorek má 149 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: regulace transkripce templátové DNA.

DN2OI8 Crystal structure of putative regulatory protein SCO4313, pochází z organismu Streptomyces coelicolor, dále je klasifikován jako protein účastnící se transkripce, vzorek má 216 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: transkripce templátové DNA, regulace transkripce templátové DNA.

DN2RVJ NMR structure of Epithelial splicing regulatory protein 1, pochází z organismu Homo sapiens, dále je klasifikován jako protein účastnící se transkripce, vzorek má 104 aminokyselinových reziduí, v sekvenci se nachází 1 mutace, vzorek se neúčastní žádného biologického procesu.

DN3D5L Crystal structure of regulatory protein RecX, pochází z organismu Lactobacillus reuteri, dále je klasifikován jako SIGNALING PROTEIN, vzorek má 442 aminokyselinových reziduí, účastní se následujících biologických procesů: regulace oprav DNA.

DN16VP Conserved Core of the Herpes Simplex Virus Transcriptional Regulatory Protein VP16, pochází z organismu Human herpesvirus 1, dále je klasifikován jako protein regulující transkripci, vzorek má 366 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: regulace transkripce templátové DNA**.**

Mezi proteiny s regulační funkcí jsou zařazeny následující vzorky:

RG1A6J Nitrogen Regulatory Bacterial Protein IIA-Nitrogen, pochází z organismu Escherichia coli, dále je klasifikován jako fosfotransferáza, vzorek má 326 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: fosforylace, transmembránový transport karbohydrátů, negativní regulace katalytické aktivity.

RG1GLD Cation Promoted Association (CPA) of a Regulatory and Target Protein is Conterolled by Phosphorylation, pochází z organismu Escherichia coli, dále je klasifikován jako fosfotransferáza, vzorek má 669 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: transport karbohydrátů, fosforyzace, negativní regulace katalytické aktivity.

RG1QB3 Crystal Structure of the Cell Cycle Regulatory Protein CKS1, pochází z organismu Saccharomyces cerevisce, dále je klasifikován jako protein účastnící se buněčného cyklu, vzorek má 450 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: regulace transkripce RNA polymerázy, buněčný cyklus, regulace miotického buněčného cyklu, aktivace proteinové aktivity kinázy, buněčné dělení.

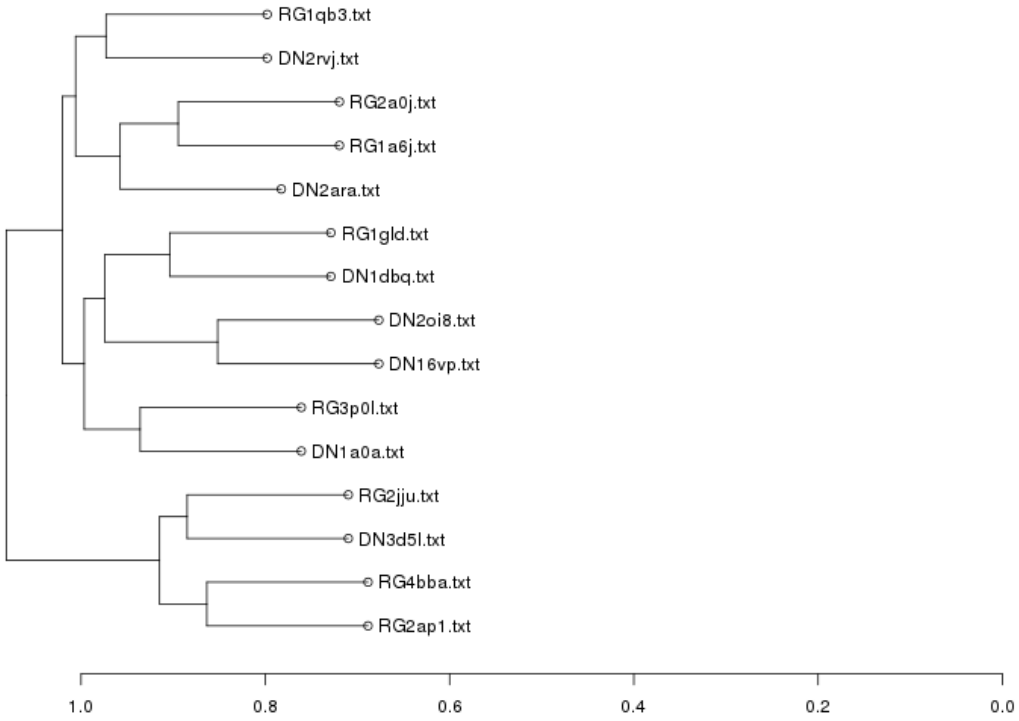
RG2A0J Crystal Structure of Nitrogen Regulatory Protein IIA-Ntr from Neisseria meningitidis, pochází z organismu Neisseria meningitidis serogroup B, dále je klasifikován jako transferáza, vzorek má 149 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: fosforylace, transmembránový transport karbohydrátů.

RG2AP1 Crystal structure of the putative regulatory protein, pochází z organismu Salmonella typhimurium, dále je klasifikován jako transferáza, vzorek má 327 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: metabolický proces karbohydrátů, fosforylace, karbohydrátová fosforylace.

RG2JJU Structure of human signal regulatory protein (sirp) beta, pochází z organismu Homo sapiens, dále je klasifikován protein, který je součástí imunitního systému, vzorek má 254 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, databáze RCSB PDB neuvádí, jakých biologických procesů se vzorek účastní.

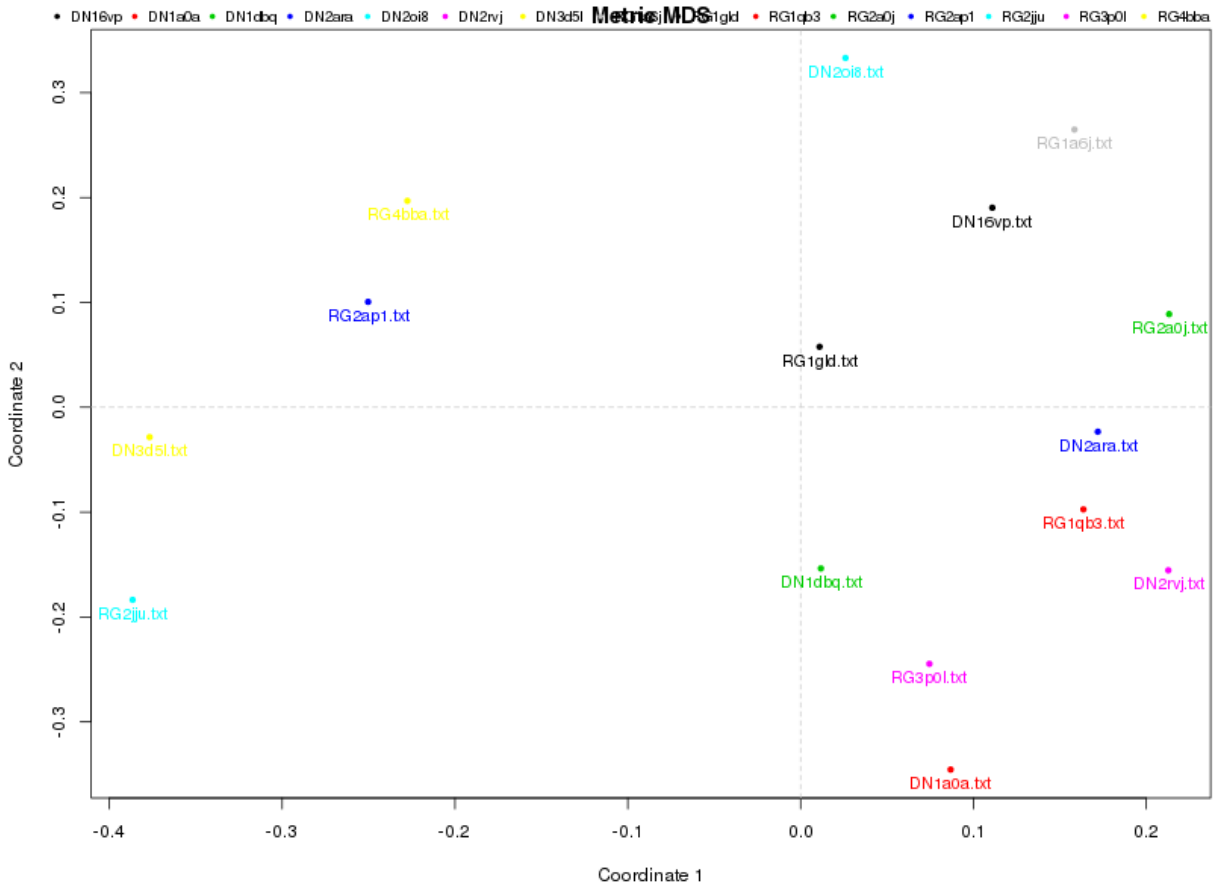
RG3P0L Human steroidogenic acute regulatory protein, pochází z organismu Homo sapiens, dále je klasifikován jako transportní protein, vzorek má 884 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: biosyntetický proces steroidů.

RG4BBA Crystal structure of glucokinase regulatory protein complexed to phosphate, pochází z organismu Homo sapiens, dále je klasifikován jako proteiny vázající protein, vzorek má 663 aminokyselinových reziduí, v sekvenci se nachází 2 mutace, účastní se následujících biologických procesů: import proteinu do nukleové translokace, nukleová translokace, metabolický proces karbohydrátů, regulace glykotického procesu, responze na fruktózu, negativní regulace aktivity glukokinázy.



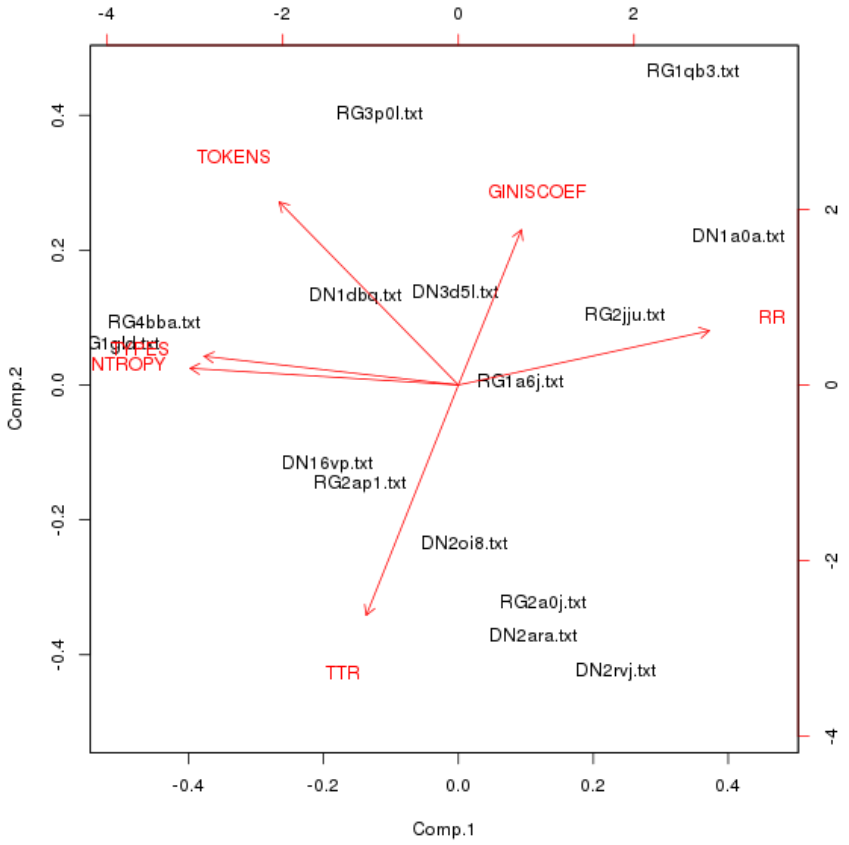
Obr. 32 Dendrogram (hierarchické shlukování)

Graf hierarchického škálování člení regulační proteiny do dvou hlavních větví. Ve spodní větvi vidíme tři zástupce regulačních proteinů a jeden regulační protein, který se váže na DNA. Další větev tvoří zbývající zástupci regulačních proteinů a regulačních proteinů, které se vážou na DNA. Z takového výsledku škálování můžeme obecně usoudit, že vzhledem k biologické roli regulace vykazují vybrané vzorky podobnost.



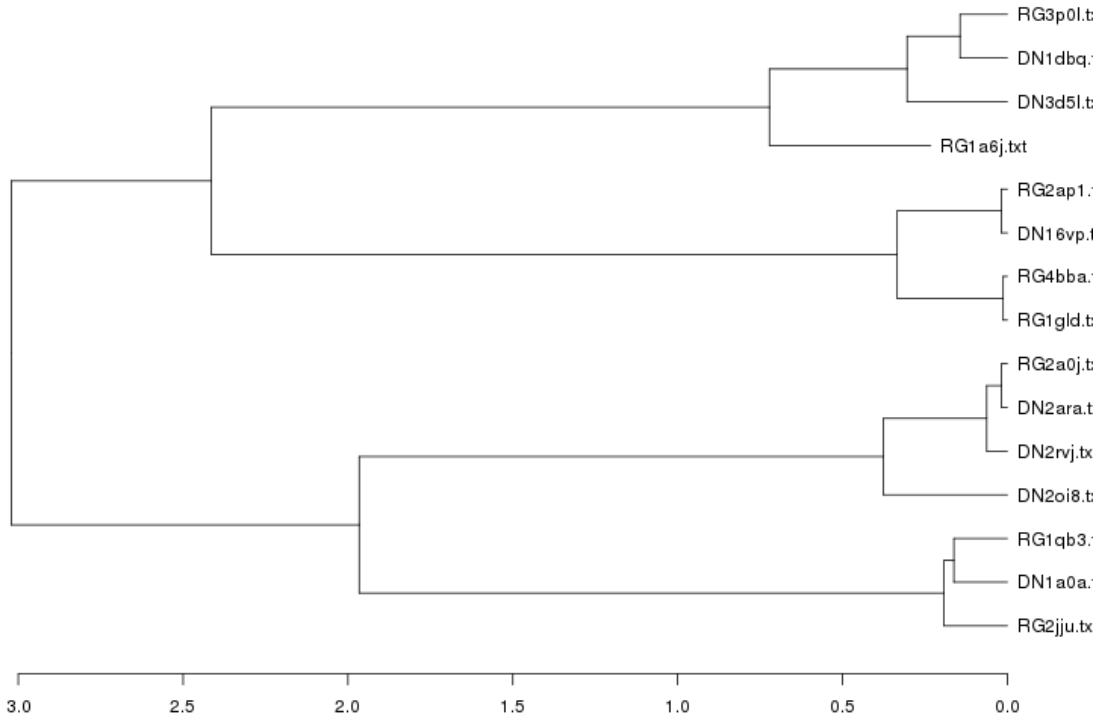
Obr. 33 MDS (vícerozměrové škálování)

V grafu multidimenzionálního škálování můžeme opět pozorovat dva shluky se smíšenými skupinami regulačních proteinů a proteinů vázajících se na DNA. V levé části grafu se nacházejí proteiny umístěné v předchozím grafu na první větvi, napravo se pak vyskytuje další skupina, která odpovídá druhé větvi.



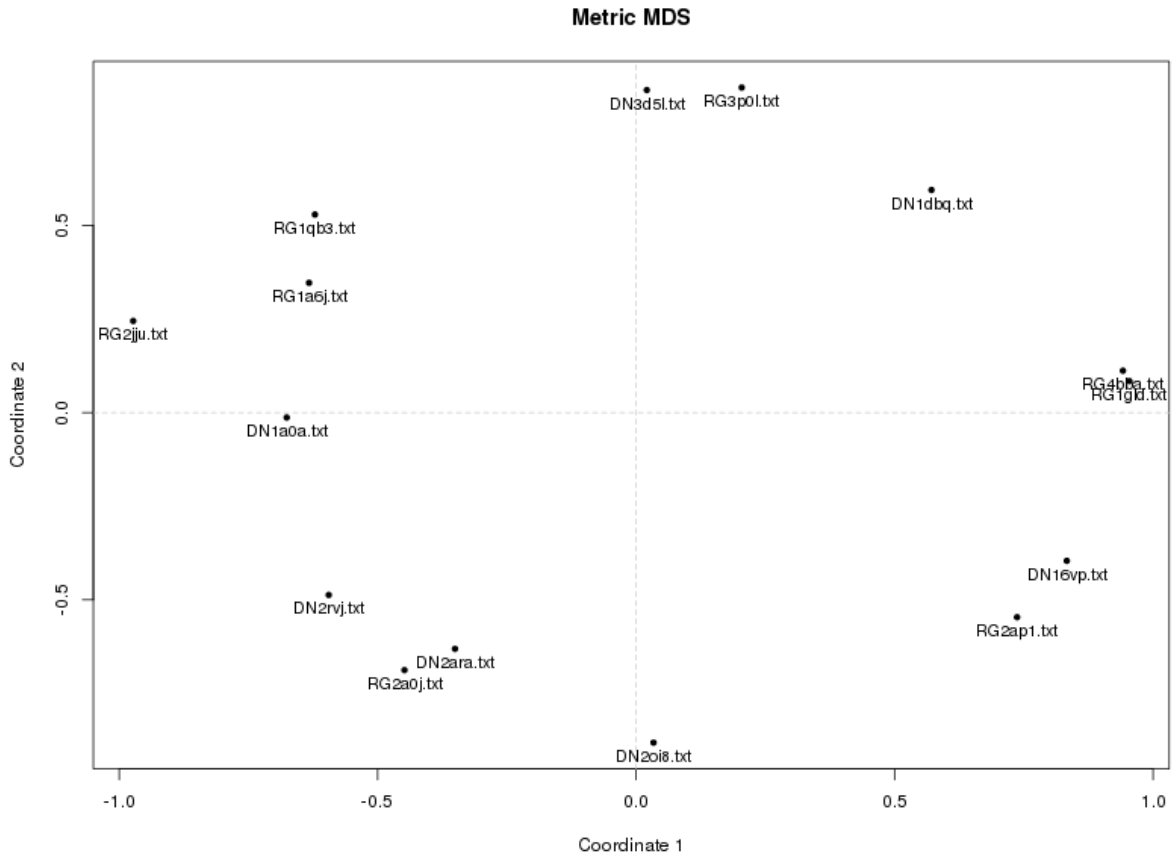
Obr. 34 PCA (analýza hlavních komponent)

V grafu analýzy hlavních komponentů můžeme vidět podobnosti vzorků v rámci hodnot jejich indexů. Nejblíže jsou si v grafu vzorky RG4BBA a RG1GLD. Jejich podobnost odráží hodnoty entropie a hodnoty typů. Dále jsou si v grafu blízké dva vzorky regulačních proteinů, které se vážou na DNA, konkrétně DN1DBQ a DN3d5L, které mají podobný počet tokenů a hodnotu Giniho koeficientu. Podobnost hodnot Giniho koeficientu a indexu RR pak můžeme pozorovat u vzorků RG1A6J a RG2JJU.



Obr. 35 Dendrogram (hierarchické shlukování)

Tento graf hierarchického škálování odráží podobnosti vzorků na základě indexů. Tento graf se skládá ze dvou větví, přičemž v obou větvích se vedle sebe vyskytují jak regulační proteiny, tak ty které se vážou na DNA. I tento graf podporuje tvrzení o tom, že všechny vybrané vzorky regulačních proteinů vykazují podobnost bez ohledu na to, do jaké skupiny jsou dále zařazeny.



Obr. 36 MDS (vícerozměrové škálování)

Posledním z grafů, který vizualizuje vztahy regulačních proteinů, je model multidimenzionálního škálování, který reflektuje podobnosti vzorků na základě zvolených indexů. I v tomto grafu lze vidět, jak se překrývají vzorky RG1A6J a RG2JJU, které jsou si podobné na základě dvou indexů, jak bylo zmíněno v interpretaci předchozího grafu. Dále můžeme pozorovat blízkost proteinů DN2ARA a RG2AJO, které se v předchozím grafu hierarchického škálování vykytovaly vedle sebe na dvou koncích rozdílných větví. Tyto vzorky jsou si podobné hodnotami indexů TTR a RR.

4.1.8 Vzorky toxických proteinů a jejich popis na základě modelu BoW a vybraných indexů

Pro analýzu toxických proteinů využívá tato práce patnácti vzorků, které jsou řazeny do jedné skupiny a dále slouží jako příklady zástupců proteinů s toxickou funkcí. Všechny použité vzorky patří mezi toxiny.

TX1LJP Crystal Structure of beta-Cinnamomin Elicitin, pochází z organismu Phytophthora cinnamomi, dále je klasifikován jako toxin, vzorek má 196 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: transport lipidů, patogeneze.

TX3TSS Toxic Shock Syndrome Toxin-1 Tetramutant, P2(1) Crystal Form, pochází z organismu Staphylococcus aureus, dále je klasifikován jako toxin, vzorek má 194 aminokyselinových reziduí, v sekvenci se nachází 4 mutace, účastní se následujících biologických procesů: patogeneze.

TX1A8D Tetanus Toxin C Fragment, pochází z organismu Clostridium tetani, dále je klasifikován jako neurotoxin, vzorek má 452 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: patogeneze, negativní regulace sekrecí neurotransmiteru.

TX1BCG Scorpion Toxin BJXTR-IT, pochází z organismu Hottentotta judaicus, dále je klasifikován jako excitatorní neurotoxin, vzorek má 77 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: patogeneze.

TX1ETN Molecular Structure of the Toxic Domain of Heat-Stable Enterotoxin Produced by a Pathogenic Strain of Escherichia Coli, pochází z organismu Escherichia coli, dále je klasifikován jako enterotoxin, vzorek má 13 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: patogeneze.

TX1EXF Exfoliative Toxin A, pochází z organismu Staphylococcus aureus, dále je klasifikován jako komplexní toxin/peptid, vzorek má 242 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: proteolýza, patogeneze.

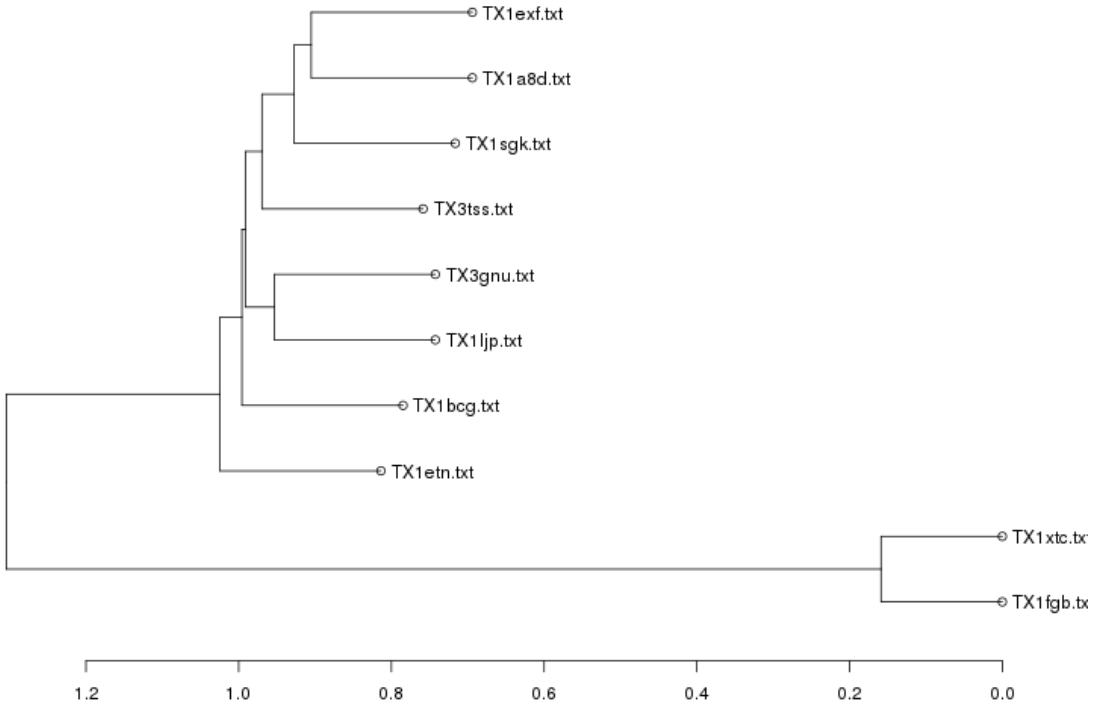
TX1FGB Toxin, pochází z organismu Vibrio cholerae serotype O1, dále je klasifikován jako enterotoxin, vzorek má 515 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: patogeneze, hubení buněk cizích organismů.

TX1SGK Nucleotide-Free Diphtheria Toxin, pochází z organismu Corynephage beta, dále je klasifikován jako toxin, vzorek má 535 aminokyselinových reziduí, v sekvenci

se nenachází žádná mutace, účastní se následujících biologických procesů: patogeneze, proteinový transmembránový transport.

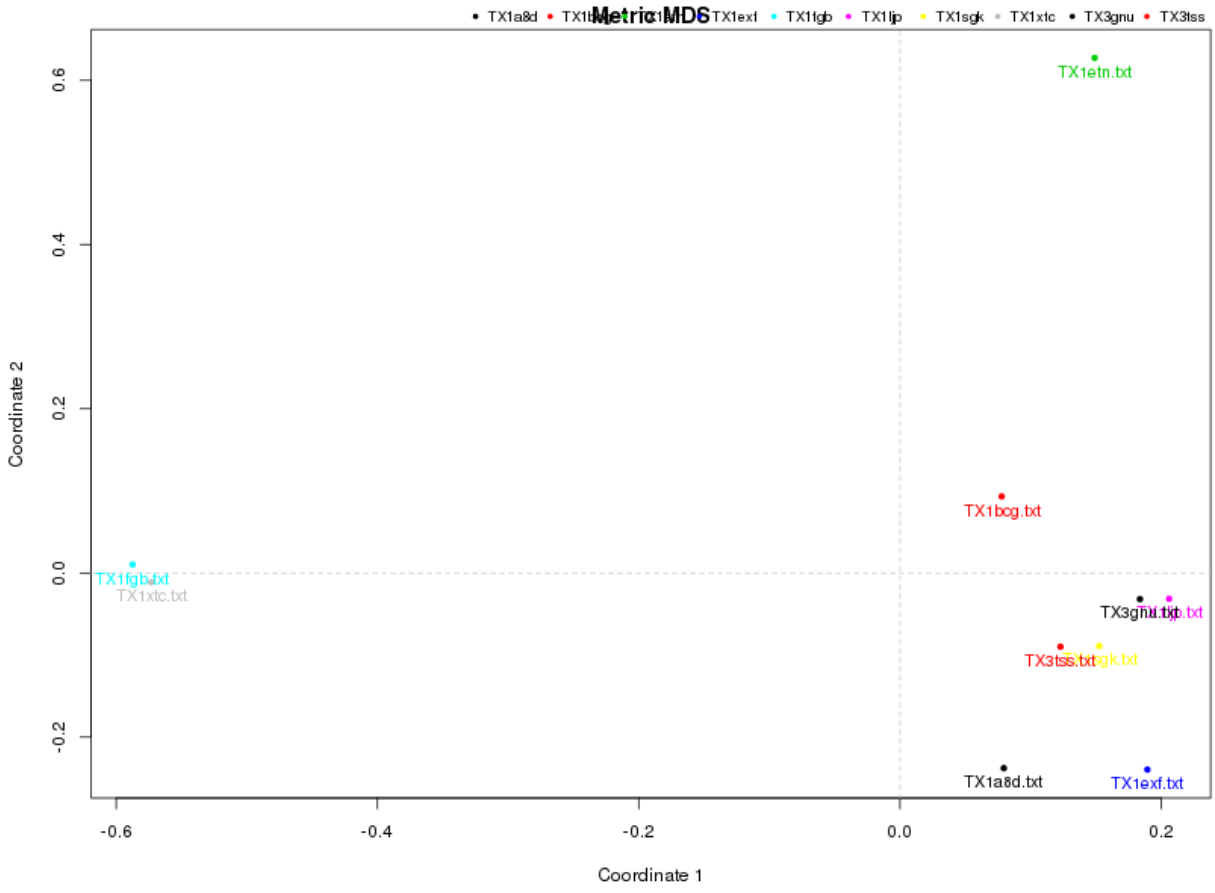
TX1XTC Cholera Toxin, pochází z organismu Vibrio cholerae serotype O1, dále je klasifikován jako toxin, vzorek má 755 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, účastní se následujících biologických procesů: patogeneze, hubení buněk cizích organismů.

TX3GNU Toxin fold as basis for microbial attack and plant defense, pochází z organismu Pythium aphanidermatum, dále je klasifikován jako toxin, vzorek má 213 aminokyselinových reziduí, v sekvenci se nenachází žádná mutace, neúčastní se žádného biologického procesu.



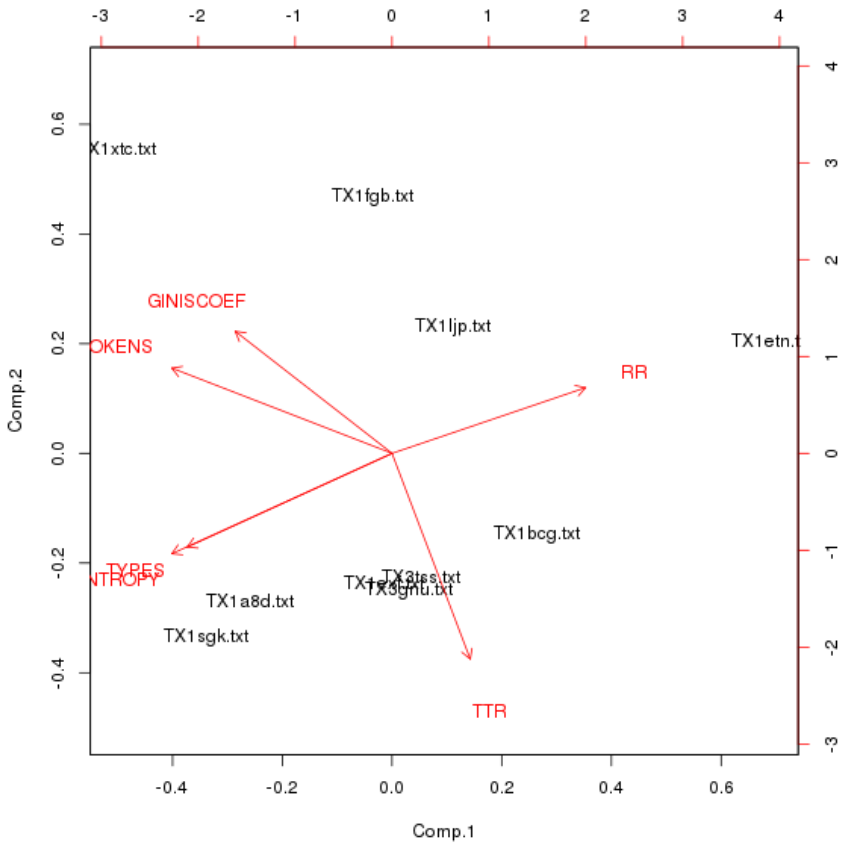
Obr. 37 Dendrogram (hierarchické shlukování)

V grafu hierarchického shlukování toxických proteinů můžeme sledovat dvě větve. Na první z nich se nachází pouze dva toxiny, TX1FGB a TX1XTC, z čehož můžeme usuzovat značnou podobnost. Oba tyto proteiny se na rozdíl od ostatních toxinů účastní biologického procesu hubení cizorodých organismů. Na druhé větvi grafu se vyskytují všechny ostatní vzorky toxinů. I u nich můžeme vysledovat podobnost na základě biologických procesů, kterých se účastní. Například se jedná o vzorky TX1A8D a TX1EXE, které se oba účastní patogeneze.



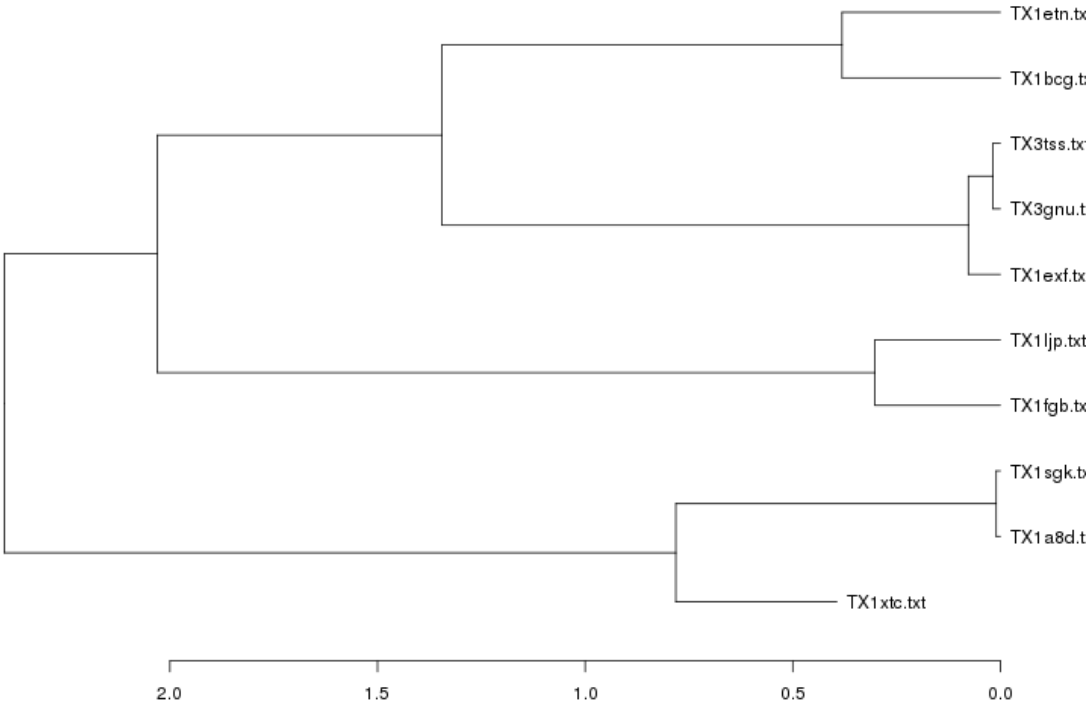
Obr. 38 MDS (vícerozměrové škálování)

V grafu multidimenzionálního škálování je rovněž patrné rozmístění vzorků do dvou skupin. V levé části grafu se nacházejí toxiny TX1FGB a TX1XTC, které v předchozím grafu tvořily samostatnou větev. Napravo pak vidíme druhou skupinu vzorků, která obsahuje zbylé toxiny. V této skupině se nacházejí dvě dvojice vzorků, které se v grafu překrývají. Jedná se o dvojici TX3TSS a TX1SGK, které se oba účastní patogeneze, takže jejich podobnost můžeme vykládat účastí na stejném biologickém procesu. Dalšími dvěma vzorky, které v grafu velmi blízce sousedí, jsou toxiny TX3GNU a TX1LJP. Tyto vzorky se vedle sebe nachází i ve výše uvedeném dendrogramu, podobnost je možné vysvětlit tím, že se v obou případech jedná o toxiny rostlinných proteinů, které mají obrannou funkci.



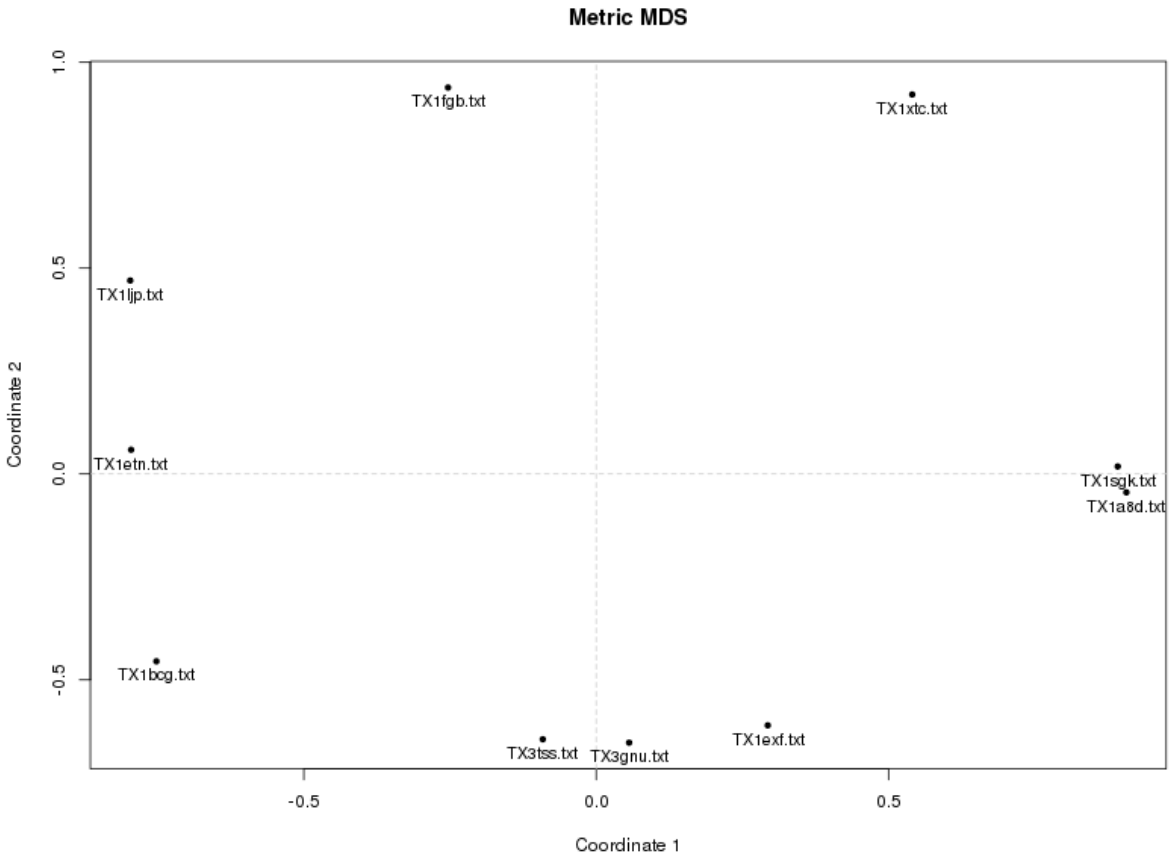
Obr. 39 PCA (analýza hlavních komponent)

V grafu analýzy hlavních komponentů můžeme sledovat hned několik podobností v rámci naměřených hodnot indexů. V grafu se nachází skupina proteinů, které jsou si tak blízké, že se překrývají. Jedná se o proteiny TX3TSS, TX1EXF a TX3GNU, které mají podobnou hodnotu TTR. V blízkosti této skupiny se nachází další dva toxiny, TX1A8D a TX1SGK, ty se však v grafu blíží spíše k indexu entropie a typů. Dále můžeme pozorovat vzorek TX1TC, jemuž je nejblíže právě výše zmíněná dvojice TX1A8D a TX1SGK, ale který se dále výrazně odlišuje od dalších vzorků skupiny toxických proteinů.



Obr. 40 Dendrogram (hierarchické shlukování)

Tento dendrogram dokládá interpretaci provedenou u předchozího grafu. I zde je možné pozorovat blízkost proteinů TX3TSS, TX1EXF a TX3GNU, které se v grafu nacházejí na společné větvi. Dále je i v tomto grafu zřetelná podobnost toxinů TX1A8D a TX1SGK, které rovněž leží vedle sebe. Těmto vzorkům je na základě zvolených indexů podobný i vzorek TX1TC, který s nimi leží na stejné samostatné větvi, ačkoli se od nich odděluje.



Obr. 41 MDS (vícerozměrové škálování)

I graf multidimenzionálního škálování znovu zobrazuje výše zmíněné vzorky, které jsou si ve všech grafech blízké, což potvrzuje jejich podobnost na základě hodnot zvolených indexů. Jedná se o trojice proteinů TX3TSS, TX1EXF, TX3GNU a TX1A8D, TX1SGK a TX1TC.

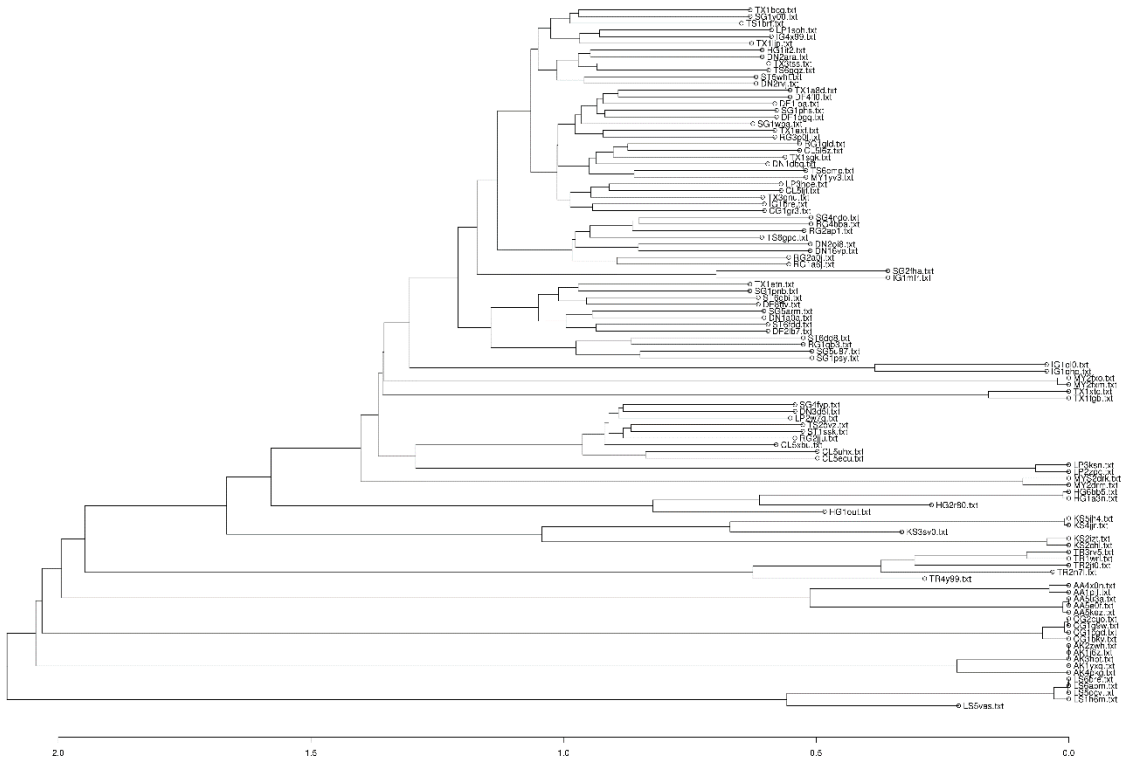
4.2 Analýza všech skupin a její interpretace

V grafech celkové analýzy se nacházejí všechny vzorky použité v analýze, tedy zástupci enzymů (α-amyláza značena AA, celuláza značena CL, lysozym značený LS), strukturních proteinů (kolagen značený CG, další strukturní proteiny značeny ST), transportních proteinů (hemoglobin značený HG, lipoprotein značený LP, další transportní proteiny značeny TS), nutričních proteinů (kasein značen KS, další nutriční proteiny značeny SG), kontrakčních a pohybových proteinů (aktin značený AK, myozin značený MY,

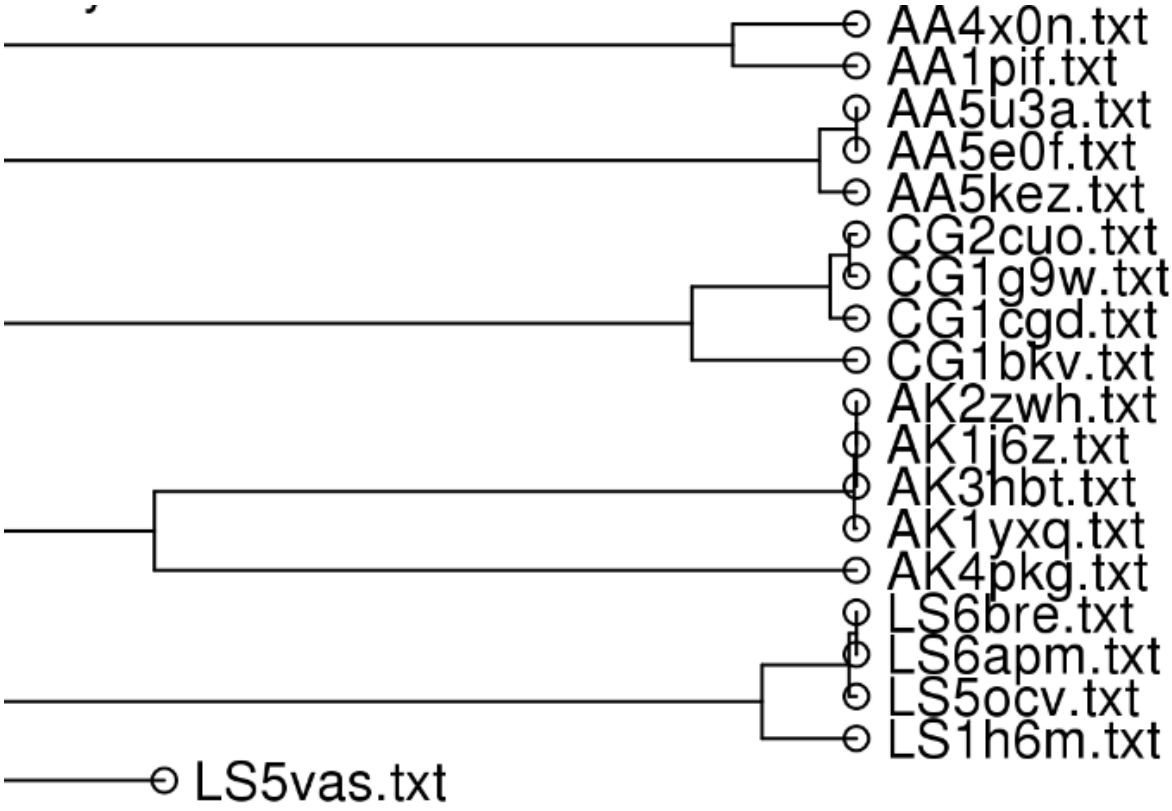
troponin značený TR), defenzivních proteinů (imunoglobulin značený IG, další proteiny s defenzivní funkcí značeny DF), regulačních proteinů (regulační proteiny, které se vážou na DNA značeny DN a další regulační proteiny značeny RG) a toxických proteinů (toxiny značeny TX).

Vzhledem k velkému počtu vzorků v jednom grafu bylo nutné následující dendrogram a některé další grafy vygenerovat ve velikosti A2. Z tohoto důvodu budou jednotlivé analyzované úseky z grafu vyříznuty a zvětšeny tak, aby byly lépe čitelné. Právě kvůli čitelnosti nebylo možné zvolit u vyříznutých obrázků stejnou velkost. Z téhož důvodu není v této části uveden graf hierarchického klastrování vytvořený z hodnot zvolených indexů. Stejné informace je možné vyčíst z grafu multidimenzionálního škálování, který je čitelný lépe.

4.2.1 Hierarchické shlukování

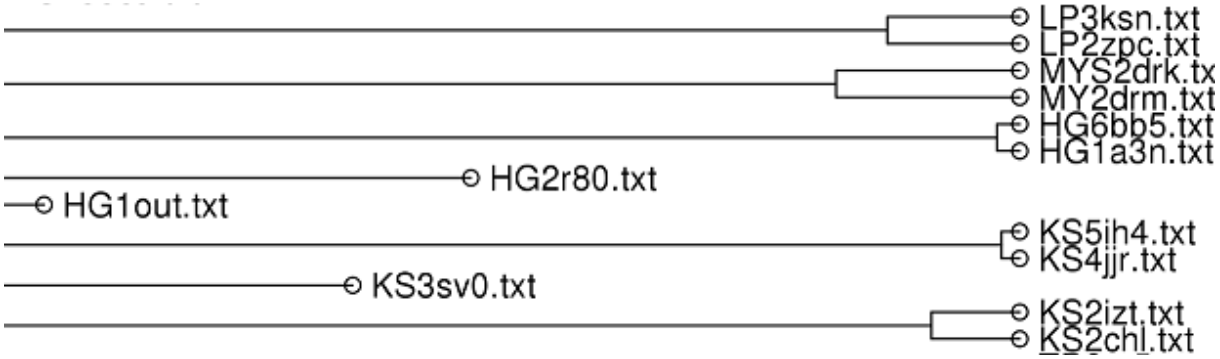


Obr. 42 Dendrogram (hierarchické shlukování)



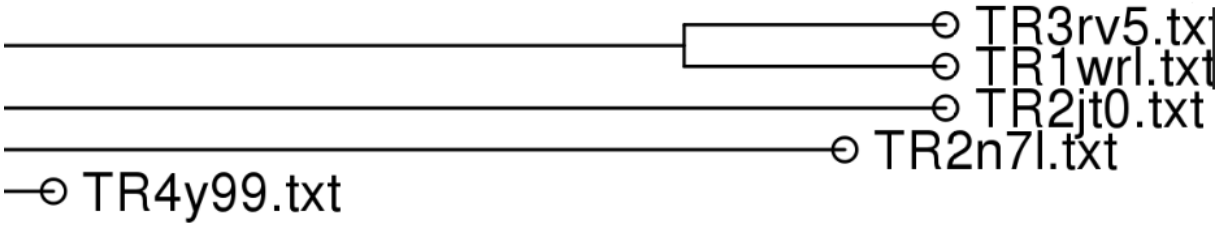
Obr. 43 Výsek z dendrogramu 1

Z dendrogramu vytvořeného pomocí metody hierarchického shlukování je možné vyvodit několik závěrů. Jedním z nich je specifičnost všech pěti vybraných vzorků enzymu lysozym, který je v grafu vykreslen na samostatné spodní větvi. To by mohlo vypovídat o výjimečnosti funkce lysozymu jak mezi ostatními proteiny, tak mezi samotnými enzymy, jak bylo zmíněno v interpretaci dendrogramu samostatné skupiny enzymů. Specifičnost katalytické funkce enzymů je možné dále potvrdit tím, že v grafu se v blízkosti lysozomu nachází všech pět vzorků dalšího enzymu, α-amylázy.

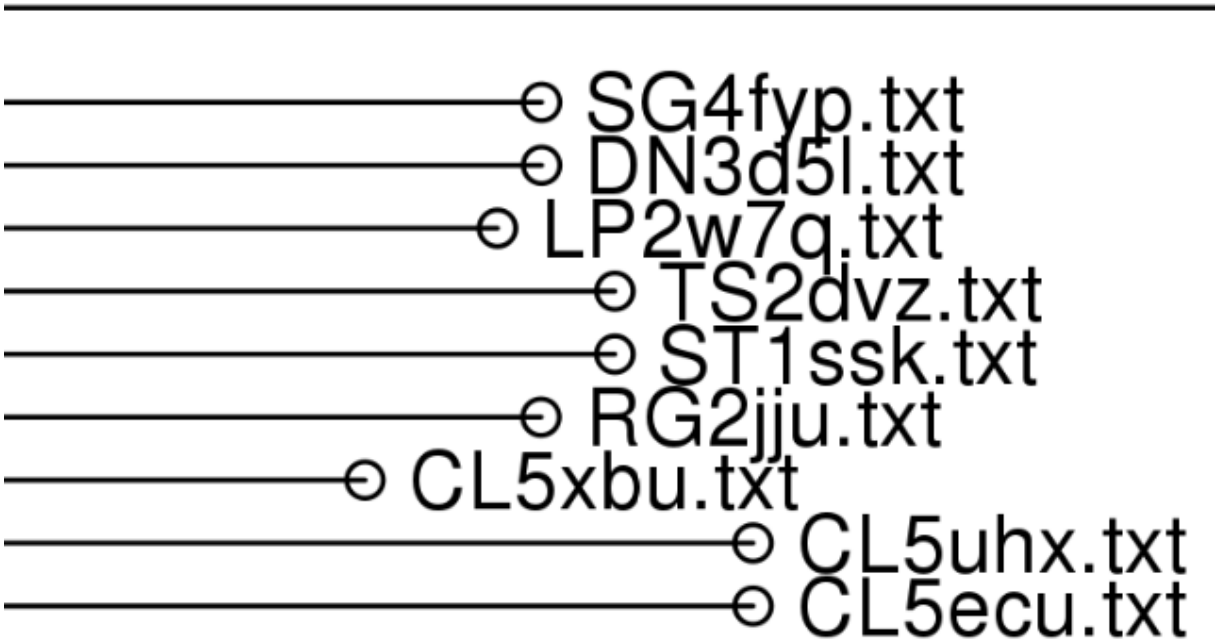


Obr. 44 Výsek z dendrogramu 2

V grafu je dále viditelné, že čtyři z pěti vzorků hemoglobinu rovněž tvoří ucelenou a samostatnou skupinu, což lze odůvodnit speciální funkcí hemoglobinu zároveň na sebe vázat a přenášet kyslík v organismu. Odlišnost hemoglobinu i to, že se vzorky vyskytují v jednom shluku, může být ovlivněno také jeho kvartérní strukturou. V blízkosti hemoglobinu se v grafu nachází další zástupci proteinů s transportní funkcí, lipoproteiny, lze tedy konstatovat to, že i transportní proteiny se nacházejí v poměrně blízkém shluku a vykazují tedy jasnou podobnost. Dalším proteinem, jehož vzorky se nacházejí v grafu ve shluku, je kasein, zástupce nutričních proteinů. Vzorky kaseinu se nacházely v bezprostřední blízkosti i v grafech skupiny nutričních proteinů, což vypovídá o výrazné podobnosti zvolených vzorků.

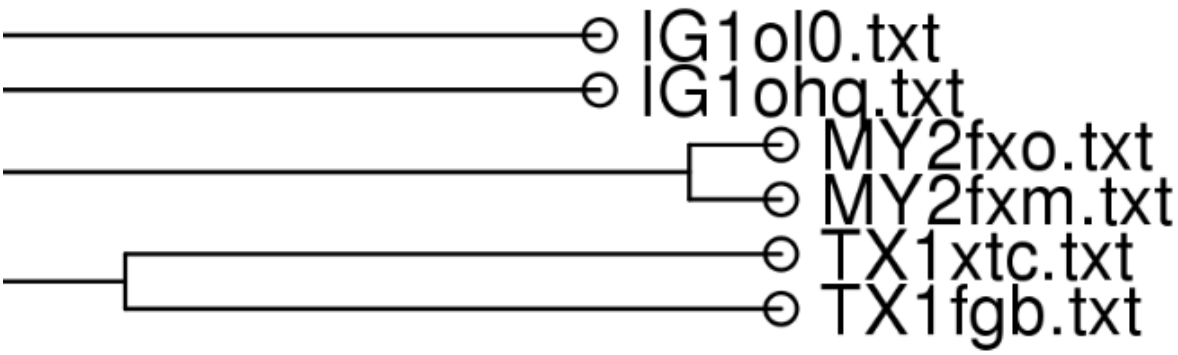


Stejný případ, jako je protein kasein, vidíme v grafu i u vzorků troponinu, který patří mezi nutriční a pohybové proteiny. Troponin byl stejně jako kasein výrazně odlišný i v rámci své skupiny. To lze interpretovat jako znak podobnosti vybraných vzorků.



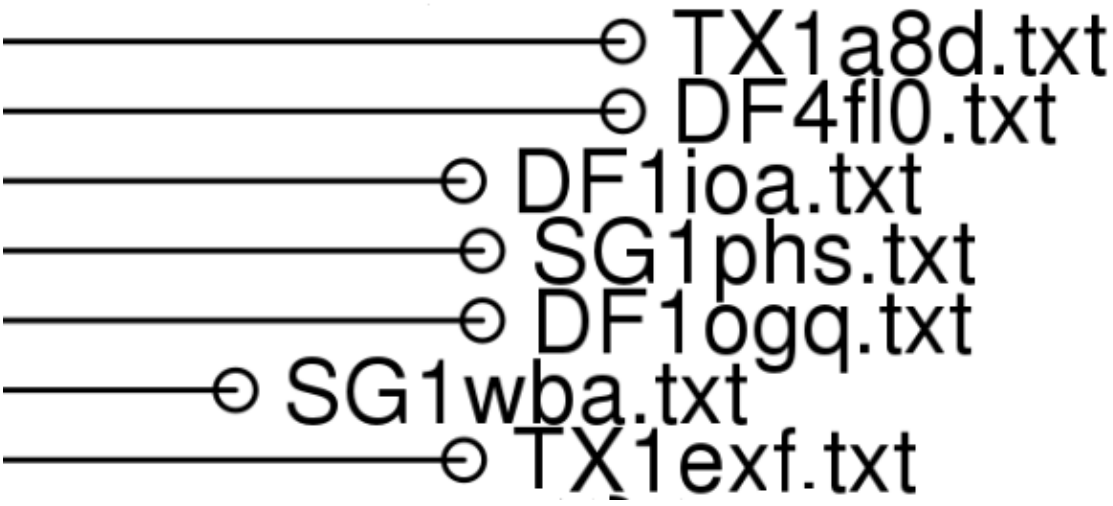
Obr. 45 Výsek z dendrogramu 3

Ve střední části grafu se nachází shluk proteinů s různými biologickými funkcemi. Stěžejní je však shluk tří z pěti vzorků posledního enzymu využitého v analýze, celulázy. Jedná se o stejné vzorky, které se vyskytovaly v trojici i v analýze enzymů jako skupiny, což podtrhuje podobnost těchto vzorků.



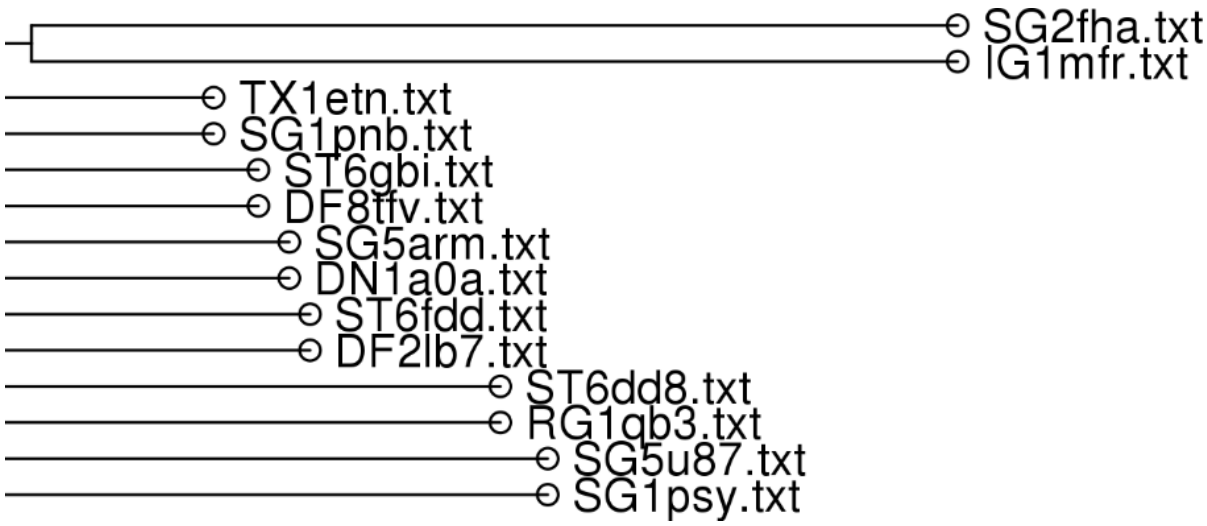
Obr. 46 Výsek z dendrogramu 4

V grafu můžeme dále nalézt shluk dvojic proteinů s odlišnými funkcemi. Jedná se o dva vzorky imunoglobulinu, tedy proteinu s defenzivní funkcí. Dále je to dvojice myozinu, proteinu s kontrakční a pohybovou funkcí. A v neposlední řadě se v tomto shluku nachází dvojice toxinů, tedy proteinů s toxickou funkcí. Blízkost proteinů s toxickou a defenzivní funkcí je možné interpretovat tak, že obranná a ochranná funkce proteinů vykazuje společné znaky.



Obr. 47 Výsek z dendrogramu 5

Shluky proteinů s toxickou a defenzivní funkcí je možné nalézt v grafu na více místech, což podporuje tvrzení o jejich možné podobnosti.



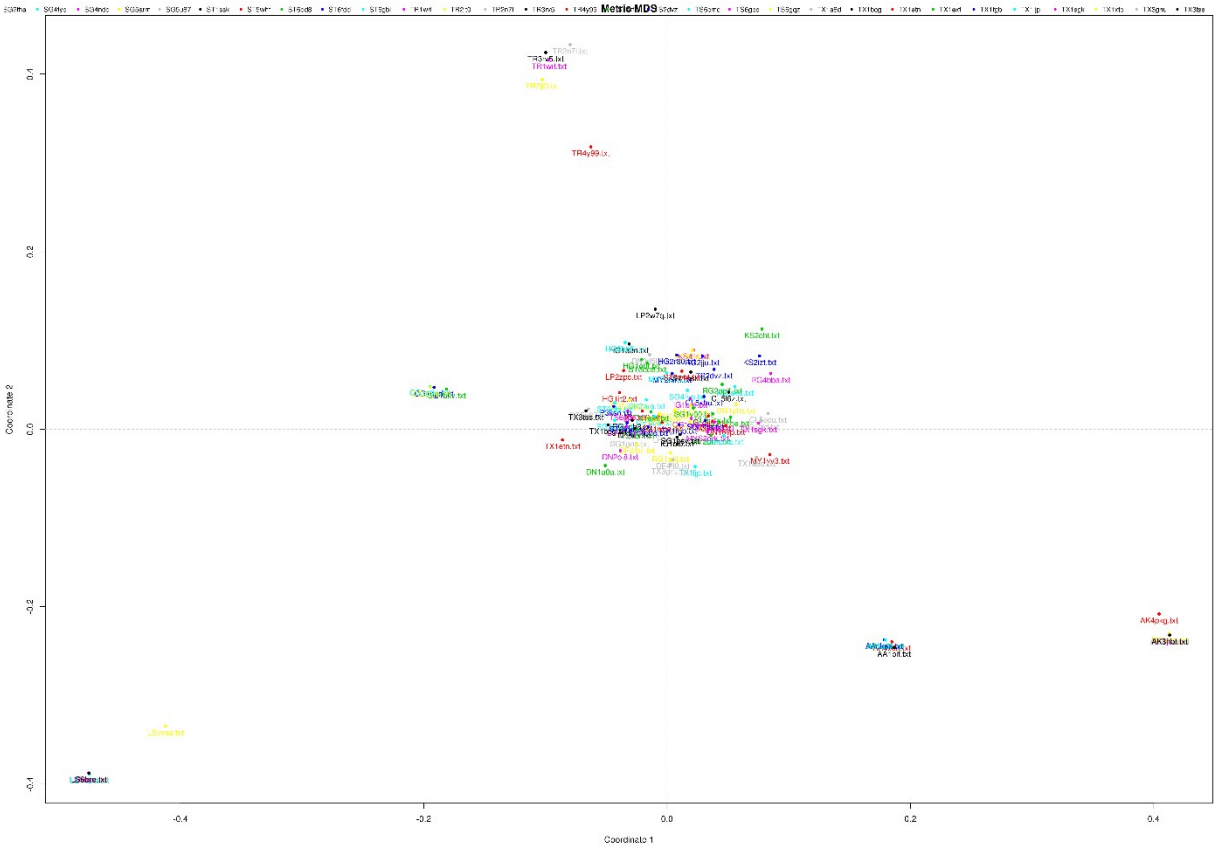
Obr. 48 Výsek z dendrogramu 6

V další části grafu můžeme pozorovat shluk skládající se převážně z regulačních a defenzivních proteinů, což by rovněž mohlo vypovídat o podobnosti těchto biologických funkcí.

Obr. 49 Výsek z dendrogramu 7

Podobný shluk regulačních a defenzivních vzorků proteinů můžeme nalézt v grafu opakovaně, což podporuje výše uvedené tvrzení o možné podobnosti těchto dvou funkcí.

4.2.2 Multidimenzionální škálování



Obr. 50 MDS (vícerozměrové škálování)

Graf multidimenzionálního škálování je vzhledem k původnímu barevnému nastavení málo čitelný i ve zvětšené podobě jednotlivých úseků. Z toho důvodu je možné z grafu vyříznout pouze ty části, kde se vzorky příliš nepřekrývají.

Obr. 51 Výřez z MDS 1

V pravé spodní části grafu se nachází vzorky enzymu α-amyláza, které se v grafu částečné překrývají. Vhledem tomu, že jsou v grafu umístěny v samostatné skupině,

můžeme konstatovat, že vzorky jsou si velmi podobné a zároveň, že jejich katalytická funkce je signifikantně specifická.

Obr. 52 Výřez z MDS 2

V levé horní části grafu můžeme vidět shluk vzorků troponinu, který patří mezi kontrakční a pohybové proteiny. Tato skutečnost opětovně dokazuje specifičnost tohoto proteinu, a to jak v rámci skupiny, do níž funkčně náleží, tak mezi všemi vzorky použitých proteinů.

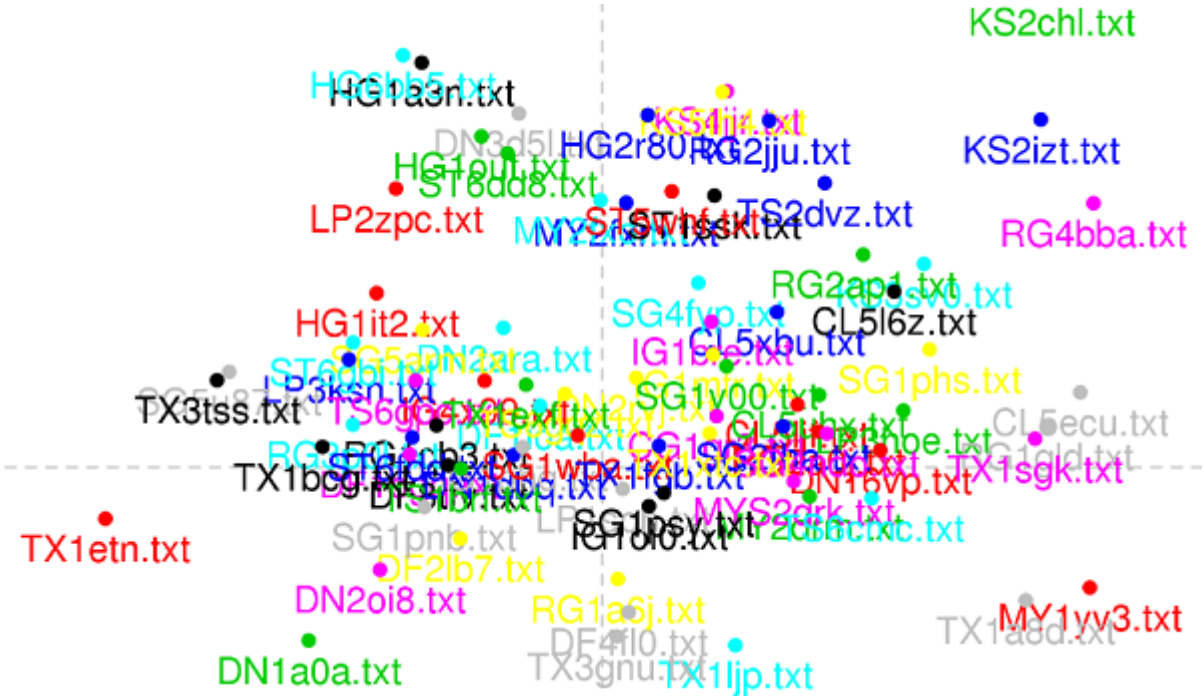
Obr. 53 Výřez z MDS 3

Výraznou odlišnost proteinů s kontrakční a pohybovou funkcí dokazuje také shluk vzorků aktinu, který se nachází v pravém dolním rohu grafu. Skupina vzorků aktinu byla

výrazně oddělena i v grafech obsahujících pouze kontrakční a pohybové proteiny, což znovu potvrzuje, že je specifický, jak je uvedeno v předchozí interpretaci.

Obr. 54 Výřez z MDS 4

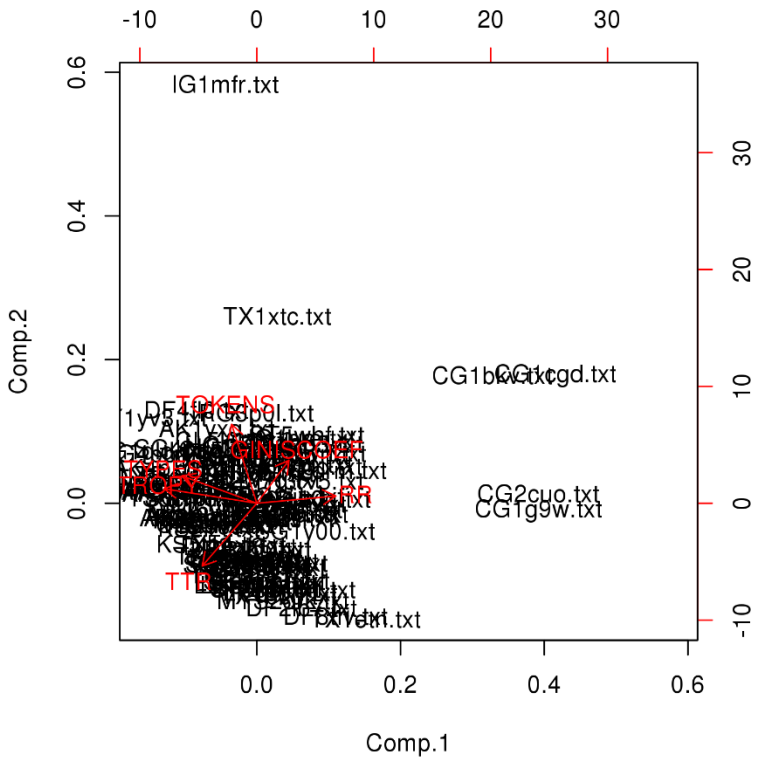
V dolní pravé části grafu se nachází vzorky enzymu lysozym. Tato skupina je opět ojedinělá, což znovu potvrzuje specifičnost tohoto proteinu mezi všemi ostatními a můžeme to přikládat charakteristické katalytické funkci enzymů.



Obr. 55 Výřez z MDS 5

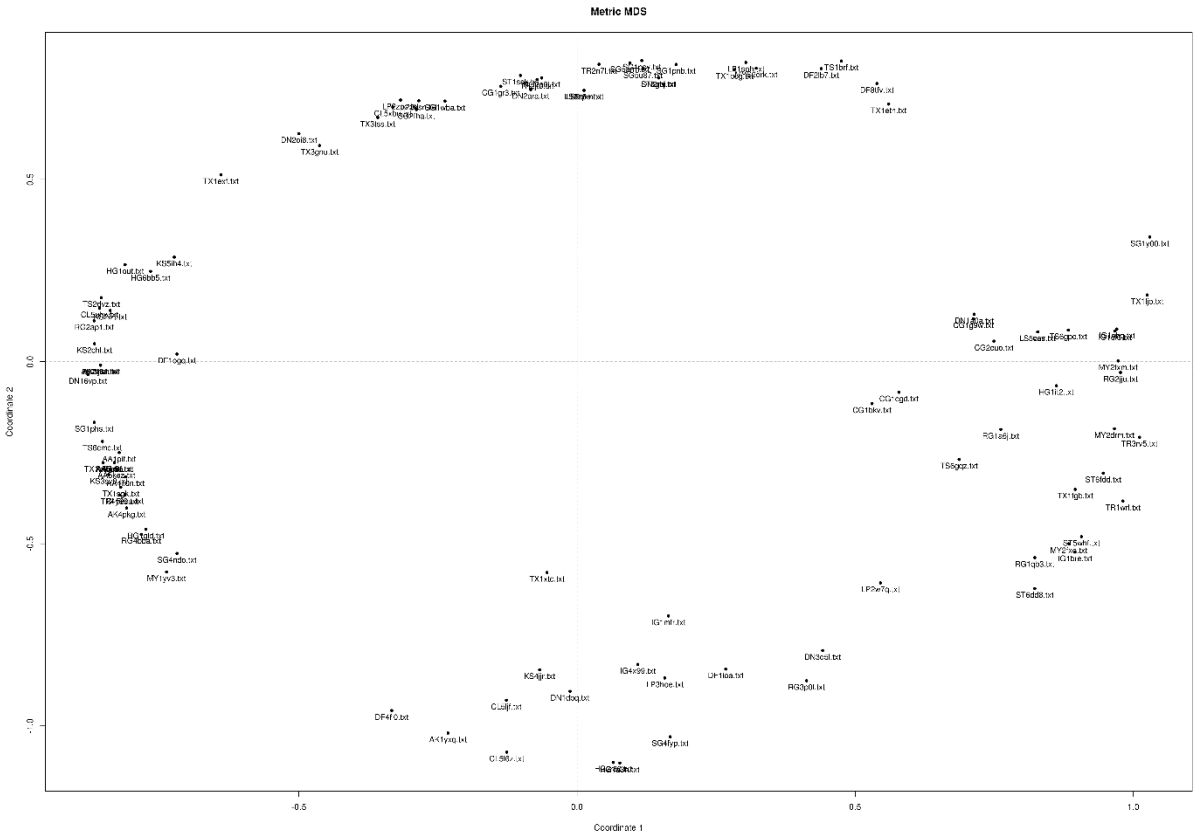
Ve střední části grafu multidimenzionálního škálování můžeme pozorovat velký shluk různých vzorků proteinů, které mají různé funkce. V levém horním rohu grafu vidíme

skupinu vzorků hemoglobinu, což podporuje předchozí tvrzení o tom, že tento protein se vymezuje nejen v rámci skupiny transportních proteinů, ale i mezi všemi vzorky. Toto pozorování bylo již výše odůvodněno speciální funkcí hemoglobinu zároveň na sebe vázat a přenášet kyslík v organismu. Znovu můžeme také vidět shluky proteinů s toxickou a defenzivní funkcí, což podporuje tvrzení o jejich možné podobnosti. Stejný vztah nastává u proteinů s regulační a defenzivní funkcí.



Obr. 56 PCA (analýza hlavních komponent)

Z grafu analýzy hlavních komponent, který využívá zvolených indexů je patrné, že mezi vzorky se vyskytuje několik proteinů, které se vydělují z hlavní skupiny. Konkrétně se jedná například o vzorek imunoglobulinu IG1MRF, který byl výrazně odlišný od ostatních imunoglobulinů, stejně jako od ostatních proteinů s defenzivní funkcí. Další vzorek vyčleněný ze skupiny je toxin TX1XTC, který se od ostatních vzorků výrazně liší počtem tokenů. V grafu rovněž můžeme pozorovat skupinu čtyř vzorků kolagenu, z nichž dva a dva se částečně překrývají. První dvojice se podobá hodnotou Giniho koeficientu, zatímco druhá má podobnou hodnotu indexu RR.



Obr. 57 MDS (vícerozměrové škálování)

V grafu multidimenzionálního škálování vidíme, že jsou vzorky rozmístěny v blízkosti okrajů, což lze interpretovat tak, že napříč hodnotami naměřenými u všech použitých vzorků neexistuje výrazný spojovací faktor. Můžeme pozorovat některé vzorky, které se přibližují ke středu grafu více než ostatní.

Obr. 58 Výřez z MDS 6

Jedná se o vzorky, které byly výrazně jiné než ostatní proteiny již v předchozím grafu. Toto pozorování podporuje tvrzení, že vzorky IG1MRF a TX1XTC jsou odlišné od skupiny všech vzorků i od vzorků skupin, do kterých funkčně náleží.

Závěr

Proteinová lingvistika umožňuje propojení biologie a lingvistiky, což poskytuje nové možnosti nejen této pomezní disciplíně, ale i oběma vědám, ze kterých vznikla. Těchto možností je využíváno například při popisu gramatických pravidel pro báze DNA, molekul RNA a podobně. Studium proteinů pomocí metod lingvistiky však není v biolingvistické a bioinformatické komunitě příliš časté. Proteinová lingvistika promítá jazykové plány na hierarchii molekulární biologie tak, že sekvence odpovídá lexikální rovině, struktura rovině syntaktické, funkce rovině sémantické a role pak rovině pragmatické. Tato experimentální magisterská práce se zabývala kvantitativní analýzou typologie proteinů klasifikovaných na základě jejich funkce, což odpovídá sémantické rovině textu.

V úvodu byla vytyčena výzkumná otázka, zda a jakým způsobem je možné provést kvantitativní lingvistickou analýzu proteinů klasifikovaných podle jejich biologické funkce a popsat tak charakteristické znaky jejich genetického textu pomocí programu QUITA (Quantitative Text Index Analyzer), modelu Bag of Words a vybraných kvantitativně lingvistických indexů. Protože účelem práce byla kvantitativní lingvistická analýza, byly tomu přizpůsobeny teoretické kapitoly pojednávající o proteinu, které slouží jako úvod do problematiky tohoto tématu. Kapitoly čerpaly z vhodně zvolené proteomické literatury, přičemž chemické a biologické procesy, kterými proteiny procházejí, byly popsány obecně s případným doplněním informací v poznámkovém aparátu, neboť práce neusilovala o hloubkový popis na úrovni makromolekulární biologie. Zpětně nahlíženo obsahuje práce velké množství teorie, která je popsána příliš podrobně a v následné analýze není plně využita. Celkově však kombinace teoretické a praktické části umožňuje získat základní orientaci v problematice.

Práce byla rozdělena do čtyř kapitol, z nichž tři se věnovaly teoretickému uchopení a popisu proteinu, který je základní biologickou jednotkou. Čtvrtá kapitola pak byla věnována samotné kvantitativní analýze. V první kapitole byly popsány základní informace o proteinu, jeho vlastnostech, složení, skládání, tvaru jeho molekul atd. Dále tato kapitola popsala aminokyseliny a peptidy.

Druhá kapitola sloužila jako podkladový materiál pro samotnou analýzu, protože popsala klasifikaci proteinů podle publikace *Fundamentals of Biochemistry* (2005) od J. L.

Jaina. Na základě této klasifikace byly proteiny vyděleny do čtyř kategorií, a to podle zdroje molekuly proteinu, podle tvaru molekuly proteinu, podle složení a rozpustnosti nebo na základě jejich biologické funkce. Pro tuto práci byla využita klasifikace dle biologické funkce, protože analýza využila právě proteiny rozdělené podle biologické funkce.

Třetí kapitola doplnila předchozí teorii o popis struktury proteinu, která má částečný vliv na jeho funkce. Tato kapitola, stejně jako v širším měřítku i tato diplomová práce, poskytla podklad pro navazující akademickou činnost, týkající se popisu syntaxe genetického textu proteinů, tedy právě struktury proteinů. Kapitola obsahuje základní informace o primární, sekundární, terciární a kvartérní struktuře.

Čtvrtá kapitola je tvořena analytickou částí, která popsala metodologické uchopení analýzy, vybrané vzorky proteinů klasifikovaných do skupin podle biologické funkce, jejich charakterizaci na základě modelu Bag of Words a vybraných indexů kvantitativní lingvistiky, a v neposlední řadě celkovou analýzu všech výše vytyčených skupin proteinů. Celá tato kapitola byla odpovědí na výzkumnou otázku vytyčenou výše, přičemž tato část závěru poskytuje celkové shrnutí analytické kapitoly.

Při provádění experimentu v softwaru QUITA byla provedena mnohá pozorování, než bylo zvoleno konečné nastavení. Klíčovým krokem byla tokenizace, protože při prvních pokusech v nainstalovaném softwaru nebyla možnost zvolit tokenizér určený speciálně pro aminokyseliny. V online verzi softwaru tato možnost již existuje, což umožnilo dosáhnout přesnějších výsledků měření. Další klíčovou okolností, která měla vliv na přesnější měření, bylo nastavení kosinovy vzdálenosti, která eliminuje odchylky způsobené různou délkou analyzovaných vzorků. Pro porovnávání stringů byly použity trigramy. Volba n-gramů s touto délkou závisela na několika faktorech. Trigramy byly zvoleny jako vhodné na základě pozorování při tvorbě samotného experimentu. Dalším důvodem pro volbu trigramů byla velikost abecedy, která se skládá z 20 písmen označujících základní aminokyseliny, které se řadí do sekvence proteinu. Pokud zohledníme možnosti kombinací utvořených 20 písmeny v sekvenci o délce desítek až stovek dílčích jednotek, dojdeme k závěru, že trigramy oproti bigramům ve vyšší míře diverzifikují jednotlivé vzorky. Výskyt trigramů je zároveň dostatečně častý na to, aby reflektoval společné znaky vzorků. Vstupní data byla nejdříve tokenizována a poté z nich byly vytvořeny trigramy. Z takto upravených dat byly vygenerovány grafy hierarchického shlukování, víceúrovňového škálování a analýzy hlavních komponent. Stejné typy grafů pak byly vytvořeny z tokenizovaných trigramů,

z nichž software QUITA vypočítal indexy. V kvantitativní analýze je využito šesti indexů. Jedná se o tokeny, typy, TTR, entropii, Giniho koeficient a RR. Indexy byly zvoleny při tvorbě experimentu, protože jejich hodnoty vykazovaly znatelné rozdíly.

Celkově je v analýze použito 105 vzorků pro osm skupin proteinů klasifikovaných podle biologické funkce. Analýza tedy využívá 15 vzorků enzymů, transportních proteinů, nutričních proteinů, kontrakčních a pohybových proteinů a regulačních proteinů, a 10 vzorků pro strukturní, defenzivní a toxické proteiny. Vzorky proteinů, které byly použity v analýze, pocházejí z databáze Research Collaboratory for Structural Bioinformatics Protein Data Bank, dále RCSB PDB. Tato data banka proteinů byla zvolena kvůli své organizaci a přehlednosti anotací, které při popisu vzorků využívá. Vzorky byly vybrány ve snaze reflektovat podskupiny výše vytyčených skupin proteinů, rozdělených podle biologických funkcí, které plní. Počet vzorků každé podskupiny se snažil brát v potaz velikost jednotlivých skupin, stejně jako jejich rozmanitost, byl však omezen možností výběru použité data banky i dalších existujících databank. Při výběru vzorků bylo bráno v úvahu několik faktorů, jejich velikost, tedy počet aminokyselinových reziduí, klasifikace podle jejich biologické funkce a další klasifikace, které využívá RCSB PDB pro popis vzorků. Při popisu každého ze 110 vzorků je využita část původní anotace RCSB PDB. Jedná se o kódové označení vzorku proteinu, originální název vzorku, jeho další klasifikace, organismus, ze kterého pochází, počet jeho aminokyselinových reziduí, případný počet mutací a biologické role, které plní. Tato anotace dále sloužila pro snazší orientaci mezi vzorky a rovněž využívala části anotace jako distinktivních rysů při porovnání vzorků. Pokud to bylo možné a příslušné informace byly známy, byly dále vzorky zařazeny do klasifikace, která byla vytyčena ve druhé kapitole této práce. Kvůli transparentnosti práce zachovala původní kódové označení vzorků a pro přehlednost byla přidána zkratka názvu podskupiny proteinu. Původní název kódového označení byl použit také z důvodu zachování datové cesty, jejímž prostřednictvím je možné vzorky nalézt. Z téhož důvodu nejsou citace vzorků v práci uvedeny.

Každá z výše vytyčených skupin proteinů byla popsána v samostatné podkapitole, která uvedla seznam vzorků a jejich anotací. Poté byly pro každou podkapitolu uvedeny grafy, které byly následně interpretovány. Pro každou podskupinu práce uvedla grafy hierarchického shlukování a multidimenzionálního škálování, které byly vytvořeny v modelu Bag of Words. Poté následovaly grafy analýzy hlavních komponent, hierarchického klastrování a multidimenzionálního škálování, které vykreslily hodnoty zvolených indexů.

Posledním krokem byla celková analýza a její interpretace. V této části analýzy byly použity všechny vzorky. Z toho důvodu byly na začátku této podkapitoly vypsány všechny skupiny proteinů a jejich dílčí vzorky spolu se zkratkami, kterými byly označeny. Vzhledem k velkému počtu vzorků v jednom grafu bylo nutné dendrogram celkové analýzy a některé další grafy vygenerovat ve velikosti A2. Z tohoto důvodu byly jednotlivé analyzované úseky z grafu vyříznuty a zvětšeny tak, aby byly lépe čitelné. Kvůli čitelnosti nebylo u vyříznutých obrázků možné zvolit stejnou velkost. Kvůli čitelnosti v této části není uveden graf hierarchického klastrování vytvořený z hodnot zvolených indexů. Stejné informace je možné vyčíst z grafu multidimenzionálního škálování, který je lépe čitelný.

Z celkové analýzy bylo možné vyvodit několik závěrů. Jedním z nich bylo potvrzení pracovní hypotézy o tom, že katalytická funkce enzymů je natolik specifická, že se v analýze projeví. Dále bylo z analýzy možné vyvodit signifikantní podobnost mezi transportními proteiny a výjimečnost funkce hemoglobinu. Dále se v analýze několikrát projevila možná souvislost obranné a ochranné funkce proteinů, protože proteiny s toxickou a defenzivní funkcí vykazovaly podobné znaky. Další podobnosti bylo možné nalézt mezi regulačními a defenzivními proteiny, což bylo taktéž možné interpretovat jako podobnost těchto biologických funkcí. Celkově lze říci, že pomocí nástrojů softwaru QUITA, modelu Bag of Words a vybraných kvantitativně lingvistických indexů bylo možné provést analýzu proteinů klasifikovaných podle jejich biologické funkce a popsat tak charakteristické znaky jejich genetického textu.

Literatura a prameny

„Amide“ Oxford dictionaries, (online). Dostupné z: < https://en.oxforddictionaries.com/definition/amide> (21. 8. 2018)

„Cellulase“ Worthington Biochemical Corporation, (online). Dostupné z: < http://www.worthington-biochem.com/cel/default.html> (21. 8. 2018)

„Dalton“ Oxford Dictionaries, (online). Dostupné z: <https://en.oxforddictionaries.com/definition/dalton> (21. 8. 2018)

„Electrophoresis“ Britannica, (online). Dostupné z: < https://www.britannica.com/science/electrophoresis> (21. 8. 2018)

„Hydration of proteins“ Britannica, (online). Dostupné z: <https://www.britannica.com/science/protein/Hydration-of-proteins#ref593785> (21. 8. 2018)

„Myosins“ U.S. National Library of Medicine, (online). Dostupné z: <https://ghr.nlm.nih.gov/primer/genefamily/myosins> (21. 8. 2018)

„Peptids“ Oxford Dictionaries, (online). Dostupné z: < https://en.oxforddictionaries.com/definition/peptide> (21. 8. 2018)

„Protein“ Britannica, (online). Dostupné z: <https://www.britannica.com/science/protein#ref593739> (21. 8. 2018)

„Škálování“ Sociologická encyklopedie, (online). Dostupné z: < https://encyklopedie.soc.cas.cz/w/%C5%A0k%C3%A1lov%C3%A1n%C3%AD> (21. 8. 2018)

Abdulla, Ali. 2012. „*Bachelor Thesis in Textual Similarity.“* Kongens Lyngby: Technical University of Denmark.

Alberts, Bruce. 2007. *Molecular Biology of the Cell*. Oxford: Garland Science.

Altmann, Gabriel. Čech, Radek.,Ján Mačutek. 2017. *„ENTROPIE“*. In: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), CzechEncy - Nový encyklopedický slovník češtiny, (online). Dostupné z: < https://www.czechency.org/slovnik/ENTROPIE > (21. 8. 2018)

Bragulla, Herman. Dominique Homberger. 2009. „Structure and functions of keratin proteins in simple, stratified, keratinized and cornified epithelia“, 214 (4), 516-59.

Bragulla, Herman. Dominique Homberger. 2009. „Structure and functions of keratin proteins in simple, stratified, keratinized and cornified epithelia“, 214 (4), 516-59.

Cvrček, Václav. 2017 „TOKEN“ In: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), CzechEncy - Nový encyklopedický slovník češtiny, (online). Dostupné z: < https://www.czechency.org/slovnik/TOKEN> (21. 8. 2018)

Cvrček, Václav. 2017 „TYPE-TOKEN“ In: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), CzechEncy - Nový encyklopedický slovník češtiny, (online). Dostupné z: < https://www.czechency.org/slovnik/TYPE-TOKEN> (21. 8. 2018)

Cvrček, Václav. 2017 „TYPE-TOKEN“ In: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), CzechEncy - Nový encyklopedický slovník češtiny, (online). Dostupné z: < https://www.czechency.org/slovnik/TYPE-TOKEN> (21. 8. 2018)

Čech, Radek. 2014. Metody kvantitativní analýzy (nejen) básnických textů. Olomouc: Univerzita Palackého v Olomouci.

Garret, Reginald. Charles Grisham, Charles. 2012. *Biochemistry.* Boston: Cengage Learning.

Jian, J.L., Sunjay Jain, Nitin Jain. 2005. *Fundamentals of Biochemistry.* New Dehli: S.Chand & Company Ltd.

Johnson, Stephen. 1967. *„Hierarchical clustering schemes,“* Psychometrika 32, 241.

Malcolm Dixon, Edwin C. Webb. 1964. Enzymes. London: Academic Press.

Markoš, Antonín. 2007. *Biosémiotika 1.*Olmouc: Univerzita Palackého v Olomouci.

Matlach, Vladimír. 2014. „*Kvantitativně lingvistický software.“* Olomouc: Univerzita Palackého v Olomouci.

McKee, Trudy. Jamese McKee. 2015. *Biochemistry: The Molecular Basis of Life.* Oxford: Oxford University Press.

Montalbetti, Christian. Falque, Virginie. 2005. *„Amide bond Formation and Peptide Coupling.“* Tetrahedron, 46, 10819-11046.

Rosenbloom, J. 1984. **„***Elastin: relation of protein and gene structure to disease***,“** Laboratory Investigation, 51(6), 605-23.

Searls, D. B. 2001. *„Reading the book of life.“* Bioinformatics, 17, 579–580.

Shoulders, MD, RT Raines. 2009. *„Colagen structure and stability, “* Annual review of Biochemistry, 78, 929-58.

Svoboda, Jan. 2005. *Organická chemie I.* Praha: Vysoká škola chemicko-technologická v Praze.

Tonhauserová, Zuzana. 2013. *„Metoda hlavních komponent a její aplikace.“* Olomouc: Univerzita Palackého v Olomouci.

Twymynn, Richard. 1999. *Advanced Molecular Biology.* Oxford: BIOS Scientific Publishers Limited.

Whitford, David. 2005. *Proteins: Structure and Function*. Chichester: John Wiley & Sons Ltd.