Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Identifikácia proteínových sektorov ako evolučných jednotiek bielkovín

Bakalárska práca

2022
Martina Babinská

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Identifikácia proteínových sektorov ako evolučných jednotiek bielkovín

Bakalárska práca

Študijný program:	Bioinformatika
Študijný odbor:	Informatika a Biológia
Školiace pracovisko:	Katedra informatiky
Školiteľ:	prof. RNDr. Ľubomír Tomáška, DrSc.
Konzultant:	doc. Mgr. Bronislava Brejová, PhD.

Bratislava, 2022
Martina Babinská

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta:	Martina Babinská
Študijný program:	bioinformatika (Medziodborové štúdium, bakalársky I. st., denná forma)
Študijné odbory:	informatika biológia
Typ záverečnej práce:	bakalárska
Jazyk záverečnej práce:	slovenský
Sekundárny jazyk:	anglický

Názov: Identifikácia proteínových sektorov ako evolučných jednotiek bielkovín
Identification of protein sectors as evolutionary units of protein structures

Anotácia: Charakterizácia trojrozmernej (3D) štruktúry proteínov je dôležitá pre porozumenie ich funkcie. Samotná 3D štruktúra však neumožňuje identifikáciu všetkých pozícii v sekvencii aminokyselín proteínu, ktoré sa podieľajú na realizácii jeho biochemických funkcií. Tieto pozície vytvárajú tzv. proteínové sektory, ktorých identifikácia je založená na kombinácii bioinformatických a biochemických metód. Cieľom bakalárskej práce bude porovnať bioinformatické nástroje pre identifikáciu proteínových sektorov a využiť ich na analýzu domén vybraných proteínov.

Vedúci:	prof. RNDr. Ľubomír Tomáška, DrSc.
Konzultant:	doc. Mgr. Bronislava Brejová, PhD.
Katedra:	FMFI.KI - Katedra informatiky
Vedúci katedry:	prof. RNDr. Martin Škoviera, PhD.

Dátum zadania: 15.10.2021

Dátum schválenia: 15.10.2021 doc. Mgr. Bronislava Brejová, PhD.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie: Chcem sa poďakovať môjmu školiteľovi prof. RNDr. Ľubomírovi Tomáškovi, DrSc. za odbornú pomoc, užitočné rady a poznámky, trpezlivosť a množstvo konzultácií počas písania práce. Vďaka patrí aj mojej konzultantke doc. Mgr. Bronislave Brejovej, PhD. za ochotu a veľkú pomoc pri riešení problémov a navrhovaní nových zaujímavých cieľov tejto práce. Ďakujem aj prof. RNDr. Jozefovi Nosekovi, DrSc. za poskytnuté dáta.

Abstrakt

Funkcia proteínu je podmienená jeho trojrozmerným usporiadaním. Evolučné adaptácie na zmeny prostredia však môžu ovplyvniť jeho biochemické vlastnosti. To, ktoré pozície v proteíne sa na týchto adaptáciách podieľajú, len zo samotnej trojrozmernej štruktúry proteínu zistiť nedokážeme. Je preto dôležité skúmať samotné zloženie proteínu a snažiť sa zistiť, aké iné, neintuitívne subštruktúry tieto zmeny podmieňujú. V našej práci sme na zamerali na hľadanie takzvaných proteínových sektorov, ktoré sú tvorené spolupracujúcimi a vzájomne sa ovplyvňujúcimi pozíciami. Na ich štúdium sme použili dva bioinformatické nástroje: (1) štatistickú analýzu prepojení a (2) analýzu GREMLIN (z angl. *Generative regularized models of proteins*) na predikciu kontaktov aminokyselín v trojrozmernej štruktúre proteínu. Metódy sme aplikovali na proteín poly (ADP-ribóza) polymeráza, ktorý zohráva dôležitú úlohu pre život bunky, pričom bol len nedávno objavený u kvasiniek. Výsledky zahŕňali identifikáciu dvojíc pozícií, ktoré na seba vzájomne pôsobia, následnú lokalizáciu a určenie rozmiestnenia proteínových sektorov v rámci pozorovaného proteínu.

Kľúčové slová: proteín, konzervovanosť, koevolúcia, proteínový sektor

Abstract

The function of a protein is mediated by its three-dimensional arrangement. Evolutionary adaptations to environmental changes can influence its biochemical properties. Nevertheless, we cannot determine which positions within the protein are responsible for these adaptations solely from the three-dimensional structure of the protein. It is therefore important to examine the composition of the protein and to try to find out what other, non-intuitive substructures are involved in these changes. In our work, we focused on identification of the so-called protein sectors, which are formed by cooperating and interacting positions. We used two bioinformatics tools to study them: (1) statistical coupling analysis and (2) GREMLIN (*Generative regularized models of proteins*) analysis to predict physical contacts between the protein residues in its three-dimensional structure. We applied these methods to the protein poly (ADP-ribose) polymerase, which plays an important role in cell life and was only recently discovered in yeasts. Our results included identification of pairs of positions that interact with each other and consequent localization and distribution of protein sectors within the protein.

Keywords: protein, conservation, coevolution, protein sector

Obsah

Úvod

Poznatky o priestorovom uložení proteínu sú veľmi užitočné, napovedajú nám totiž o jeho funkcii a vlastnostiach. S predikciou štruktúr proteínov sa začalo už v roku 1951, kedy Linus Pauling, Robert Corey a Herman Branson predpovedali prvé priestorové usporiadanie proteínu. Dnes už poznáme priestorové štruktúry mnohých proteínov a je viacero spôsobov ako ich určiť. Prvým sú rôzne laboratórne experimentálne techniky, ktoré však môžu byť časovo aj finančne náročné. Alternatívnou cestou je použitie bioinformatických nástrojov, ktoré sú založené na umelej inteligencii a strojovom učení a dnes sú na vysokej úrovni [34]. Priamy súvis medzi funkciou proteínu a jeho priestorovým usporiadaním je síce empiricky podložený, avšak proteíny nie sú vo všetkých podmienkach nemenné. V rôznych prostrediach môžu svoj tvar, dokonca aj správanie meniť a prispôsobovať, a teda toto ich prepojenie nie je úplne jednoznačné.

Vieme nejakým spôsobom zistiť, čím sú podmienené biochemické vlastnosti proteínov? Je možné nájsť pozície proteínu, vplývajúce na tieto neprebádané vlastnosti a správanie? Podobne, ako pri stanovovaní priestorového usporiadania proteínu, je možné hľadať takéto pozície experimentálne. Tento prístup však opäť má svoje obmedzenia a pre dlhý proteín by mohlo byť veľmi náročné a zdĺhavé zisťovať, ktorá konkrétna pozícia zapríčiňuje zmenu správania sa proteínu - napríklad vplyvom zmeny prostredia. V súčasnosti však existujú bioinformatické nástroje na hľadanie neintuitívnych pozícií, ktoré by sa potenciálne mohli podieľať na nejakej z vlastností proteínu. Poznatky z takýchto analýz môžu byť užitočné z toho dôvodu, že potenciálne obrovské množstvo skúšaní pri hľadaní nejakej zaujímavej pozície, prípadne skupiny pozícií, sa výrazne vymedzí. Medzi bioinformatické nástroje, ktoré sa takéto účely používajú, patrí aj štatistická analýza prepojení (z angl. *Statistical Coupling Analysis*) a analýza GREMLIN (z angl. *Generative Regularized Models of proteins*).

V našej práci sme oba nástroje preštudovali a porovnali z hľadiska ich fungovania a ich cieľa. Následne sme oba použili pre analýzu proteínu poly (ADP-ribóza) polymeráza (PARP). V rámci analýzy sme sa venovali hľadaniu neintuitívnych subštruktúr v tomto proteíne, ktoré by mohli mať významnú rolu v jeho fungovaní. Spomínané metódy sme použili pre rôzne vstupy, aby sme získané výsledky dokázali medzi sebou porovnávať. Proteín PARP sme ako model vybrali preto, že bol nedávno objavený u kvasiniek, u ktorých je jeho skutočná funkcia a význam správania sa stále

neprebádané a nové poznatky by mohli byť nápomocné pri ich ďalšom štúdiu aj s možnými biomedicínskymi implikáciami. Základným vlastnostiam proteínov, princípom ich bioinformatického študovania, ako aj pozorovanému proteínu PARP sa venujeme v prvej kapitole. V druhej kapitole popisujeme priebeh výpočtov jednotlivých nástrojov a v tretej rozoberáme a hodnotíme získané výsledky z oboch analýz.

Kapitola 1

Proteíny

V tejto kapitole sa zameriame na proteíny. Stručne popíšeme ich štruktúru, vlastnosti a funkcie. Objasníme aj význam konzervovanosti, koevolúcie a proteínových sektorov. V závere kapitoly opíšeme charakteristiky proteínu PARP (poly (ADP-ribóza) polymeráza), ktorý bude v ďalších častiach práce slúžiť na aplikáciu bioinformatických nástrojov.

1.1 Proteíny: Známe-neznáme molekuly

Proteíny charakterizujeme ako organické biomakromolekuly, ktorých základnými jednotkami sú aminokyseliny, pričom poznáme dvadsať druhov základných proteínotvorných aminokyselín. Bielkoviny predstavujú vyše 80 % všetkých organických látok v živom organizme. Ich najdôležitejšia funkcia je stavebná, ďalej zabezpečujú transport iných látok po bunke, môžu slúžiť ako katalyzátory chemických reakcií a plnia mnoho ďalších úloh pre správny chod života bunky a následne celého organizmu. Pokiaľ je proteín katalyticky aktívny, označujeme ho ako enzým [2]. Proteíny sú užitočné nástroje, no ak v nich dôjde k mutácii alebo poškodeniu, ich funkcia sa môže natoľko zmeniť, že budú pre bunku škodlivé a toxické, dokonca môžu spôsobiť aj jej smrť.

Aby sa čitateľ jednoducho orientoval v ďalšom texte, zadefinujeme pojmy ako sú proteínová doména a proteínová rodina. Proteínovou doménou (ďalej len doména) označujeme skupinu evolučne spriahnutých pozícií, ktorá zabezpečuje určitú funkciu proteínu. Doména, aj po oddelení od zvyšku proteínu, nestráca svoju funkciu. Proteín môže obsahovať viacero domén a rovnaká doména sa u rôznych organizmov môže nachádzať v iných proteínoch, a stále pri tom zabezpečovať rovnakú úlohu. Skupinu evolučne blízkych proteínov, kde je prítomná zhoda aminokyselín aspoň 30 % pre všetky dvojice proteínov, označujeme ako proteínová rodina (ďalej len rodina). Podmnožiny takejto rodiny voláme podrodiny.

Keď hovoríme o štruktúre proteínu, rozlišujeme ju na štyroch základných úrovniach.

Prvou úrovňou je primárna štruktúra, ktorú si vieme predstaviť ako lineárny reťazec - sekvenciu konkrétnych aminokyselín, tvoriacich tento proteín. Primárna štruktúra teda určuje poradie aminokyselín v proteíne a vplýva na jeho skladanie vo vyšších úrovniach. Podľa Anfinsenovej dogmy [27] existuje jediná výhodná priestorová konformácia pre daný proteín a je presne determinovaná primárnou štruktúrou, avšak môže sa meniť vplyvom vonkajšieho prostredia. Navyše aj zmena jedinej aminokyseliny môže mať závažné následky na biochemické vlastnosti proteínu. Sekundárna štruktúra je priestorová, pričom sa aminokyseliny skladajú do dvoch hlavných priestorových útvarov - takzvané α-helixy a β-skladané listy. Ďalšou úrovňou priestorového skladania proteínov je terciárna štruktúra, ktorá predstavuje konečné priestorové - trojdimenzionálne (ďalej len 3D) usporiadanie proteínu. Proces, ktorým proteín prechádza od primárnej štruktúry až po terciárnu, označujeme ako skladanie proteínu. Ak sa viaceré proteíny navzájom spájajú a tvoria tak veľké komplexy, hovoríme o ich kvartérnej štruktúre [2].

Pri uvažovaní o fungovaní proteínov musíme mať na pamäti, že proteín nadobúda svoju funkciu až po dosiahnutí optimálnej priestorovej konformácie, teda priestorového usporiadania, ktoré je pre ňu energeticky najvýhodnejšie. Keďže teda terciárna štruktúra determinuje funkciu proteínu, je zaujímavé sa ňou zaoberať a skúmať ju. Stále totiž nie je úplne jasné, čo presne funkcie proteínov podmieňuje, ako proteín „vie“, do akej konformácie sa má dostať a na základe čoho vlastne plní svoju funkciu. V súčasnosti sa v databáze proteínových sekvencií UniProt (z angl. *The Universal Protein Resource*, [47]) nachádza vyše 200 miliónov sekvencií proteínov, z čoho iba približne 184 tisíc má predikovanú 3D štruktúru, dostupnú v databáze proteínových štruktúr PDB (z angl. *Protein Data Bank*, [8]). V záujme vedcov, či už biológov alebo bioinformatikov, je skúmať a objavovať nové terciárne štruktúry proteínov, aby sme sa o ich rozličných úlohách dozvedeli čo najviac. Na tieto poznatky sa dá následné nadviazať napríklad pri študovaní vplyvu prostredia na fungovanie proteínu alebo pri dizajnovaní a vytváraní proteínov s novou cielenou funkciou. V neposlednom rade sa informácie o funkciách jednotlivých proteínov využívajú v oblasti medicíny pri návrhu nových liečiv.

1.2 Dokážeme niečo vyčítať z primárnej štruktúry?

Už sme spomenuli, že aktivita proteínu je určená priestorovou štruktúrou. Faktom však je, že veľa proteínov ju stanovenú nemá a ich pôsobenie je stále neznáme. Získanie terciárnej štruktúry je možné pomocou laboratórnych techník ako je röntgenová kryštalografia alebo NMR (z angl. *Nuclear Magnetic Resonance*) spektroskopia [48]. Proces oboch techník je však drahý a technicky náročný. V poslednom období je ako alternatíva experimentálnych techník stále viac využívaná umelá inteligencia, ktorú využíva napríklad nástroj Alphafold, ktorý dokáže s relatívne vysokou presnosťou

predikovať 3D štruktúru z primárnej sekvencie [24]. Napriek známej 3D štruktúre však nevieme predikovať, ktoré pozície sa bezprostredne podieľajú na biochemickej funkcii a evolúcii adaptácií proteínu. Jednou z možností je, že niektoré aminokyseliny tvoria akúsi „kostru“, na ktorú sú „zavesené“ ostatné aminokyseliny. Touto myšlienkou sa zaoberal okrem iných vedcov aj Rama Ranganathan s jeho kolektívom [20], na ktorých nadviazali mnohí ďalší [4].

Prvým krokom je určiť, ktoré aminokyseliny sú významnejšie ako ostatné, teda ktoré viac prispievajú k fungovaniu proteínu. Jednou z možností by bolo experimentálne skúšanie odstraňovania alebo vymieňania aminokyselín, čo však vyžaduje veľa úsilia a v závere by sa ani nemuselo dospieť k rozumnému výsledku. Preto vznikol nový nápad zamerať sa čisto na sekvencie proteínu ako na text, ktorý sa dá štatisticky vyhodnocovať. Z jednej sekvencie proteínu sa však veľa informácií zistiť nedá. Preto je lepšie zhromaždiť a zarovnať pod seba sekvencie proteínov z jednej rodiny tak, aby sme získali čo najpresnejšiu zhodu písmen v jednotlivých stĺpcoch. Takémuto súboru sekvencií hovoríme zarovnanie. Samozrejme, nebudeme zarovnávať úplne totožné sekvencie, ale sekvencie z rôznych organizmov, ktoré sa od seba aspoň mierne odlišujú, aby sme vedeli štatistické nástroje rozumne použiť a vyčítať z nich relevantné výsledky. To, že sa sekvencie rovnakého proteínu v rôznych organizmoch odlišujú, je dôsledkom evolúcie. Organizmy si ten istý proteín prispôsobili pre svoje potreby, pričom hlavná funkcia proteínu ostala zachovaná.

1.3 Zo zarovnania vieme zistiť viac

Ako zo zarovnania zistíme, ktoré aminokyseliny sú práve tie vplyvné a ktoré nemajú v proteíne významnú rolu? Jednou z možností je, že sa v stĺpci zarovnania nachádzajú všetky aminokyseliny s približne rovnakou frekvenciou výskytu. Vzhľadom na to, že viac-menej nezáleží na prítomnej aminokyseline, nepredpokladáme, že má pozícia významný vplyv. Podobne, pokiaľ sa nachádza v stĺpci príliš veľa medzier, pozícia zjavne nie je dôležitá, pretože sa u mnohých organizmov ani nevyskytuje. Naopak, ak je nejaká aminokyselina v stĺpci do väčšej miery zachovaná, dá sa o jej pozícii uvažovať ako o podstatnej, keďže si viaceré organizmy konkrétnu aminokyselinu alebo jej funkčne podobnú ponechali. V takomto prípade označujeme pozíciu za konzervovanú. Hodnota konzervovanosti pozície nám teda hovorí, do akej miery sa frekvencie výskytov jej aminokyselín odlišujú od náhodných ([10], [20]). Zároveň platí, že čím sú tieto frekvencie odlišnejšie, tým je silnejšie takzvané evolučné obmedzenie tejto pozície. Z neho vyplýva odolnosť pozície voči zmenám, teda tendencia zachovania znaku pozície, následkom čoho môžeme predpokladať jej výraznejší vplyv na samotnú funkciu proteínu. Zistilo sa, že najviac konzervované pozície sú tie, čo sa podieľajú na katalytickej

činnosti proteínu a tie, čo sa nachádzajú hlavne vo vnútri proteínu v okolí jeho jadra, zatiaľ čo málo konzervované pozície sú lokalizované pri povrchu proteínu a väčšinou plnia iba podpornú funkciu [20].

Štatistické nástroje nám umožnia viac ako len pozrieť sa na jednu pozíciu. Vzhľadom na to, že aminokyseliny spoločne vytvárajú funkciu proteínu, je zaujímavé preskúmať ich spolupôsobenie a zistiť, či zmena aminokyseliny na pozícii i nevyvolá zmenu na pozícii j. Vďaka koreláciám dvojíc by sa dali nájsť potenciálne spolupracujúce aminokyseliny, ktoré by sa mohli podieľať na spoločnej funkcii proteínu. Bolo by samozrejme zaujímavé pozrieť sa so štatistikou na väčšie skupiny pozícií naraz, ale experimentálne sa zistilo, že hľadanie korelácií u dvojíc, teda takzvaná štatistika druhého rádu, je najvhodnejšia na zisťovanie vzájomných vzťahov a nie je potrebné zaoberať sa štatistikou vyšších rádov [20]. Hovoríme, že dvojica pozícií koevolvuje, pokiaľ má zmena jednej vysoký vplyv na druhú. Znamená to, že zmena prvej aminokyseliny pravdepodobne vyvolá aj zmenu druhej z dvojice s cieľom obnoviť fungovanie proteínu. Keďže na seba takéto pozície vplývajú, môžeme ich označiť ako korelujúce pozície.

1.4 Neviditeľné štruktúry proteínov?

Výhodou nájdenia korelujúcich dvojíc je, že informácie o nich sú smerodajné, pokiaľ chceme hľadať väčšie, funkčne spolupracujúce celky, teda skupiny silno koevolvujúcich pozícií. Takúto skupinu nazývame proteínový sektor (ďalej len sektor) [20]. Sektory v skutočnosti nemôžeme zaradiť medzi klasické úrovne proteínovej štruktúry. Nedajú sa totiž zo žiadnej zo štruktúr jednoducho odvodiť a nie je možné ich na prvý pohľad bez štatistických analýz nájsť. Pre sektor je podstatné aj to, že v rámci neho pozície koevolvujú, avšak s pozíciami z ostatných sektorov, prípadne s tými, čo neboli priradené do žiadneho, nejavia vysokú koevolúciu. Keďže sektory obsahujú primárne evolučne zachované a navzájom koevolvujúce pozície, môžeme ich označiť ako evolučné jednotky proteínov. Sektory napokon prestavujú potenciálne zaujímavé skupiny pozícií pre vykonávanie rôznych empirických skúmaní. Vďaka výsledkom bioinformatickej analýzy nemusíme tieto skupiny hľadať pracným experimentálnym skúšaním.

1.5 PARP

PARP proteíny sú významné enzýmy, ktoré sa podieľajú na opravách poškodení DNA [26]. Skupina PARP proteínov je kódovaná sedemnástimi rôznymi génmi, pričom zachovaná v nich je katalytická doména. Konkrétne PARP1 patrí medzi najlepšie preskúmaný PARP proteín a zistilo sa, že sa zameriava na opravy jedno- a dvojvláknových zlomov DNA v bunke [5]. Obsahuje aj viacero iných domén, ktoré sa v závislosti od

organizmu môžu meniť. Podstatnou je však spomínaná katalytická doména, zabezpečujúca jeho hlavnú funkciu, spočívajúcu v pridávaní jedného alebo viacerých zvyškov adenozíndifosfát ribózy (ďalej len ADP-ribózy) na postranné reťazce cielených proteínov. Po nájdení poškodenia DNA sa na DNA naviaže a ADP-ribozyluje proteíny, ktoré sa v tomto mieste nachádzajú. Túto ADP-ribózovú značku rozpoznajú ďalšie enzýmy, ktoré budú následne v opravovaní poškodenej DNA pokračovať [30]. Enzymatická aktivita bola dokázaná napríklad u človeka, ale predpokladá sa, že u organizmov, ktoré tento proteín majú, s vysokou pravdepodobnosťou funguje na princípe, ktorý sme popísali, hoci táto funkcia nemusela byť u nich testovaná. Proteíny PARP sa vyskytujú v rôznych organizmoch, boli objavené v baktériách, rastlinách aj živočíchoch ([19], [35]).

Znamená to, že proteíny PARP plnia esenciálnu funkciu pre život a delenie bunky, čo by, v prípade prítomnosti priveľkého množstva poškodení DNA, bunka nedokázala. Pre zdravé bunky sú teda nevyhnutné. Rovnako to však platí aj pre nádorové bunky, ktoré sú výnimočné tým, že sú viac-menej nesmrteľné a vedia sa nekontrolovateľne deliť. Takéto rýchle delenie buniek však spôsobuje veľa mutácií a zmien v genetickom kóde bunky, následkom čoho majú nádorové bunky extrémne vysoké nároky na svoje opravné mechanizmy DNA, pričom na týchto opravách sa podieľa aj proteín PARP1. V záujme liečenia rakoviny preto vznikla myšlienka zakomponovať do liečby práve znefunkčnenie proteínu PAPR1 tak, že sa nebude schopný viazať na DNA, čím sa zabráni efektívnej oprave poškodenia DNA ([42], [43]). Takýmto látkam, ktoré bránia funkcií enzýmov, hovoríme inhibítory [33]. PAPR inhibítory v kombinácií so štandardnou chemoterapiou zabezpečia takzvanú syntetickú letalitu nádorových buniek - bunky budú mať príliš veľa poškodení DNA a nefunkčný PARP1 enzým, následkom čoho nebudú schopné opravovať zmutovanú a poškodenú DNA a odumrú. Zistilo sa, že takáto kombinovaná liečba je veľmi účinná. Ako asi žiadna liečba rakoviny nie je úplne dokonalá, tak isto nie je ani liečba PARP inhibítormi, a preto je vo všeobecnom záujme naďalej skúmať fungovanie PARP proteínov pre ich efektívne a cielené ovládanie.

Napriek tomu, že PARP proteíny sú dobre preskúmané, je prínosné sa o nich dozvedieť ešte viac. Dajú sa pozorovať vlastnosti celej PARP rodiny naprieč viacerými taxonomickými skupinami a potom porovnávať správanie a znaky týchto proteínov u rôznych organizmov. Do nedávna sa myslelo, že proteín PARP u kvasiniek chýba, a preto u nich nebol študovaný. Na Prírodovedeckej fakulte Univerzity Komenského sa však tento proteín podarilo objaviť u druhu Yarrowia lipolytica, a teda je možné zaradiť ho do štúdie spolu s ostatnými PARP proteínmi a zistiť o ňom nové informácie. Fungovanie tohto proteínu je síce obdobné ako fungovanie spomínaného PARP1, avšak nie je úplne jasné, aká je motivácia pridávania značky v podobe ADP-ribózy a či sa skutočne podieľa na opravách DNA. Preto sme sa v našej práci zamerali na objavenú katalytickú doménu a práve jej sekvenciám z kvasiniek sme prispôsobili celú analýzu, aby bola čo najrelevantnejšia práve pre ne. Kvasinky sú navyše výborným modelovým

organizmom aj v laboratóriách a získané poznatky z bioinformatickej analýzy sa dajú relatívne ľahko využiť aj experimentálne.

Hlavným cieľom našej práce bolo porozumenie fungovania a použitie bioinformatických nástrojov pre hľadanie koevolvujúcich pozícií a proteínových sektorov. Skúmali sme taktiež odlišnosti a spojitosti týchto nástrojov. Keďže pre ľudský PARP1 už máme stanovených viacero terciárnych štruktúr, ktoré sú blízke aj ku nájdeným doménam kvasiniek, na mapovanie pozícií sme použili jednu z nich, konkrétne 1WOK, získanú z verejnej databázy RCSB PDB (z angl. *The Research Collaboratory for Structural Bioinformatics Protein Data Bank*, [8]).

Kapitola 2

Bioinformatické metódy

V tejto kapitole rozoberieme princípy fungovania dvoch bioinformatických metód na stanovenie konzervovanosti jednotlivých pozícií, prípadne na následné určenie polohy proteínového sektora. Vysvetlíme vzťahy medzi jednotlivými metódami, ich spoločné ciele, ale aj rozdielne postupy a stanoviská.

2.1 Statistical coupling analysis

Štatistická analýza prepojení (ďalej len SCA, z angl. *Statistical Coupling Analysis*) je bioinformatická metóda na nájdenie proteínových sektorov [38]. Funguje v dvoch základných krokoch: štatistika prvého a druhého rádu. Jej základnou podstatou je určenie miery konzervovanosti všetkých pozícií proteínu a s jej pomocou nájdenie tých pozícií, ktoré spolu koevolvujú. Predstavitelia SCA navyše tvrdia, že nie všetky aminokyseliny na pozíciách, ktoré spolu koevolvujú, musia byť aj v priamom fyzickom kontakte v terciárnej štruktúre. Finálnym krokom je samotná lokalizácia sektoru a definovanie jeho polohy, teda ktoré pozície a k nim prislúchajúce aminokyseliny ho tvoria. V nasledujúcom texte ilustrujeme priebeh programu obrázkami, ktoré vznikli pri samotnej analýze nášho pozorovaného proteínu PARP pre náhodnú množinu 500 sekvencií z celej jeho rodiny (vstup nahodny_vyber, viď. Tab. 3.1, pre kód viď. Dodatok B: elektronická príloha: Kody/printResultsSK.py).

2.1.1 Príprava vstupných dát

Vstupom SCA je M zarovnaných sekvencií proteínov s L pozíciami v zarovnaní. Môžeme predpokladať, že sekvencie, ktoré majú veľa medzier a nedefinovaných aminokyselín, sa v tomto zarovnaní nenachádzajú. Aby sme získané výsledky vedeli vizualizovať, budeme požadovať, aby sme pre aspoň jednu sekvenciu poznali atomickú 3D štruktúru proteínu. Zároveň predpokladáme, že M predstavuje dostatočne veľký počet sekvencií pre dôveryhodné stanovenie frekvencií aminokyselín na jednotlivých pozíciách.

Predtým, ako sa samotné frekvencie vypočítajú, je potrebné určiť „váhu“ každej sekvencie. Pridávanie váh sekvenciám je potrebné z toho dôvodu, že vo vstupnom zarovnaní máme rôzne sekvencie proteínov, ktoré sú si navzájom podobné. V niektorých prípadoch však môže byť podobnosť reťazcov veľmi vysoká, čo by mohlo viesť ku skresleným a nepresným výsledkom. Nech w_s je váha sekvencie s a nech N_s je počet sekvencií, ktoré majú zhodu aspoň 80 % so sekvenciou s [38]:

$$w_s = \frac{1}{N_s} \tag{2.1}$$

Pridaním váh w_s zabezpečíme, že všetky príliš podobné sekvencie zoberieme do úvahy pri počítaní ďalších hodnôt iba ako 1 celok, teda iba ako 1 sekvenciu. Znamená to, že podobným sekvenciám, ktoré budú v zarovnaní vo väčšom počte priradíme menšiu váhu ako tým, ktoré budú mať v zarovnaní menej podobných sekvencií. Pomocou vypočítaných váh vieme zároveň vyrátať efektívny počet sekvencií M_{eff}:

$$M_{eff} = \sum_s w_s \tag{2.2}$$

Výskyt aminokyseliny a v sekvencii s na pozícii i označíme premennou x^a_{si}, pričom $x^a_{si} = 1$ pokiaľ sa a na pozícii i v sekvencii s vyskytuje, inak platí, že $x^a_{si} = 0$. Frekvenciu aminokyseliny a na pozícii i v rámci celého zarovnania (teda v rámci všetkých sekvencií) označíme ako f^a_i. Zavedieme ešte regularizačný parameter λ, ktorý zabezpečí nenulové hodnoty f^a_i, potom [38]:

$$f^a_i = \frac{1-\lambda}{M_{eff}} \sum_s w_s x^a_{si} + \frac{\lambda}{21} \tag{2.3}$$

Priemernú frekvenciu výskytu aminokyseliny a vo všetkých existujúcich proteínových sekvenciách označíme q^a. Túto hodnotu získame z dát z NCBI databázy (z angl. *National Center for Biotechnology Information*, [32], [41]). Ak q^a chceme spresniť, potrebujeme vyjadriť celkovú frekvenciu medzier vo všetkých sekvenciách nášho zarovnania, $\bar{q}^0$, pričom platí rovnosť [41]: $\bar{q}^0 = \sum_i f^0_i / L$, kde f^0_i vyjadruje frekvenciu medzier v i-tom stĺpci a platí: $f^0_i = 1 - \sum_{a=0}^{20} f^a_i$. Frekvenciu f^0_i môžeme chápať ako opačnú pravdepodobnosť výskytu hocijakej aminokyseliny na danej pozícii. Spresnenú hodnotu q^a vieme využiť napríklad pre výpočet celkovej konzervovanosti nejakej pozície (2.9) a zapíšeme ju ako $\bar{q}^a$:

$$\bar{q}^a = (1 - \bar{q}^0) q^a \tag{2.4}$$

Keďže sa však v SCA pracuje najmä s dvojicami pozícií, potrebujeme vyjadriť aj spoločnú frekvenciu dvoch aminokyselín a a b na pozíciách i a j, pričom opäť použijeme rovnaký regularizačný parameter λ, ako [38]:

$$f^{ab}_{ij} = \frac{1-\lambda}{M_{eff}} \sum_s w_s x^a_{si} x^b_{sj} + \frac{\lambda}{21^2} \tag{2.5}$$

2.1.2 Relatívna entropia ako nástroj na vypočítanie konzervovanosti pozície

Prvým krokom SCA je štatistika prvého rádu - stanovenie konzervovanosti (viď. Kap. 1.2) každej pozície proteínu. Túto hodnotu vyjadríme pomocou relatívnej entropie. Entropia v teórii pravdepodobnosti predstavuje mieru neurčitosti alebo neistoty náhodného procesu. Entropia pre diskrétnu náhodnú premennú X s pravdepodobnosťami $p(x_i)$ pre elementárne udalosti $x_1, \ldots, x_n$ je udávaná v bitoch a je definovaná Shannonovým vzorcom [13] :

$$H(X) = -\sum_{i}^{n} p(x_i) \log_2 p(x_i) \tag{2.6}$$

Relatívnu entropiu vyjadrujeme ako mieru vzdialenosti medzi dvoma pravdepodobnostnými distribúciami. Keďže však táto miera nie je symetrická, nedá sa o nej hovoriť ako o vzdialenosti v pravom slova zmysle. Napriek tomu sa často o relatívnej entropii uvažuje práve týmto spôsobom. Pre pravdepodobnostné distribúcie p a q je nasledovne definovaná [13]:

$$D(p \parallel q) = \sum_{i}^{n} p(x_i) \ln \frac{p(x_i)}{q(x_i)} \tag{2.7}$$

Môžeme si všimnúť, že relatívna entropia nadobúda len nezáporné hodnoty a klesá, ak sa $p(x_i)$ blíži ku $q(x_i)$ a rovná nule je iba v prípade, že platí $p(x_i) = q(x_i)$ pre všetky $i \in \{1, ..., n\}$.

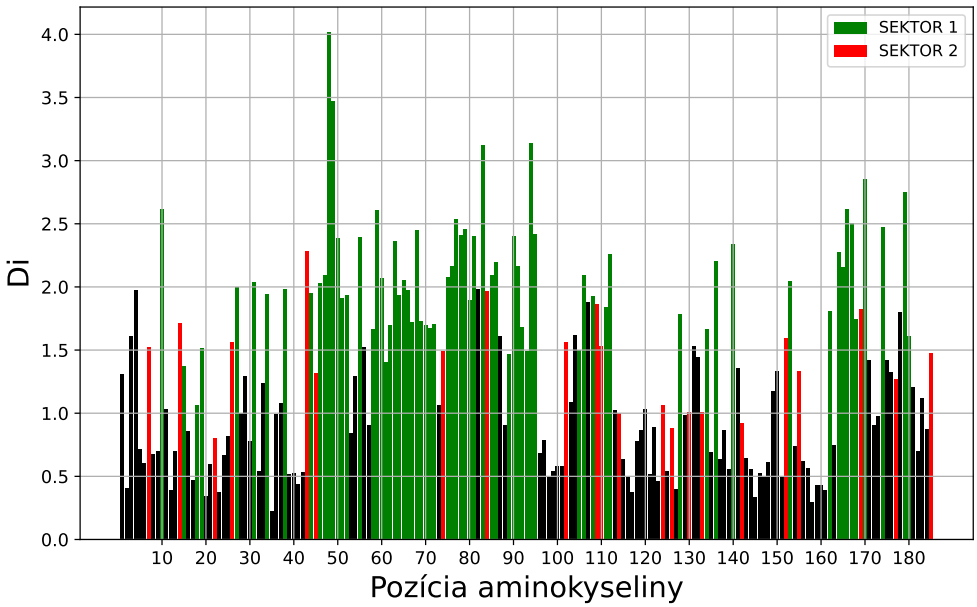
Konzervovanosť aminokyseliny a na pozícii i označíme D_i^a a na jej výpočet použijeme práve spomínanú relatívnu entropiu. Vyjadrujeme teda vzdialenosť medzi dvoma pravdepodobnostnými distribúciami, konkrétne medzi frekvenciou aminokyseliny a na pozícii i, teda f_i^a a jej priemernou frekvenciou vo všetkých proteínoch q^a. D_i^a zapíšeme nasledovne:

$$D_i^a = f_i^a \ln \frac{f_i^a}{q^a} + (1 - f_i^a) \ln \frac{1 - f_i^a}{1 - q^a} \tag{2.8}$$

Vidíme, že vzorec D_i^a je viazaný iba na jednu aminokyselinu. Preto sa aj hodnota entropie odvíja iba od toho, či sa daná aminokyselina na danej pozícii nachádza (s pravdepodobnosťou f_i^a) alebo nie (s pravdepodobnosťou $1 - f_i^a$). Ak chceme vyjadriť konzervovanosť pozície i vzhľadom na všetkých 20 aminokyselín, je navyše potrebná práve hodnota $\bar{q}^a$ (2.4) a platí:

$$D_i = \sum_{a=0}^{20} f_i^a \ln \frac{f_i^a}{\bar{q}^a} \tag{2.9}$$

Pomocou D_i hodnôt vieme napríklad vizualizovať mieru konzervovanosti každej pozície na histograme (viď. Obr. 2.1).



Obr. 2.1: Histogram zobrazujúci mieru celkovej konzervovanosti jednotlivých pozícií v katalytickej doméne proteínu PARP. Príslušnosť pozícií do nájdených proteínových sektorov je vyjadrená odlišnými farbami, pričom čierne stĺpce predstavujú pozície, ktoré nie sú priradené do žiadneho sektora.

2.1.3 Korelácie konzervovaných pozícií

Druhý krok SCA je štatistika druhého rádu - vytvorenie takzvanej váhovanej korelačnej matice konzervovanosti (ďalej len matica konzervovanosti alebo korelačná matica), ktorá bude určovať korelácie medzi jednotlivými dvojicami pozícií (viď. Obr. 2.2). Vďaka nej budeme vedieť povedať, ktoré dvojice pozícií navzájom koevolvujú (viď. Kap. 1.3). Jej prvok, koreláciu dvoch aminokyselín (a, b) na pozíciách (i, j), vyjadríme ako rozdiel spoločnej frekvencie dvoch pozícií f_{ij}^{ab} a $f_i^a f_j^b$, čo predstavuje očakávaný spoločný výskyt aminokyselín (a, b) na pozíciách (i, j) bez prítomnosti ich vzájomnej korelácie [21]:

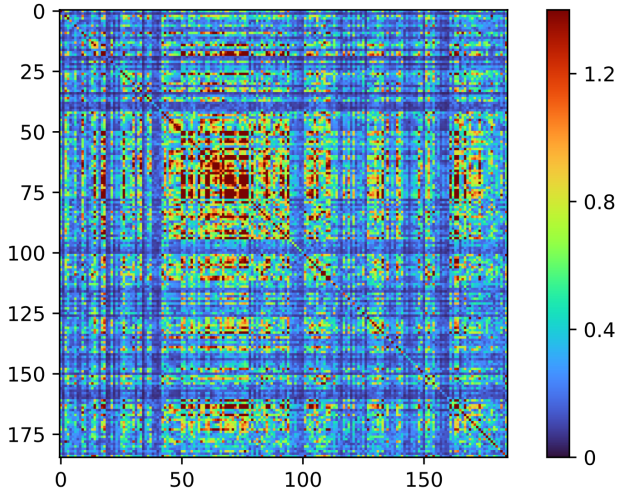
$$C_{ij}^{ab} = f_{ij}^{ab} - f_i^a f_j^b \tag{2.10}$$

Na to, aby sa zobrala do úvahy aj konzervovanosť samostatných pozícií, je potrebné pridať hodnotám C_{ij}^{ab} váhu. Miera váhy danej pozície i s aminokyselinou a, teda ϕ_i^a, sa dá vyjadriť ako zmena (derivácia) konzervovanosti D_i^a vzhľadom na f_i^a ([21], [41]). Potom:

$$\phi_i^a = \frac{\partial D_i^a}{\partial f_i^a} = \ln\left[\frac{f_i^a(1-q^a)}{(1-f_i^a)q^a}\right] \tag{2.11}$$

Prvok matice konzervovanosti bude potom definovaný ako:

$$\tilde{C}_{ij}^{ab} = \phi_i^a \phi_j^b C_{ij}^{ab} \tag{2.12}$$



Obr. 2.2: Korelačná matica $\tilde{C}_{ij}$. Na osi x a y sú pozície domény proteínu PARP a hodnota jej prvku je hodnota korelácie medzi dvoma pozíciami. Vidíme silno korelovaný zhluk v hornej časti matice. V rámci celej matice pozorujeme aj v primárnej štruktúre vzdialené, no napriek tomu silno korelované dvojice.

V praxi sa najčastejšie stretávame s dvojrozmernými maticami, pretože práca s nimi je jednoznačná a oveľa jednoduchšia ako s viacrozmernými. Keďže je aktuálna matica konzervovanosti $\tilde{C}_{ij}^{ab}$ štvorrozmerná, na jej prevod do dvoch rozmerov použijeme Frobeniovu normu [12] pre každú maticu veľkosti 20x20, ktorá predstavuje maticu konzervovanosti pre dvojicu pozícií (i, j), následkom čoho dostávame [41]:

$$\tilde{C}_{ij} = \sqrt{\sum_{a,b} \left(\tilde{C}_{ij}^{ab}\right)^2} \tag{2.13}$$

Navyše, pri vykonávaní štatistických analýz často dochádza k chybám, spôsobeným rôznymi faktormi. V SCA sa pri vytváraní matice konzervovanosti objavuje takzvaný štatistický a historický šum. Štatistický šum vzniká v dôsledku použitia konečného množstva dát a historický šum sa tvorí kvôli fylogenetickým príbuznostiam medzi organizmami, ktorých sekvencie proteínov pozorujeme. Oba druhy šumov zapríčiňujú nerelevantnosť niektorých hodnôt matice, a preto je žiadúce sa šumom vyhnúť, respektíve ich nejakým spôsobom potlačiť. Historický šum je možné eliminovať vhodným výberom sekvencií tak, aby sme mali zastúpenie rôznych organizmov v našej vzorke približne rovnaké. Taktiež váhovanie sekvencií pomáha tento šum minimalizovať. Ostáva nám preto vyriešiť štatistický šum, ktorého príčinou vzniku v našom prípade je práca s konečným počtom vstupných sekvencií. Jeho potlačenie zabezpečíme dekompozíciou matice podľa vlastných hodnôt. Tomuto problému sa venujeme v nasledujúcej podkapitole (viď Kap. 2.1.4).

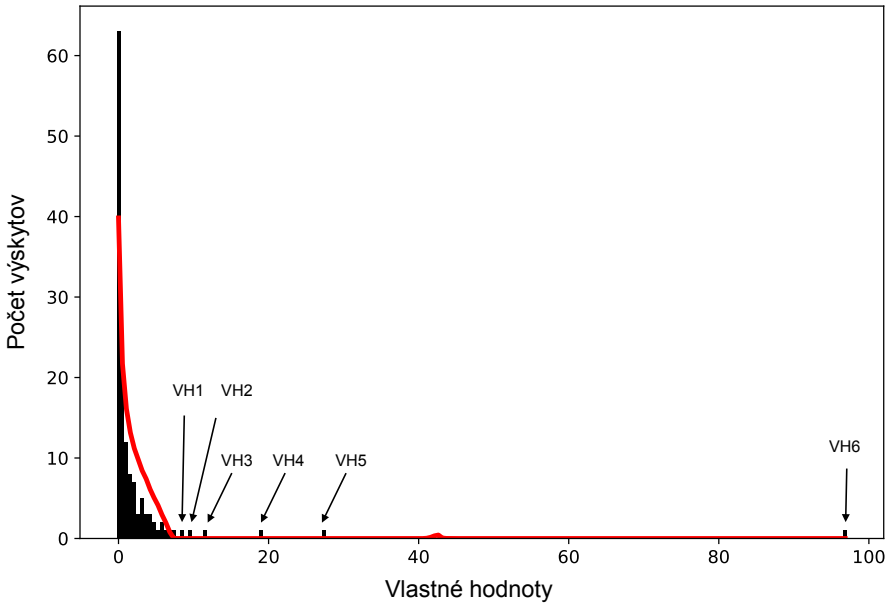
2.1.4 Poloha sektora

Nájdenie a určenie polohy sektora je primárnym cieľom celej SCA analýzy. Považujeme ju za najšpecifickejšiu a výpočtovo najnáročnejšiu časť SCA. Ako získame z matice skupiny štatisticky nezávislých korelujúcich pozícií? Hľadanie sektorov využíva rôzne metódy z algebry a štatistiky a spočíva v troch na seba nadväzujúcich krokoch. Prvým z nich je dekompozícia matice $\tilde{C}_{ij}$ pomocou vlastných hodnôt, pričom dekompozíciu zapíšeme pomocou násobenia matíc [21]:

$$\tilde{C} = \tilde{V}\tilde{\Delta}\tilde{V}^T \tag{2.14}$$

Matica $\tilde{\Delta}$ je diagonálna matica, obsahujúca vlastné hodnoty. Pozície matice $\tilde{C}_{ij}$ lineárne kombinujeme a premietame ich do nových dimenzií - vlastných vektorov, ktoré sú uložené do stĺpcov matice $\tilde{V}$. Vlastné hodnoty nám potom hovoria o dôležitosti príslušných vlastných vektorov. Tým, že potrebujeme získať iba štatisticky významné vlastné hodnoty, slúži táto dekompozícia zároveň aj ako nástroj na elimináciu štatistického šumu. Znamená to, že sa zbavíme takých vlastných hodnôt, ktoré by sme získali aj z náhodného vstupného zarovnania. Takúto náhodnú vzorku vytvoríme vertikálnym premiešaním stĺpcov pôvodného vstupného zarovnania, teda premiešame aminokyseliny v rámci každého stĺpca zvlášť. Takto odstránime všetky korelácie a zároveň zachováme jednotlivé frekvencie pozícií. Z novozískaného zarovnania vytvoríme novú maticu konzervovanosti a z nej nové vlastné hodnoty. Toto premiešanie spravíme veľa krát a pre každú vlastnú hodnotu vypočítame jej priemernú hodnotu. Hľadané signifikantné vlastné hodnoty z našej prvotnej matice konzervovanosti nájdeme tak, že ponecháme všetky jej vlastné hodnoty, ktoré sú väčšie ako druhá najväčšia priemerná vlastná hodnota získaná z matíc konzervovanosti náhodných zarovnaní. Najväčšia vlastná hodnota je totiž vždy považovaná za triviálny dôsledok zachovania nezávislej konzervovanosti každej pozície, keďže sa aminokyseliny premiešavajú iba v rámci stĺpcov a nie riadkov (viď. Obr. 2.3, [21]). K vybraným vlastným hodnotám potom dostaneme aj výber vlastných vektorov, ktoré nám aktuálne tvoria prvotné rozdelenie pozícií - pre každú pozíciu máme skóre príslušnosti do každého z vlastných vektorov [38].

Prečo nám však nestačí táto dekompozícia a je nutné vykonať ďalšiu? Problém nastáva v tom, že dekompozícia matice pomocou vlastných hodnôt nie je dostatočne silná na rozdiel od úplnej štatistickej nezávislosti. Pri jej hľadaní je totiž potrebné nie len zbaviť sa korelácií medzi dvojicami pozícií, ale aj zbavenie sa korelácií vyšších rádov, na čo využijeme takzvanú analýzu nezávislých komponentov (ďalej len ICA, z angl. *Independent component analysis*). ICA zabezpečí pomocou numerických optimalizácií transformáciu najlepších vlastných hodnôt a ich vektorov do maximálne nezávislých komponentov (ďalej len komponentov) [7]. Zahŕňa iteratívne vytvorenie špeciálnej matice W, pomocou ktorej túto transformáciu zabezpečí [21].



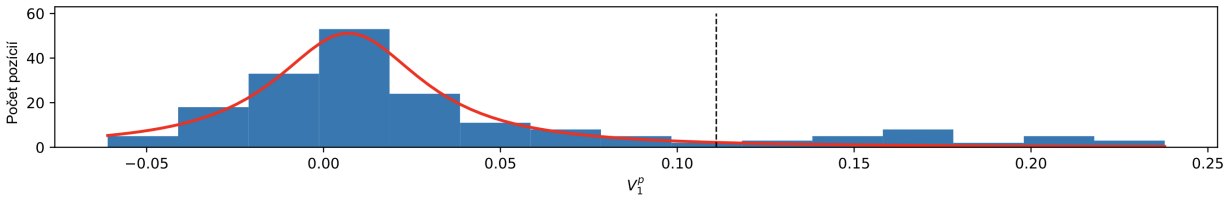
Obr. 2.3: Graf znázorňujúci distribúciu priemerných vlastných hodnôt z náhodných matíc (červená čiara) a z pôvodnej matice $\tilde{C}_{ij}$ (čierne stĺpce). Signifikantné vlastné hodnoty (VH) sú označené šípkou.

Nakoniec získame maticu $\tilde{V}^p$, ktorej jednotlivé stĺpce odvodíme nasledovne, pričom x predstavuje počet signifikantných vlastných hodnôt:

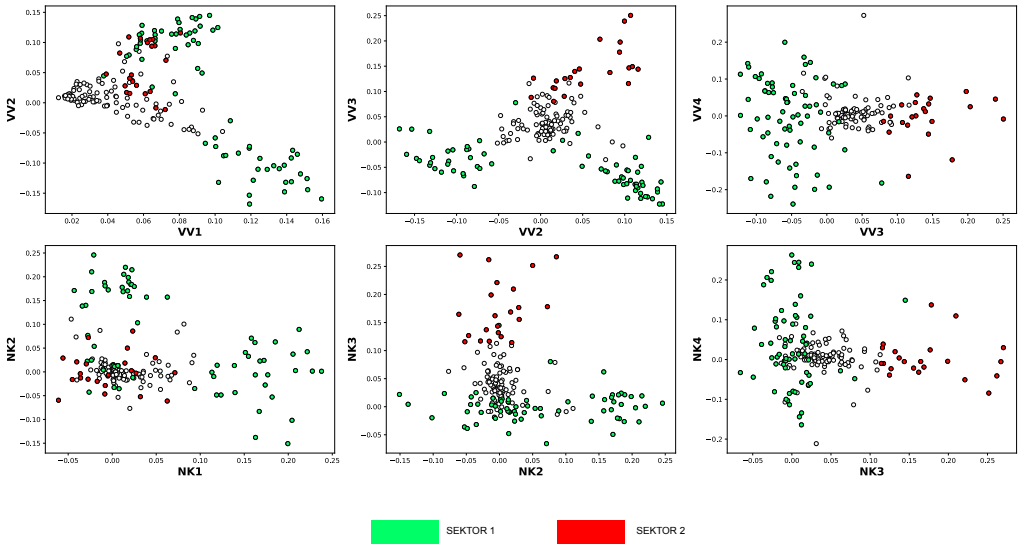
$$\tilde{V}^p_{1\ldots x} = W\tilde{V}_{1\ldots x} \tag{2.15}$$

Prvky v matici $\tilde{V}^p$ nám určujú dostatočne presné a dôveryhodné skóre príslušnosti pozície do náležitého nezávislého komponentu. Distribúcia týchto hodnôt pre každý nezávislý komponent sa dá namapovať na Studentovo rozdelenie, pričom do daného nezávislého komponentu sú nakoniec priradené tie pozície, ktorých skóre príslušnosti patrí medzi 5 % najvyšších skóre pre daný komponent (viď. Obr. 2.4). ICA navyše zabezpečuje, že ak je jedna pozícia v rámci najlepších 5 % vo viac než jednom nezávislom komponente, pomocou jednoduchých algoritmov správne priradí pozíciu len do jedného z nich. Pri rozhodovaní o priradení do nezávislého komponentu sa snaží dosiahnuť čo najvyššiu mieru koevolúcie pozícií v rámci komponentu [40]. Z medzivýsledkových vizualizácií je vidieť, že dekompozícia do nezávislých komponentov lepšie rozdeľuje pozície do štatisticky nezávislých skupín (viď. Obr. 2.5).

Posledným krokom je na základe nájdených nezávislých komponentov určiť umiestnenie proteínových sektorov. Bohužiaľ, nie vždy jeden nezávislý komponent determinuje práve jeden sektor, pretože ani samotná ICA nezabezpečí stopercentnú nezávislosť. Preto treba získané komponenty hlbšie preskúmať.

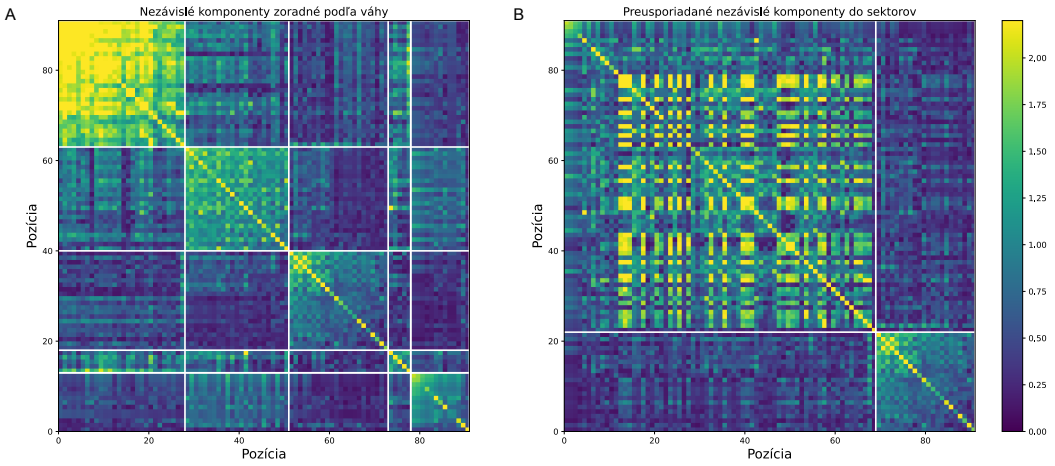


Obr. 2.4: Mapovanie hodnôt príslušnosti všetkých pozícií do nezávislého komponentu č.1 ku Studentovmu rozdeleniu. Prerušovaná čiara ukazuje minimálnu hodnotu pre zaradenie pozície do daného komponentu (5 % najlepších). Červená krivka znázorňuje priebeh Studentovho rozdelenia.



Obr. 2.5: Grafy v hornom rade znázorňujú závislosti príslušností pozícií do vlastných vektorov(VV). Z grafov nie je úplne jasné, ktoré pozície sú príznačnejšie k nejakému z vektorov. Naopak, na grafoch v spodnom rade, kde sú zobrazené závislosti medzi skóre príslušností pozícií do nezávislých komponentov (NK), je príslušnosť ku jednotlivým komponentom jednoznačnejšia (usporiadanie sa pozícií do tvaru písmena L)

Každý takýto nájdený komponent môže mať dve možné vysvetlenia: (1) skutočne predstavuje nájdený nezávislý sektor, alebo (2) predstavuje iba istú časť nejakého sektora. Z toho vyplýva, že treba systematicky hľadať a odlíšiť také komponenty, ktoré samy tvoria sektor. Následne treba správne pospájať tie komponenty, ktoré spoločne tvoria ten istý sektor. Na rozdiel od predošlých fáz, tento krok sa robí ručne, na základe hľadania závislostí a korelácií pozícií naprieč jednotlivými komponentami pomocou matice, v ktorej vidíme mieru príslušnosti všetkých pozícií ku každému nezávislému komponentu a silu koevolúcie dvojíc pozícií. Jej preusporiadaním a zoskupením komponentov do nových celkov vieme vizuálne overiť aj závislosti získaných sektorov (viď. Obr. 2.6) [38].



Obr. 2.6: Tepelné mapy rozdelenia pozícií do nezávislých komponentov (NK) a následne sektorov. Jednotlivé riadky a stĺpce sú pozície, zaradené do nejakého nezávislého komponentu (resp. sektora). Mriežka zvislých a vodorovných bielych čiar predstavuje rozdelenie do nezávislých komponentov (resp. sektorov), teda napríklad prvý riadok a prvý stĺpec predstavuje prvý nezávislý komponent (resp. sektor). Nezávislé komponenty sú usporiadané podľa ich dôležitosti zhora (resp. zľava), od najvýznamnejšieho po najmenej významný. Jeden štvorček na pozícii $[i, j]$ predstavuje výšku korelácie dvojice pozícií i a j v matici. Miera korelácie je značená farebnou škálou od modrej až po žltú, pričom čím je farba žltšia, tým je korelácia silnejšia. Matica A zobrazuje pozície zoradené podľa príslušnosti do daného komponentu. Matica B zobrazuje preusporiadanú maticu A po zoskupení komponentov 1, 2, 4 a 5 a zachovania komponentu 3 v pôvodnom stave.

2.2 GREMLIN

GREMLIN (z angl. *Generative REgularized ModeLs of proteINs*) predstavuje metódu, ktorá bola vyvinutá skupinou Davida Bakera, využívajúcu učenie pre hľadanie štatistických modelov pre danú proteínovú rodinu [25]. Svoje výpočty rieši pomocou globálnych štatistických modelov - Markovove náhodné polia [6]. Jej hlavným cieľom je nájdenie dvojíc pozícií, ktoré sú vo fyzickom kontakte v terciárnej štruktúre proteínu, v konečnom dôsledku tak predikovať celkovú 3D štruktúru a zlepšovať takzvané komparatívne modely pre túto predikciu, pričom sa navyše môže používať aj pri dizajnovaní nových proteínov, s podobnými štatistickými vlastnosťami. Metóda je založená na tom, že práve koevolúcia je silnou a smerodajnou informáciou pre hľadanie priamych kontaktov [25].

V tejto práci nás v prípade metódy GREMLIN viac zaujíma chápanie a interpretácia výsledkov ako implementačné detaily a samotný beh programu, v ďalšej podkapitole však priblížime princípy jeho fungovania. Stretávame sa tu totiž aj s rozdielnymi

názormi predstaviteľov metódy SCA a GREMLIN na tému o vzdialenosti koevolvujúcich pozícií v terciárnej štruktúre proteínu. Na rozdiel od autorov SCA analýzy, Baker a kol. tvrdia, že takmer všetky dvojice, ktoré spolu koevolvujú je možné nájsť priamo vo fyzickom kontakte v aspoň jednej 3D štruktúre pozorovaného proteínu. V prípade SCA boli takto vzdialené koevolvujúce pozície práve niečím pochopiteľným, očakávaným a vysvetľovaným napríklad vzdialenou spoluprácou aminokyselín v proteínoch [37].

Hlavnou výhodou GREMLIN-u oproti SCA je to, že dokáže rozlíšiť takzvané priame a nepriame korelácie medzi pozíciami. Priame korelácie sú medzi pozíciami, ktoré naozaj koevolvujú. Avšak správanie takýchto algoritmov je tranzitívne, čo spôsobí, že ak existuje priama korelácia $i - j$ a zároveň $j - k$, tak potom vzniká aj takzvaná nepriama korelácia $i - k$, ktorá však v skutočnosti nemusí odrážať skutočný vzťah medzi pozíciami i a k. Riešeniu tejto problematiky sa venujeme v nasledujúcej podkapitole.

2.2.1 Matematické pozadie algoritmov GREMLIN-u

Vstupom pre GREMLIN je, podobne ako v SCA, súbor zarovnaných sekvencií proteínu. GREMLIN na rozdiel od SCA nevyužíva algebraické operácie na maticiach, ale svoje výpočty rieši pomocou globálnych štatistických modelov. Využíva princíp maximálnej entropie, čo znamená že je snaha dosiahnuť čo najvyššiu entropiu pri danom pravdepodobnostnom rozdelení [6].

Pravdepodobnostný model je definovaný pomocou diskrétnej náhodnej premennej X_i, ktorá predstavuje kompozíciu aminokyselín na pozícii i. Hodnoty, ktoré X_i nadobúda sú $1, \ldots, 21$, keďže existuje 20 aminokyselín a jedna medzera. Spolu máme teda 21 stavov X_i. Na to, aby sme reprezentovali celé naše zarovnanie s počtom stĺpcov L, označíme ho ako $\mathbf{X}$ a definujeme ako $\mathbf{X} = X_1, X_2, \ldots, X_L$. Pravdepodobnostný model, ktorý potom GREMLIN používa, je definovaný ako $P(\mathbf{X})$ a počíta pravdepodobnosť kompozície nášho zarovnania [25]:

$$P(\mathbf{X} = \mathbf{x}) = \frac{1}{Z} exp\left(\sum_{i=1}^{L}\left[\mathbf{v}_i(x_i) + \sum_{j>i}^{L}\mathbf{w}_{i,j}(x_i, x_j)\right]\right) \tag{2.16}$$

Vo vzorci pre $P(\mathbf{X})$ máme tri dôležité parametre. Prvým je pole $\mathbf{v}$, pričom jeho prvok $\mathbf{v}_i$ je množina pravdepodobností výskytu pre každú aminokyselinu na pozícii i v proteíne. Druhý parameter je matica $\mathbf{w}$. Jej prvkami sú taktiež matice $\mathbf{w}_{i,j}$, ktoré pre dvojicu pozícií i a j určujú ich vzájomnú koreláciu pre všetky dvojice aminokyselín. Znamená to teda, že $\mathbf{w}$ je matica matíc veľkosti 21×21 pre všetky dvojice pozícií v proteíne. Posledným je Z, čo je normalizačná konštanta, ktorá zabezpečuje to, že súčet všetkých pravdepodobností je 1. Zarovnanie reprezentované pomocou Markovového náhodného poľa, si môžeme v jednoduchosti predstaviť ako neorientovaný úplný graf, pričom má L vrcholov, teda jeden vrchol pre každú pozíciu, respektíve stĺpec. Keďže

hovoríme o úplnom grafe, hrany sú medzi všetkými dvojicami vrcholov. Súvislosť medzi touto reprezentáciou a parametrami $\mathbf{v}$ a $\mathbf{w}$ je v tom, že váha hrany medzi vrcholmi i a j je reprezentovaná maticou $\mathbf{w}_{i,j}$ a vrchol i má svoje parametre podľa poľa $\mathbf{v}_i$. Hodnoty v matici $\mathbf{w}_{i,j}$, resp. parameter hrany, určuje do akej miery vrcholy, teda pozície, spojené touto hranou korelujú. Čím vyššia je hodnota parametra, tým je korelácia silnejšia. Ak by sme mali vyjadriť, že medzi vrcholmi i a j nie je žiadna korelácia, všetky prvky $\mathbf{w}_{i,j}$ by boli nulové.

Cieľom pravdepodobnostného modelu je maximalizovať pravdepodobnosť nášho zarovnania, teda optimálne určiť hodnoty parametrov $\mathbf{v}$ a $\mathbf{w}$. Znamená to, že v našom grafe potenciálne chceme mať čo najvyššiu vierohodnosť, aby sme dosiahli čo najväčšiu pravdepodobnosť nášho zarovnania. Je to teda proces učenia sa samotnej štruktúry grafu a jeho parametrov. Prvým intuitívnym riešením by bolo maximalizovanie všetkých parametrov v grafe. Lenže to spôsobí vznik falošných a nepriamych korelácií, o ktorých vieme, že sa ich GREMLIN snaží odfiltrovať. Východiskom pre riešenie parametrizácie hrán aj vrcholov je takzvaná regularizácia [6].

Regularizácia modelu funguje nasledovným spôsobom. Počas procesu učenia sa pre dvojicu parametrov $\mathbf{v}$ a $\mathbf{w}$ spočíta hodnota R, ktorá predstavuje celkovú hodnotu penalizácie, ktorá sa od pravdepodobnosti zarovnania musí odčítať. Hodnota R sa získava ako súčet penalizácií za každú hranu, ktorú model vypočíta. Navyše, čím väčšiu hodnotu parametra v matici $\mathbf{v}$ alebo $\mathbf{w}$ chce model pridať, tým vyššiu penalizáciu za ňu musí „platiť “. On však vie, že chce dosiahnuť čo najvyššiu pravdepodobnosť pre zarovnanie, je preto v jeho záujme získať čo najnižšiu penalizáciu a zároveň najvyššiu pravdepodobnosť. Algoritmus penalizácie je nastavený tak, že pre tranzitívne hrany, ktoré by predstavovali nepriame korelácie, je výška penalizácie príliš vysoká, preto ju do grafu nepridá a namiesto nej pridá hrany priamych korelácií ([6], [25]).

V našej práci nás primárne zaujímajú hodnoty korelácií medzi dvojicami pozícií, ktoré sú obsiahnuté v matici $\mathbf{w}$. Táto matica prechádza ešte ďalšou korekciou, pričom hodnoty tejto upravenej matice predstavujú naše hľadané korelácie ([11], [25]). GREMLIN navyše škáluje tieto hodnoty tak, aby pre výsledné skóre platilo, že pokiaľ nadobúda hodnotu väčšiu ako 1, znamená to, že korelácia medzi pozorovanými dvoma pozíciami je značne vyššia ako priemerná korelácia medzi ostatnými.

2.2.2 Ďalšia práca s modelom

Okrem vysvetľovania vzdialených korelujúcich pozícií sa autori GREMLIN-u zaoberajú aj ďalšími záležitosťami, a to porovnávaním už existujúcich homologických štruktúr s takzvanými komparatívnymi modelmi, ktoré sú užitočné hlavne pri dizajnovaní nových proteínov a pri vytváraní terciárnych štruktúr. Dnes už poznáme veľa priestorových štruktúr proteínov a stretávame sa tak so skutočnosťou, že mnohé proteíny, ktoré sa

podrobujú analýzam, ako je GREMLIN, už vlastne majú stanovenú svoju 3D štruktúru. Vyvstáva nám preto otázka, na čo je vlastne dobré robiť predikcie kontaktov pozícií a pomocou nich zlepšovať takéto komparatívne modely, čo nanovo predpovedajú terciárnu štruktúru. Prvým dôvodom je overenie správnosti už nájdenej štruktúry. Vzhľadom na to, že väčšina nových štruktúr je taktiež predikovaná iba pomocou bioinformatických technológií, jej potvrdenie pomocou inej metódy je určite nápomocné a pridáva štruktúre vyššiu dôveryhodnosť. Faktom však ostáva, že ak sa komparatívny model mapuje na GREMLIN-om nájdené kontakty rovnako dobre ako už existujúca homologická štruktúra, nič nové sa nedozvieme a komparatívny model nie je možné ďalej zlepšiť. Avšak pokiaľ sa homologická štruktúra mapuje na nájdené kontakty lepšie ako doteraz nájdený komparatívny model, je zrejmé že vieme tieto informácie využiť na jeho zlepšovanie.

GREMLIN teda využíva aj metódy HHsearch a HHblits [45], pomocou ktorých vieme zistiť, ako sa predikcia kontaktov zhoduje s informáciami z už existujúcich štruktúr. Metóda HHblits nájde v databáze proteínov UniProt [47] také sekvencie, ktoré pridá do nášho zarovnania a vytvorí tak obohatené nové zarovnanie. Pomocou HHsearch sa z databázy PDB [8] nájdu homologické štruktúry pre náš referenčný pozorovaný proteín a vyrobí sa pre neho predpočítané zarovnanie a príslušný Skrytý Markovov Model. Predpočítané zarovnanie sa potom porovná s našim vstupným zarovnaním (upraveným pomocou HHblits). Podobnosť predpočítaného zarovnania a nášho zarovnania vyjadruje $HH\Delta$. Pokiaľ sú sekvencie týchto dvoch zarovnaní veľmi podobné, $HH\Delta$ má nízku hodnotu a znamená to, že pre náš referenčný proteín existuje dobrá 3D štruktúra a nepredpokladáme, že GREMLIN-om nájdené kontakty prinášajú nové informácie. Naopak $HH\Delta$ rastie s väčšou rozdielnosťou týchto dvoch zarovnaní. Maximálnu hodnotu 1 nadobudne, keď k pozorovanej sekvencii homologický vzor so štruktúrou neexistuje. Experimentálne sa zistilo, že $HH\Delta$ je smerodajná vtedy, ak platí, že $HH\Delta > 0.5$. Znamená to, že vo vypočítanej korelačnej mape je dokonca viac užitočných informácií o kontaktoch medzi aminokyselinovými zvyškami než v danej homologickej štruktúre [25].

2.2.3 GREMLIN a vzdialené koevolvujúce pozície

Ako už bolo spomenuté vyššie, autori GREMLIN-u sa snažia poukázať na to, že takmer všetky skutočne korelujúce dvojice sa dajú nájsť v priamom kontakte. Ako si GREMLIN vysvetľuje vzniknuté vzdialené koevolvujúce pozície, teda také, ktoré v štruktúre netvoria kontakt? Jedným z možných dôvodov je použitie chybnej štruktúry. Výskyt takýchto dvojíc klesá pri použití lepšieho rozlíšenia použitej štruktúry. Naopak, vysoký počet je možný pri proteínoch so štruktúrou s nízkym rozlíšením a v proteínoch s vysokým počtom repetitívnych sekvencií.

V tomto prípade totiž symetria týchto sekvencií zvyšuje vznik korelácií medzi aminokyselinami, ktoré ale v skutočnosti nie sú v kontakte. Preto sa takýmto proteínom s vysokým počtom opakujúcich sa sekvencií chceme vyhnúť [3]. Stále sa však v niektorých prípadoch objavujú skutočne korelujúce dvojice napriek tomu, že v štruktúre sú vzdialené. Ich pôvod je potrebné hľadať v biologickom pozadí vlastností a správania sa proteínov.

Prvé vysvetlenie sa opiera o také formy proteínov, ktoré sú tvorené viacerými monomérmi. Je dokázané, že napriek veľkým vzdialenostiam korelujúcich dvojíc v samostatných monoméroch, sú nájdené ich fyzické kontakty práve medzi monomérmi navzájom. Ich interakcie sú najčastejším dôvodom vzniku vzdialených korelácií. Druhé vysvetlenie sa odvoláva na štruktúrnu variabilitu proteínov. Vďaka nej sa môže stať, že korelujúca dvojica sa v jednej namapovanej štruktúre danej proteínovej rodiny môže javiť ako vzdialená a v inej štruktúre tej istej rodiny ako priamy kontakt. Zároveň s ňou súvisia aj konformačné zmeny proteínu, ktoré sa nedajú vidieť v jednej štruktúre všetky naraz. To znamená, že ak máme dve štruktúry pre jeden proteín v dvoch rôznych konformačných stavoch, je možné, že v jednej štruktúre bude koevolvujúca dvojica vzdialená a v druhej bude mať menšiu vzdialenosť alebo až tvoriť kontakt.

GREMLIN vo svojich výpočtoch zahŕňa aj takéto vysvetliteľné vzdialené korelujúce dvojice a majú skóre približne také vysoké, ako správne nájdené v samostatných monoméroch. Stále však ostávajú tie, ktoré nie sú nijakým spôsobom vysvetliteľné. GREMLIN im však pridá relatívne nízke skóre a budú tak považované za nie úplne relevantné. Práve takýmto spôsobom sa GREMLIN snaží viac menej takéto dvojice ani nepripúšťať [3].

2.2.4 Výsledky, ktoré GREMLIN ponúka

Pre získanie výsledkov z tejto metódy je možné použiť pracovné webové rozhranie http://gremlin.bakerlab.org. Ponúka vyhotovenie GREMLIN analýzy priamo v prehliadači s prehľadným výstupom vizuálne zaujímavých výsledkov. Jedným z nich je napríklad spomínaná korelačná mapa, ktorá pre vybrané dvojice aminokyselinových zvyškov proteínu označí ich mieru koevolúcie. Podstatnými výsledkami sú aj porovnania s už existujúcimi atomickými štruktúrami spolu s hodnotou $HH\Delta$ a s vizuálnou mapou prekrytia modelovej štruktúry a homologickej štruktúry z PDB.

2.3 Porovnanie metód

V predošlých podkapitolách sme predstavili dve z mnohých bioinformatických metód, ktoré sa zaoberajú analýzou proteínov na úrovni primárnej štruktúry. Zistili sme, že obe pracujú len na základe informácií získaných zo vstupného zarovnania, GREMLIN

však toto zarovnanie dodatočne obohacuje o nové sekvencie. Obe metódy skúmajú konzervovanosť každej pozície a následne vyhodnocujú mieru koevolúcie dvojíc pozícií. Používajú výpočtové techniky z oblasti pravdepodobnosti a štatistiky. SCA vo svojich krokoch pre stanovenie konzervovanosti pozícií používa entropiu, pri ďalších výpočtoch zahŕňa hlavne operácie s maticami. GREMLIN využíva pravdepodobnostné modely a na rozdiel od SCA na výstup vráti iba určitý počet štatisticky najvýznamnejších koevolvujúcich dvojíc. SCA pokračuje hľadaním nezávislých komponentov a smeruje ku nájdeniu proteínových sektorov, zatiaľ čo GREMLIN počíta pravdepodobnosti kontaktu dvojíc pozícií, mapuje a porovnáva svoje výsledky s existujúcimi štruktúrami.

Nás zaujíma ako a či môžeme výsledky medzi nimi prepájať a porovnávať. Vidíme, že SCA sa skôr zameriava na hľadanie „neviditeľných“ subštruktúr - sektorov a z týchto informácií získať nové poznatky o spolupráci pozícií a ich aminokyselín, naopak GREMLIN predikuje kontakty v novej štruktúre. Pri porovnávaní výsledkov oboch metód sa môžeme bližšie pozrieť na to, či pozície v sektoroch, ktoré sme našli vďaka SCA, tvoria fyzické kontakty aj podľa metódy GREMLIN. Ak totiž tieto pozície nenájdeme v zozname blízkych kontaktov, znamená to, že v sektore máme pozície ktoré spolu síce koevolvujú, ale nie sú v priamom kontakte, čo by potom GREMLIN samozrejme spochybnil. Tým, že aj koevolúcia medzi dvojicami pozícií je počítaná odlišne, je zaujímavé sa pozrieť na porovnanie týchto hodnôt a zistiť, či je medzi nimi nejaké prepojenie, resp. súvis. Na túto tému nadviažeme v Kapitole 3.

Medzi ďalšie metódy, s ktorými však v tejto práci nepracujeme, patria napríklad DCA (z angl. *Direct coupling analysis*) a PSICOV (z angl. *Protein Sparse Inverse COVariance*), ktoré sa obe, podobne ako GREMLIN, zameriavajú na vyhľadávanie fyzických kontaktov medzi pozíciami a predikciu terciárnej štruktúry ([23], [31], [49]).

Kapitola 3

Analýza proteínu PARP

V nasledujúcej kapitole sa budeme venovať praktickej časti bakalárskej práce. Vysvetlíme postup získavania dát, ich následného spracovania a popíšeme použitie jednotlivých metód. Zároveň zanalyzujeme, vizualizujeme a porovnáme získané výsledky.

3.1 Zisk a spracovanie dát

Ako každej dátovej analýze, aj tejto predchádzala podrobná a precízna príprava dát. Základom našich vstupných zarovnaní boli voľne prístupné sekvencie katalytickej domény PARP z verejnej databázy Pfam [29]. Tieto sekvencie boli vybraté z organizmov, pre ktoré bol známy celý ich proteóm. Celý zoznam týchto sekvencií domény proteínu PARP obsahuje niečo vyše dvanásťtisíc záznamov a v ďalšom texte sa na neho odvolávame ako na rodinu.

Vzhľadom na to, že naša analýza bola zameraná na PARP domény získané práve z kvasiniek, ich sekvencie tvorili podstatnú zložku pripravovaných vstupných dátových množín. Nájdené sekvencie proteínu PARP z genómov rôznych druhov kvasiniek nám poskytol pán profesor Jozef Nosek z Prírodovedeckej fakulty Univerzity Komenského. Pretože nás zaujímala iba katalytická doména tohto proteínu, potrebovali sme ju v každej sekvencii identifikovať. Sekvencie kvasinkových proteínov však neboli úplne totožné, teda bolo treba hľadať pozície domény pre každú sekvenciu zvlášť. Na to sme použili bioinformatický nástroj *hmmsearch*, ktorý využíva Skryté Markovove Modely na nájdenie polohy všetkých domén, ktoré sa v danej sekvencii nachádzajú [16]. Výsledkom *hmmsearch* bol zoznam týchto súradníc priradených k jednotlivým kvasinkám. Na následné vystrihnutie tejto domény sme použili nástroj *bedtools*, ktorý podľa tohto zoznamu vybral správny podreťazec zo sekvencie prislúchajúcej kvasinky [36]. Všetky takto získané sekvencie katalytických domén sme si rozdelili na dve samostatné skupiny: (1) sekvencie všetkých rodov kvasiniek spolu a (2) sekvencie iba z kvasiniek rodu Yarrowia. V tejto fáze sme mali 2 súbory kvasinkových domén proteínu PARP.

S týmito množinami sme pracovali vo všetkých krokoch zvlášť. V ďalšom texte sa pre prehľadnosť odvolávame na iba jednu z nich (ďalej len kvasinkové sekvencie).

Keďže kvasinkových sekvencií bolo na dátovú analýzu príliš málo, potrebovali sme k nim pridať aj nejaké iné sekvencie z rodiny proteínu PARP. Tým, že nás zaujímali výsledky predovšetkým zamerané na tie kvasinkové, prispôsobili sme tomu celú vstupnú dátovú množinu tak, že z celej rodiny sme vyberali iba tie sekvencie, ktoré sú kvasinkovým aspoň na 30 % podobné (ďalej len podrodina). Na hľadanie podobných sekvencií z rodiny pre každú kvasinku sme využili nástroj *blastp* ([1], [9]), pričom ako databázu sme použili kvasinkové sekvencie a zarovnávali sme k nej sekvencie z rodiny. Z výsledného zoznamu, kde sme mali ku každému identifikátoru z rodiny priradený identifikátor kvasinkovej sekvencie s najlepším skóre, sme vybrali tie, pre ktoré platilo, že kvasinková sekvencia mala s danou sekvenciou z rodiny zhodu aspoň 30 %, E-hodnota, teda pravdepodobnosť výskytu rovnakého skóre zarovnania náhodnej sekvencie s náhodnou databázou, nesmela prekročiť prah 0,001 a skóre podobnosti muselo byť aspoň 50 %. K vybraným identifikátorom sekvencií z rodiny sme prislúchajúce sekvencie priradili pomocou *faSomeRecords* [15]. Z tejto množiny sme vyberali 500 náhodných, z ktorých nám vznikol prvý vstupný súbor. Ďalší sme získali tak, že sme k vybratým záznamom z podrodiny pridali príslušné kvasinkové sekvencie. V tomto štádiu prípravy dát sme mali 2 dátové množiny: jednu, ktorú tvorilo 500 sekvencií z podrodiny (ďalej len podrodinný vstup) a druhú, kde sme mali navyše pridané aj kvasinkové sekvencie (ďalej len kvasinkový vstup). Museli sme ale myslieť na to, že pre správne vizualizovanie výsledkov potrebujeme mať vo vstupnej množine aspoň jednu sekvenciu, ktorá má stanovenú 3D štruktúru. V našom prípade to bola sekvencia s identifikátorom PARP1_HUMAN, keďže kvasinkovým sekvenciám bola najpodobnejšia spomedzi všetkých PARP proteínov. Do oboch vstupných množín sme túto sekvenciu ručne pridali. Výsledné množiny všetkých sekvencií vo formáte FASTA sme zarovnali pomocou nástroja pre viacnásobné zarovnávanie *muscle* ([14], [28]). Takto sme dostali dve konečné zarovnania s približne 500 a 520 sekvenciami, ktoré zároveň predstavovali dva vstupné súbory pre nasledujúce analýzy.

Pripomíname, že tento postup sme zopakovali pre množinu sekvencií všetkých kvasiniek spolu a samostatne pre množinu sekvencií kvasiniek rodu Yarrowia. Spolu sme potom mali 4 vstupné súbory. Sekvencie si boli v správnej miere podobné (ideálne v rozmedzí 20 % až 40 %) a so zarovnaním sa dalo ďalej pracovať. Mohli sme však pozorovať, že vo výslednom zarovnaní boli aj takmer rovnaké sekvencie (zhoda 80 % až 100 %), čo bolo spôsobené tým, že práve kvasinkové sekvencie sú si veľmi podobné. V oboch použitých metódach bola takáto zhoda ošetrená pomocou pridávania váh (viď. Kap. 2). Podobnosť nižšiu ako 20 % sme spozorovali napriek stanovenému minimu identity, napríklad kvôli prítomnosti pridanej sekvencie PARP1_HUMAN, ktorá nemusela mať s ostatnými sekvenciami z rodiny príliš veľkú podobnosť (viď. Obr. 3.2).

Príčina tejto nízkej zhody však mohla vznikať aj tým, že niektoré kvasinkové sekvencie alebo sekvencie z rodiny si navzájom nemuseli byť podobné.

Navyše, ako piaty vstupný súbor, sme si ešte pripravili množinu úplne náhodných sekvencií z rodiny, ktorá nebola špecializovaná pre kvasinkové sekvencie. Tento vstup slúžil aj pre porovnávanie jeho výsledkov s výsledkami z oboch kvasinkových vstupov, ale hlavne na nezávislé vizualizovanie priebehu SCA analýzy v Kapitole 2. Aby sme čitateľovi uľahčili porozumenie tejto pomerne zložitej prípravy dát, uvádzame jej schematický postup a prehľadovú tabuľku s názvami všetkých vstupných súborov a krátkym vysvetlením ich obsahu (viď Tab. 3.1, Obr. 3.1).

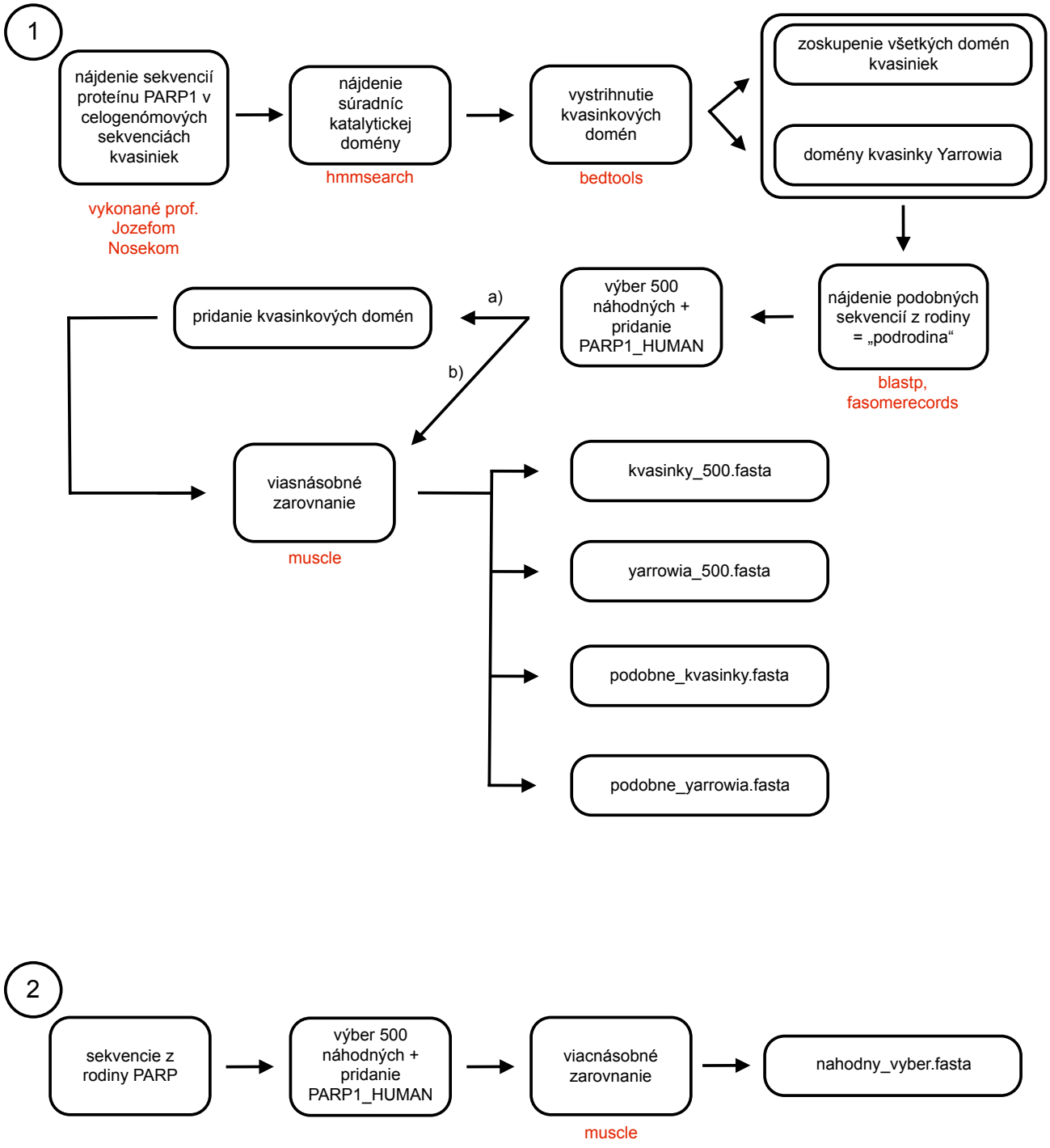
Skupina	Názov	Obsah
Kvasinkové vstupy	kvasinky_500	sekvencie všetkých kvasiniek s 500 sekvenciami príslušnej podrodiny
	yarrowia_500	sekvencie kvasiniek rodu Yarrowia s 500 sekvenciami príslušnej podrodiny
Podrodinné vstupy	podobne_kvasinky	500 sekvencií z podrodiny prislúchajúcej všetkým kvasinkám
	podobne_yarrowia	500 sekvencií z podrodiny prislúchajúcej kvasinkám rodu Yarrowia
Náhodný vstup	nahodny_vyber	500 náhodne vybratých sekvencií z celej rodiny

Tabuľka 3.1: Tabuľka vstupných dátových množín

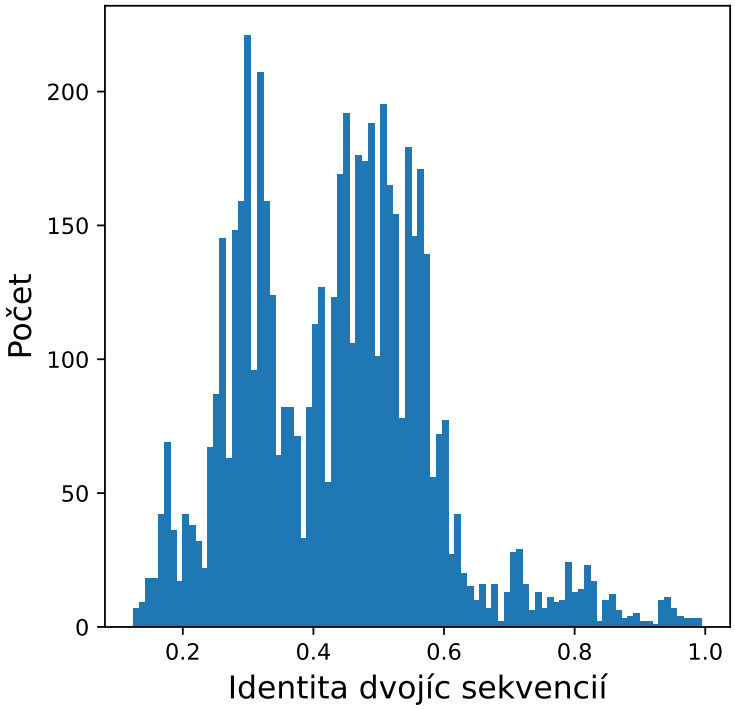
3.2 Výsledky SCA analýzy

V tejto podkapitole podrobne rozoberieme výsledky jednotlivých krokov SCA analýzy pre proteínovú doménu PARP. Pre vizualizáciu dát sme použili už spomínanú terciárnu štruktúru 1WOK (viď. Kap. 1). Obrázky boli vytvorené pomocou knižnice programovacieho jazyka Python - Matplotlib a programu PyMOL ([22], [44], kód viď. Dodatok B: elektronická príloha: Kody/printResultsSK.py). Analýzu SCA sme vykonali na všetkých vstupných množinách spomínaných v predošlej podkapitole (viď. Tab. 3.1).

Program SCA sme získali z verejného repozitáru GitHub ([18], [39]). SCA analýza bežala v piatich výpočtových krokoch, pričom každý krok sme spúšťali pomocou samostatného skriptu z príkazového riadka.



Obr. 3.1: Prehľadová schéma prípravy vstupných zarovnaní



Obr. 3.2: Histogram, znázorňujúci podobnosť všetkých dvojíc sekvencií vo vstupnom zarovnaní kvasinky_500.

Prvým krokom bola anotácia sekvencií, za ktorou nasledovalo upravovanie a filtrovanie zarovnania. Tretím krokom bolo počítanie konzervovanosti pozícií a korelácií dvojíc a posledným krokom bola identifikácia nezávislých komponentov. Výsledky jednotlivých krokov sa ukladali do databázového súboru, z ktorého sa dali použiť dáta na ich porovnávanie a vizualizáciu. Skript na zhotovenie obrázkov sme vytvorili pospájaním a upravením existujúcich kódov pre vykreslenie jednotlivých výsledkov, vizualizáciu niektorých obrázkov sme vytvorili navyše (kód viď. Dodatok B: elektronická príloha: Kody/printResultsSK.py, [40]). Pre porovnávanie získaných nezávislých komponentov sme vytvorili skript (kód viď. Dodatok B: elektronická príloha: Kody/compareICs.py), ktorého vstupom boli dve výsledné databázy, ktorých komponenty sme chceli porovnať. Spúšťanie programu compareICs.py bolo realizované opäť pomocou príkazového riadka. Pri porovnávaní výsledkov jednotlivých vstupov sme sa zamerali na samotné pozície, ktoré SCA zaradila do nejakého nezávislého komponentu. Pozorovali sme, ktoré pozície sa objavili medzi tými, čo tvorili komponenty pred pridaním kvasinkových sekvencií a po ich pridaní ku podrodinám. Kompletný zoznam týchto pozícií je možné nájsť v elektronickej prílohe (viď. Dodatok B: elektronická príloha: Tabulky/ porovnania_pozicii_vsetkych_NK.xlsx). Takéto pozície v ďalšom texte označujeme ako „nové pozície“. Pre vizuálnu predstavu ich rozmiestnenia v rámci domény ich vieme mapovať aj na jej 3D štruktúru (viď. Obr. 3.3). Výsledkom porovnávania sa venujeme v Podkapitole 3.3.2.

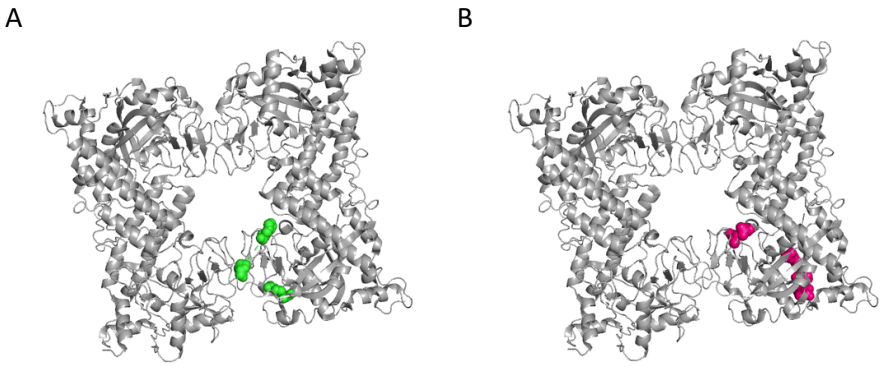
3.2.1 Úprava zarovnania

Úprava zarovnania v programe SCA prebiehala vo viacerých fázach. Začala odstraňovaním slabo zastúpených pozícií, teda stĺpcov zarovnania, kde bolo viac ako 80 % medzier. Nasledovalo filtrovanie sekvencií, ktoré obsahujú viac ako určité percento medzier (> 20 %), ďalej sekvencií, ktoré boli príliš podobné referenčnej (> 80 % zhoda) alebo boli od nej príliš vzdialené (zhoda < 20 %). Úprava zarovnania končila pridaním váh jednotlivým sekvenciám. Takto sme získali spracované vstupné zarovnania, pričom v ďalších krokoch SCA sa pracovalo už len s takýmito upravenými zarovnaniami.

Po dôslednom preštudovaní získaných zarovnaní sme však zistili, že v prípade vstupu yarrowia_500 sa nám do spracovaného zarovnania nedostali žiadne príslušné kvasinkové sekvencie. Rozhodli sme sa preto preskúmať tento problém, keďže sme vyslovene chceli mať kvasinkové sekvencie zahrnuté v ďalších výpočtoch. Prišli sme na to, že kvasinkové sekvencie sa do spracovaného zarovnania nedostanú práve kvôli vysokému percentu medzier. Po zanalyzovaní jednotlivých percent medzier pre všetky sekvencie vstupu sme zistili, že ho kvasinkové sekvencie nadobúdajú v zaokrúhlenej výške 20,05 %. Dôvodom relatívne vysokej hodnoty môže byť to, že kvasinkové sekvencie sú o niečo kratšie ako ostatné sekvencie v pôvodnom zarovnaní. Rozhodli sme sa preto mierne zasiahnuť do programu a zvýšiť prah tolerancie medzier v jednotlivých sekvenciách na 20,05 % pre vstup yarrowia_500, čím sme do spracovaného zarovnania získali väčšinu pridaných kvasinkových sekvencií. Rovnaký prah sme použili aj pri vstupnom súbore podobne_yarrowia, keďže výsledky týchto vstupov sme detailne porovnávali. Pri ostatných vstupných súboroch sme prah ponechali na 20 %, ako bol pôvodne nastavený. Prehľad o počtoch sekvencií a pozícií pred a po spracovaní zarovnania ako aj iné užitočné informácie uvádzame v Tabuľke 3.2.

Vstupné zarovnanie	S/P pred	S/P po	M_{eff}
kvasinky_500	522/574	453/188	238
podobne_kvasinky	501/555	409/189	209
yarrowia_500	513/529	449/188	205
podobne_yarrowia	499/584	432/192	203
nahodny_vyber	501/498	107/185	66

Tabuľka 3.2: Tabuľka spracovania jednotlivých zarovnaní. Stĺpce: (1) názov vstupného zarovnania, (2) počet sekvencií/počet pozícií pred filtrovaním a úpravou zarovnania (3) počet sekvencií/počet pozícií po filtrovaní a úprave zarovnania, (4) počet efektívnych sekvencií.



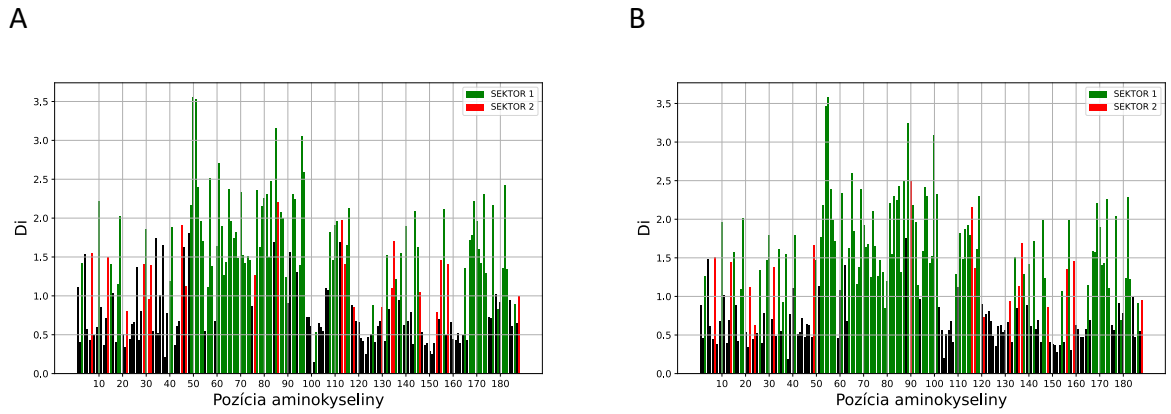
Obr. 3.3: 3D vizualizácie štruktúry PARP domény. Farebne sú označené „nové pozície“, ktoré boli navyše nájdené ako súčasť niektorého z nezávislých komponentov po pridaní kvasinkových sekvencií do podrodinných vstupov. „Nové pozície“ v súbore (A) kvasinky_500 v porovnaní s podobne_kvasinky a (B) yarrowia_500 v porovnaní s podobne_yarrowia.

3.2.2 Konzervovanosť pozícií

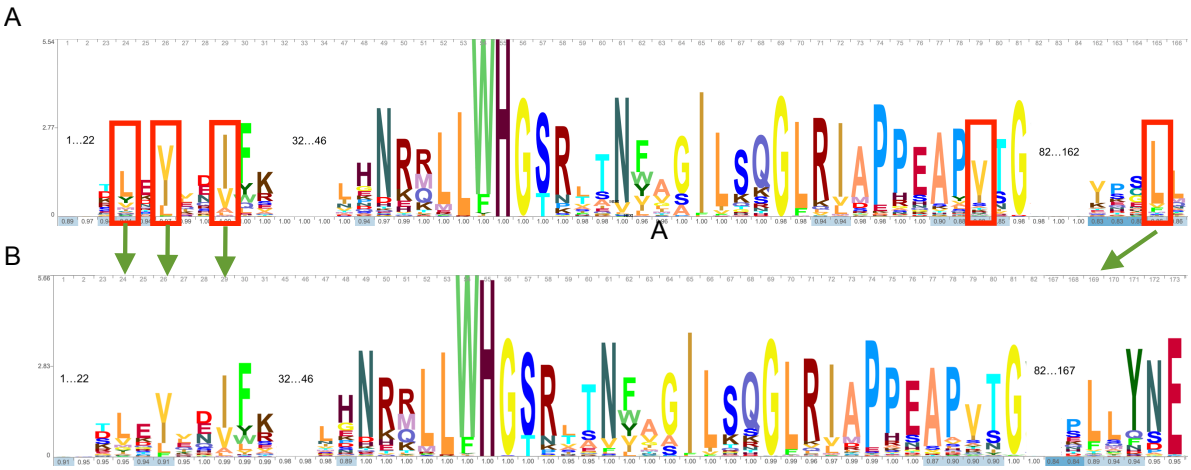
Konzervovanosť každej pozície bola kľúčovou informáciou v SCA, podstatnou pre jej ďalšie výpočty. Pre celé zarovnanie zo vstupnej množiny sme znázornili pomocou histogramu (viď. Obr. 3.4). Keďže hovorila o zachovaní určitej aminokyseliny na danej pozícii, dalo sa očakávať, že výšky jednotlivých stĺpcov budú úmerne zodpovedať výškam písmen aminokyselín v logu Skrytých Markovových Modelov (ďalej len HMM logo, z angl. *Hidden Markov Model*) pre dané zarovnanie. Pri vstupe yarrowia_500 sme mohli takýto jasný súvis pozorovať v okolí pozícií 48 až 60, navyše sa výšky písmen relatívne úmerne menili aj v originálnom HMM logu proteínu PARP ([29], viď. Obr. 3.4 A, Obr. 3.5 A, Obr. 3.7). Obdobne pri vstupe kvasinky_500, avšak pri porovnaní s originálnym HMM logom sme videli posun o 3 pozície práve v spomínanej oblasti (viď. Obr. 3.4 B, Obr. 3.6 A). V oboch prípadoch sme si všimli aj mierne zmeny písmen v porovnaní s originálnym HMM logom, čo však bolo spôsobené práve cieleným výberom podobných sekvencií.

Našu pozornosť pri vyhodnocovaní konzervovanosti sme upriamili na „nové pozície“, teda na rozdiely medzi množinami kvasinky_500 (resp. yarrowia_500) a podobne_kvasinky (resp. podobne_yarrowia). Tým, že sme v oboch prípadoch pracovali s podrodinou a jej podobnými kvasinkovými sekvenciami, ich výsledky boli podobné, odlišovali sa však v niektorých pozíciách a stĺpce sa o niekoľko pozícií posunuli. Príčinou týchto rozdielov bolo samotné pridanie kvasinkových sekvencií, a teda aj zmena samotných stĺpcov, z čoho bola ďalej odvodená zmena upravovania vstupného zarovnania, ktorá v konečnom dôsledku viedla ku iným hodnotám konzervovanosti, posunom stĺpcov a ku rozdielnemu zachovaniu aminokyselín. Ak sa bližšie zameriame na „nové

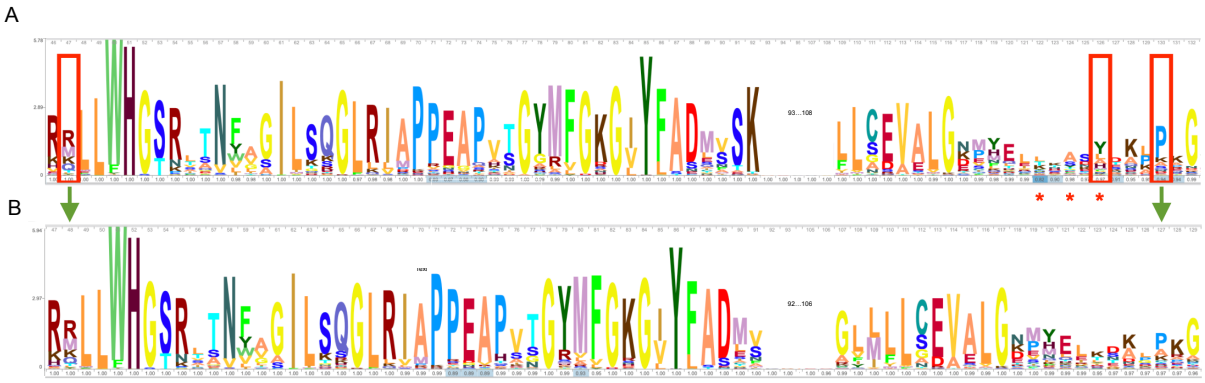
pozície“, ku každej vieme nájsť príslušnú pozíciu v logu podrodiny (viď. Obr. 3.5). Stal sa ale aj prípad, že „nová pozícia“ nájdená v kvasinkovom vstupe zo zarovnania príslušného podrodinného vstupu úplne vypadla (viď. Obr. 3.6). Pre lepšiu čitateľnosť uvádzame iba zaujímavé výseky spracovaných HMM log. Kompletné HMM logá je možné nájsť v elektronickej prílohe (viď. Dodatok B: elektronická príloha: HMM_logos/).



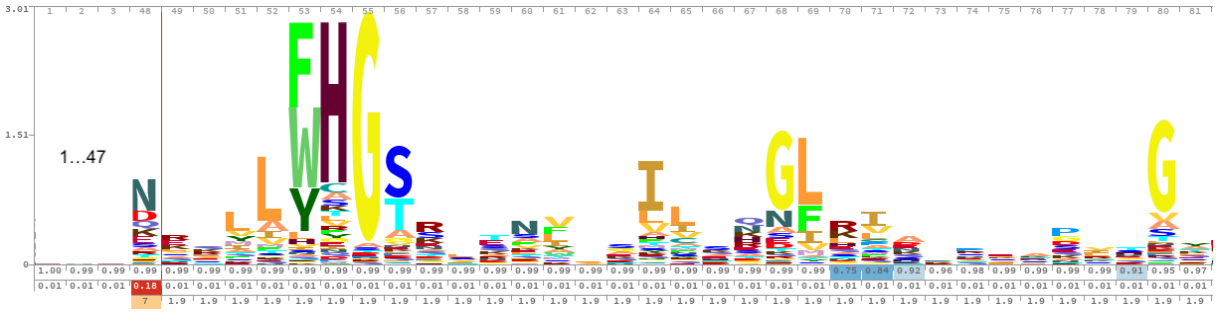
Obr. 3.4: Histogramy kozervovanosti pre množiny (A) kvasinky_500 a (B) yarrowia_500.



Obr. 3.5: Výseky z HMM log domény PARP zo spracovaného zarovnania (A) yarrowia_500 a (B) podobne_yarrowia. V červených obdĺžnikoch sú označené „nové pozície“, ktoré boli nájdené v yarrowia_500. Zelená šípka smeruje ku prislúchajúcim pozíciám zarovnania podobne_yarrowia.



Obr. 3.6: Výseky z HMM log domény PARP zo spracovaného zarovnania (A) kvasinky_500 a (B) podobne_kvasinky. Pozície, ktoré sme našli v podobne_kvasinky nenašli vôbec, sú označené hviezdičkou. Vidíme, že medzi nimi je aj jedna „nová pozícia“.



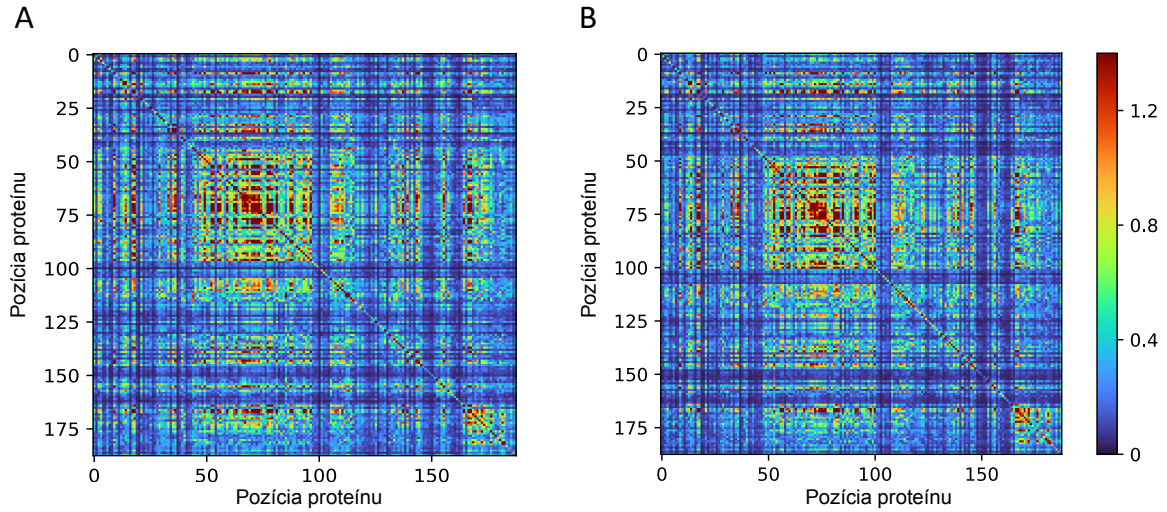
Obr. 3.7: Výsek z HMM loga pre katalytickú proteínovú doménu PARP z databázy Pfam [29].

3.2.3 Koevolúcia ako nástroj na nájdenie nezávislých komponentov

Vzťahy - korelácie medzi dvojicami pozícií boli ďalším podstatným výpočtom SCA. Vyjadrovali ako sú dvojice pozícií koevolvované (viď. Kap. 1). Korelačnú maticu konzerovanosti sme znázornili pomocou tepelnej mapy (viď. Obr. 3.8). Ako sme však spomínali v Kapitole 2, určeniu pozícií, ktoré tvoria sektor, predchádzalo hľadanie skupín nezávislých pozícií. Dozvedeli sme sa, že aj v prípade analýzy domény PARP nie je dekompozícia matice podľa vlastných hodnôt úplne postačujúca, a preto bola potrebná ICA, ktorá lepšie determinovala príslušnosť jednotlivých pozícií do takzvaných nezávislých komponentov (viď. Obr. 3.9, Obr. 3.10).

Podľa získaných obrázkov bolo zjavné, že prerozdelenie pozícií do nezávislých komponentov bolo jednoznačnejšie ako do vlastných vektorov. Body - pozície sa v grafe zoskupovali do tvaru písmena „L“, čo znamenalo, že pre jeden nezávislý komponent mala pozícia vysoké skóre príslušnosti a pre druhý nízke. Mohli sme však pozorovať, že aj pri nezávislých komponentoch, hlavne vo vyšších dimenziách, bolo aj toto rozdelenie podobne zašumené ako pri vektoroch. Prekrývanie jednotlivých bodov sa spôsobilo

hlavne tým, že grafy boli dvojrozmerné (viď. Obr. 3.9, Obr. 3.10). Hodnoty príslušnosti pozícií do daného nezávislého komponentu bolo možné mapovať na Studentovo rozdelenie, pričom skóre pozícií, ktoré boli nakoniec vybraté do samotných nezávislých komponentov, museli byť vyššie ako 95 % nájdených skóre pre daný komponent (viď. Dodatok A: Obr. A2, Obr. A3).

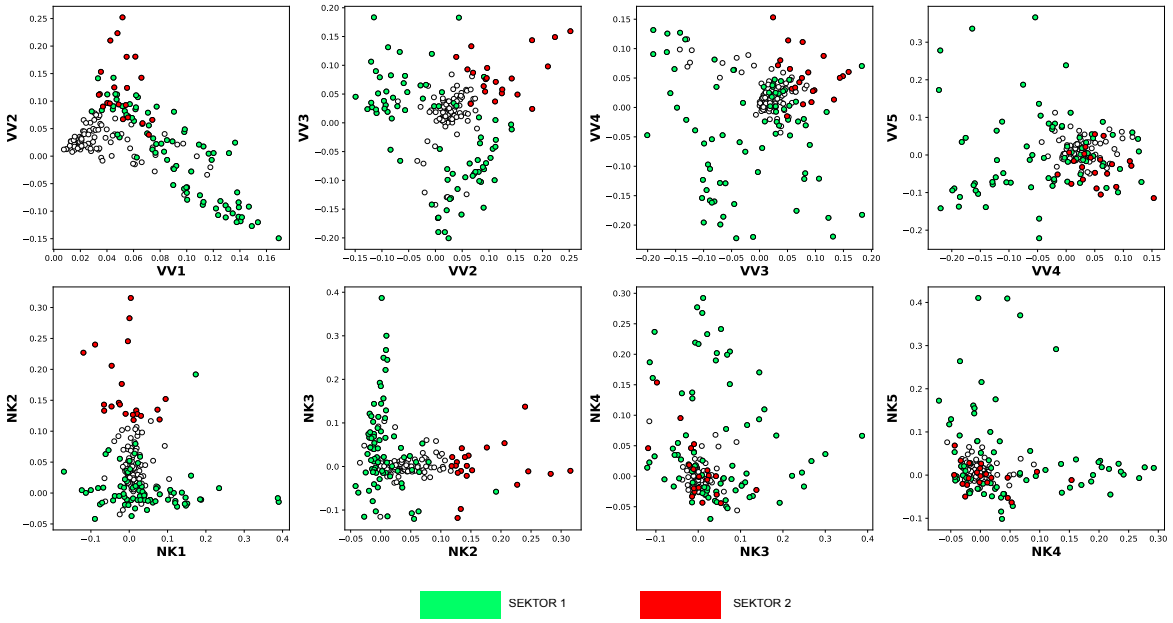


Obr. 3.8: Tepelné mapy korelačnej matice získané zo vstupných súborov (A) kvasinky_500 a (B) yarrowia_500. Miera zafarbenia určuje korelačnú silu medzi dvojicou pozícií. Čím je farba červenšia, tým je dvojica viac korelovaná. Na mapách je vidieť, že koevolvujúce dvojice môžu byť nakumulované blízko seba (oblasť medzi pozíciami 50 až 100), ale aj v relatívne veľkej vzdialenosti (okolie pozície 50 s oblasťou pri pozícii 170).

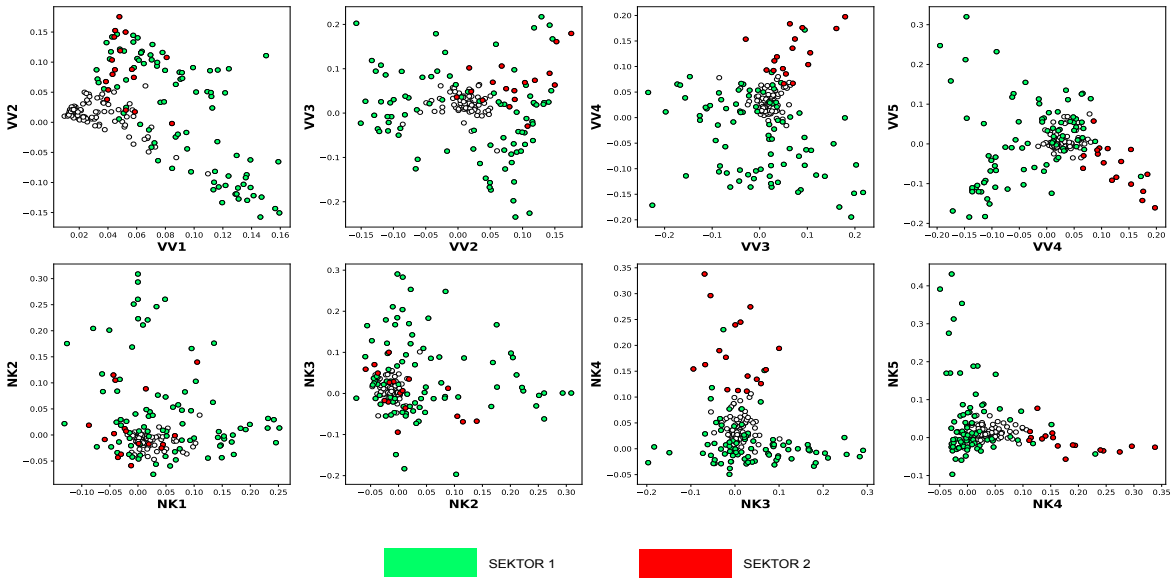
3.2.4 Cesta od nezávislých komponentov až ku sektorom

Napriek tomu, že nezávislé komponenty predstavovali štatisticky samostatné skupiny pozícií, stále platilo, že pozície naprieč komponentami mohli byť závislé. Preto sme jednotlivé nezávislé komponenty preskupili do väčších celkov - proteínových sektorov. Pozície priradené do nezávislých komponentov sme opäť vizualizovali pomocou tepelných máp, prvotne sme ich zoradili podľa sily príslušnosti k danému komponentu, od prvého po posledný. Výhodou týchto tepelných máp bola najmä priamočiara vizuálna práca s nimi. Bolo zreteľne vidieť, ktoré nezávislé komponenty boli na sebe navzájom závislé vďaka koreláciám pozícií naprieč komponentami (viď. Obr. 3.11).

Otázkou bolo, podľa čoho nezávislé komponenty pospájať. Vytváranie sektorov zatiaľ nie je algoritmicky vyriešené, a preto bolo nutné tento krok robiť ručne.



Obr. 3.9: Porovnanie zaradenia pozícií do (A) vlastných vektorov (VV) -vrchný rad grafov a do (B) nezávislých komponentov(NK) - spodný rad, pre vstup kvasinky_500.



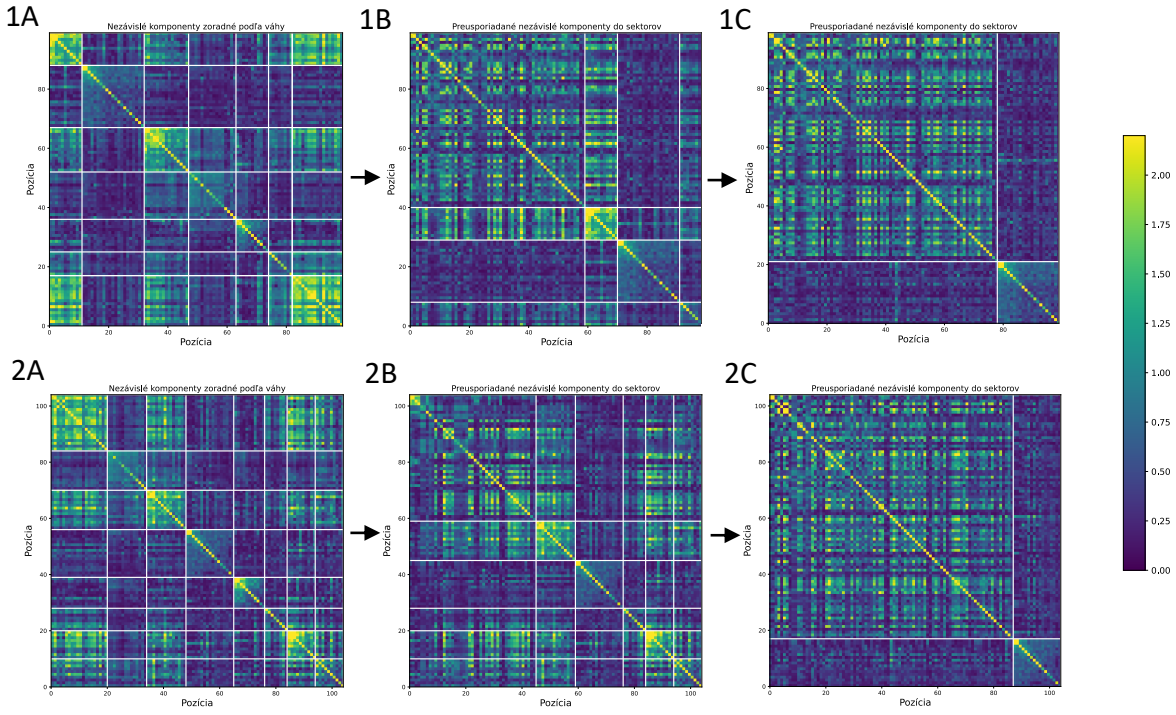
Obr. 3.10: Porovnanie zaradenia pozícií do (A) vlastných vektorov (VV) - vrchný rad grafov a do (B) nezávislých komponentov (NK) - spodný rad, pre vstup yarrowia_500. V tomto prípade si môžeme všimnúť, že hoci sa pri prvých dvoch dimenziách nezdá byť prerozdelenie do nezávislých komponentov oveľa lepšie, pri porovnaní tretej až piatej dimenzie je prerozdelenie do komponentov výrazne účinnejšie.

Mohlo sa postupovať priamo podľa tepelnej mapy a spájať nezávislé komponenty podľa pozícií, ktoré prejavujú vysokú koreláciu s pozíciami z iných komponentov. Pomohol by tiež expertný pohľad biológov, ktorí majú pozície a aminokyseliny detailnejšie preskúmané a vedia, ktoré z nich by mohli byť súčasťou jedného sektora a potenciálne sa podieľať na jednej funkcií proteínovej domény. Podľa takýchto pozícií sme sa riadili aj v našej analýze. V databáze UniProt [47] sme našli proteín PARP a vyhľadali pozície, ktoré plnia katalytickú funkciu jeho katalytickej domény. Nezávislé komponenty, ktoré tieto pozície obsahovali, sme spojili do jedného sektora a ostatné sme ponechali ako samostatné sektory (viď. Obr. 3.11 - 1B, 2B). Všimli sme si, že niektoré sektory ukazovali značnú závislosť s inými, a preto sme ich preskupili do nových celkov. Takto sme získali sektor, v ktorom sú umiestnené pozície, zúčastňujúce sa katalytickej aktivity domény a samostatné nezávislé sektory, ktorých funkciu zatiaľ určiť nevieme (viď. Obr. 3.11 - 1C, 2C). Sektory sme vizualizovali aj na 3D štruktúre proteínu, kde sme pozorovali ich umiestnenie vzhľadom na ostatné pozície, vedeli sme sa pozrieť, či sú to súvislé celky, alebo tvoria sektor aj fyzicky vzdialenejšie pozície (viď. Obr. 3.12).

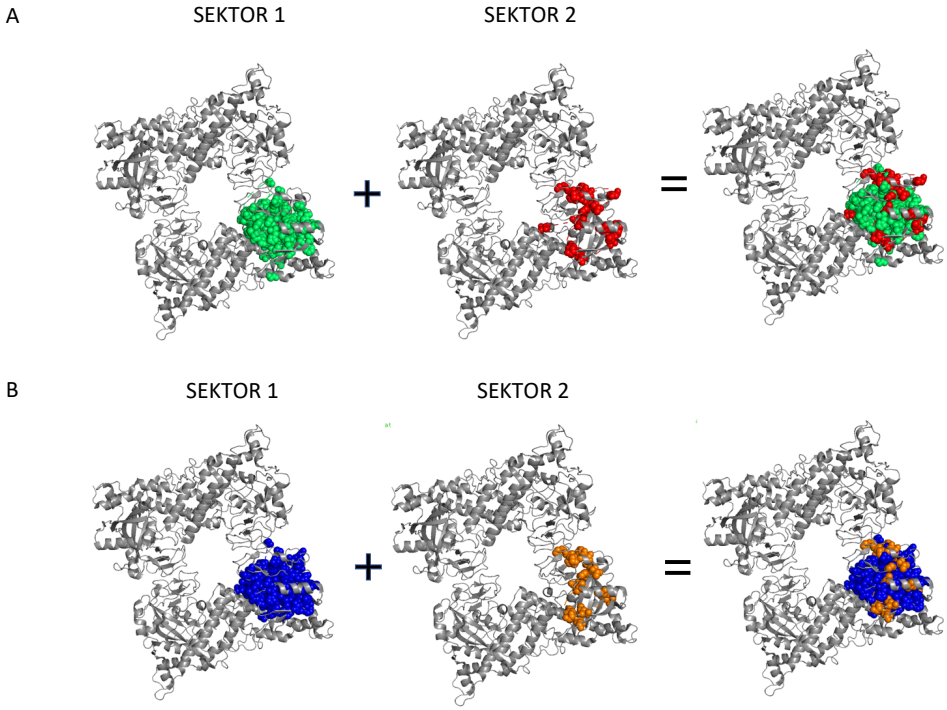
3.3 Sektory - indikátory zaujímavých pozícií

Zoskupenie nezávislých komponentov do sektorov primárne podľa pozícií, vykonávajúcich katalytickú funkciu malo svoj význam. Keďže pozície, ktoré tvorili jeden sektor, koevolvovali, teda sa vzájomne ovplyvňovali a spolupracovali, mohli sme si dovoliť predpokladať, že tento sektor bude zabezpečovať hlavnú katalytickú funkciu PAPR domény. Keďže bol tento sektor v oboch našich prípadoch väčší, boli v ňom od hlavných katalytických pozícií aj vzdialenejšie pozície, ktoré by mohli byť zaujímavé z toho pohľadu, že vplývajú na pozície vykonávajúce katalytickú funkciu, ale nie sú s nimi priamo v kontakte. Ostali nám potom zoskupené pozície do ostatných sektorov, ktorým sme funkciu nevedeli momentálne priradiť.

Pozície, ktoré tvoria samotné sektory, ako aj všetky 3D vizualizácie všetkých nezávislých komponentov, resp. sektorov a ostatné výstupné obrázky pre všetky vstupy z Tabuľky 3.1 prikladáme v elektronickej prílohe (viď. Dodatok B: elektronická príloha: SCA_OUTPUT/). SCA analýza v tomto kroku pre nás skončila. Zistili sme, do akej miery sú pozície konzervované, získali sme skupiny pozícií - nezávislé komponenty, ktoré sme následne zoskupili podľa našich kritérií do proteínových sektorov. Tie sme vizualizovali a je možné ďalej s nimi, či už experimentálne alebo v teoretických rovinách pracovať. Experimentálne by sa dalo ukázať, na aké funkcie proteínu jednotlivé nezávislé komponenty alebo sektory vplývajú, prípadne čo spôsobia zmeny ich aminokyselín, resp. ako sa zmení fungovanie proteínu po aplikácií týchto zmien. Bioinformatikom sa však otvorili ďalšie možnosti skúmania získaných pozícií.



Obr. 3.11: Tepelné mapy pozícií usporiadaných do (A) nezávislých komponentov a (B,C) sektorov. Tepelné mapy 1A a 1B ukazujú prerozdelenie pozícií do nezávislých komponentov bez zoskupovania do sektorov. Na obrázkoch 1B a 2B sú komponenty zoskupené podľa toho, či sa v nich nachádzajú pozície podieľajúce sa na katalytickej funkcii domény (komponenty obsahujúce tieto pozície sme spojili do jedného, zvyšné sme ponechali ako samostatné sektory). Je vidieť, že nezávislé komponenty, ktoré boli ponechané ako samostatné sektory, ukazujú závislosť s inými. Preto bolo vytvorené nové zoskupenie nezávislých komponentov 1C a 2C. Tu si môžeme všimnúť, že získané sektory sú od seba skutočne nezávislé a nevidíme pozície, ktoré by mali vysokú hodnotu korelácie s pozíciami z iných sektorov súčasne (modré pásy v oboch grafoch 1C a 2C). Výsledky pre vstupy (1) kvasinky_500 a (2) yarrowia_500.

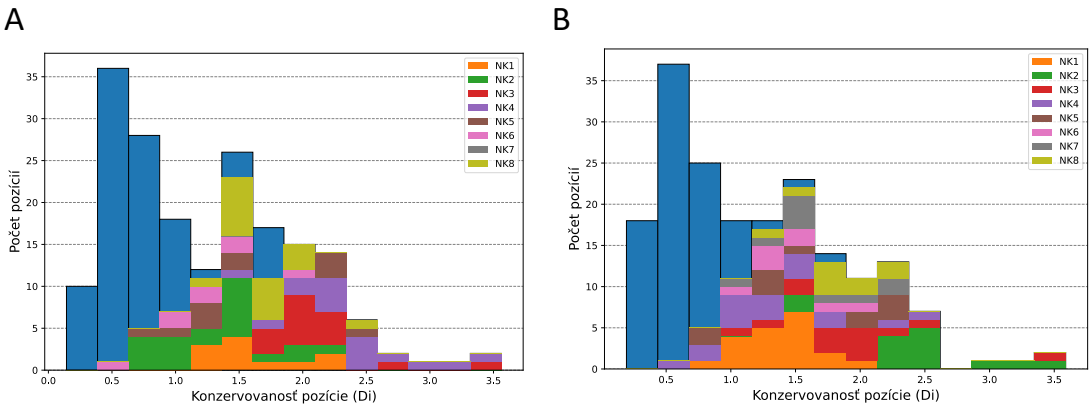


Obr. 3.12: 3D vizualizácia sektorov získaných zo vstupov (A) kvasinky_500 a (B) yarrowia_500 pomocou štruktúry 1WOK. V oboch prípadoch sú pozície katalytickej funkcie umiestnené v najväčšom prvom sektore.

3.3.1 Naozaj potrebujeme hľadať koevolvujúce dvojice?

Zo zafarbeného grafu konzervovanosti (viď. Obr. 3.4) sme vedeli vyčítať aj to, že sektory naozaj neobsahovali len také pozície, ktoré spolu susedia v primárnej štruktúre. Bolo teda jasné, že subštruktúry, ako sú sektory, nie je možné vidieť na prvý pohľad len na základe primárnej štruktúry. Ešte zaujímavejšie bolo, že väčšina pozícií, ktoré patrili do nejakého sektora, bolo vysoko konzervovaných. Bolo teda na mieste si klásť otázku, či by predsa len nestačilo na identifikáciu sektorov použiť iba informácie o konzervovanosti. Touto myšlienkou sa zaoberali aj ďalší vedci, ktorí pre proteíny, v ktorých bol nájdený len jeden sektor, tvrdia, že práve konzervovanosť pozície je kľúčom k nájdeniu sektorov a vlastne sa netreba zaoberať hľadaním koevolvujúcich dvojíc pozícií [46]. Opačne to vidia tvorcovia SCA analýzy, ktorí hľadanie koevolvujúcich dvojíc považujú za neoddeliteľnú súčasť pri hľadaní sektorov [38].

Napriek tomu, že sme v našej pozorovanej doméne vo všetkých prípadoch našli viac ako jeden sektor, rozhodli sme sa zistiť, ako veľmi závisí to, že je pozícia priradená do nejakého nezávislého komponentu od toho, ako je konzervovaná. Porovnávali sme skóre príslušnosti do niektorého z nezávislých komponentov a konzervovanosti týchto pozícií. Korelácia medzi týmito dvoma množinami hodnôt pre každú zo vstupných množín vyšla v rozmedzí od 0,6 po 0,65 (viď. Obr. 3.14).

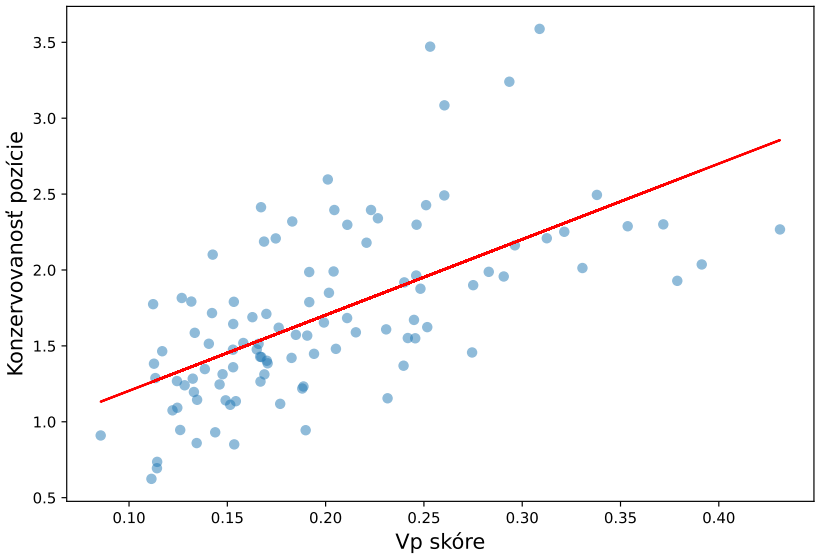


Obr. 3.13: Histogramy znázorňujúce počet pozícií pre hodnoty kozervovanosti. Pozície, ktoré boli priradené do nejakého nezávislého komponentu (NK) sú farebne odlíšené. Vidíme, že väčšina vysokokonzervovaných pozícií je súčasťou nejakého NK, avšak niektoré ostali nezaradené. Naopak v prípade (A) sa zaradila do NK6 pozícia, ktorá má relatívne nízku hodnotu konzervovanosti. Podobný prípad sa stal aj v prípade (B). Preto nemôžeme prehlásiť, že by vysoká konzervovanosť stačila na priradenie pozícií do NK. Výsledky pre vstupy (A) kvasinky_500 a (B) yarrowia_500.

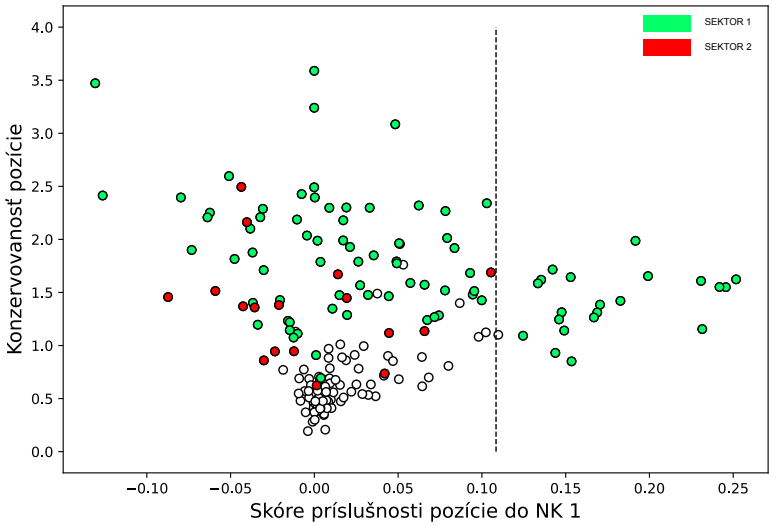
Koreláciu v takejto hodnote sme prehlásili za mierne pozitívnu. Výsledky sú teda naklonené k obom stranám, ale zdá sa, že napriek mierne pozitívnemu vzťahu, to, že je pozícia konzervovaná, nie je postačujúcou podmienkou pre to, aby patrila do nejakého nezávislého komponentu (viď. Obr. 3.13). Navyše sme spravili aj porovnanie skóre príslušnosti všetkých pozícií do jedného náhodne vybraného nezávislého komponentu a ich konzervovanosti, kde sme tak isto pozorovali zvýšenú koreláciu práve pri pozíciách s najvyšším skóre (viď. Obr. 3.15).

3.3.2 Pozície výnimočné pre kvasinkové sekvencie

Posledným bodom našej analýzy bolo zistenie vplyvu pridania kvasinkových sekvencií do vstupných množín a zistenie, v čom sa odlišovali výsledky kvasinkových a podrodinných vstupov. Keďže sektory sme vytvárali ručne pre každý vstup zvlášť, zamerali sme sa radšej na porovnanie nezávislých komponentov, ktoré algoritmicky vypočítal program SCA analýzy. Na porovnávanie získaných pozícií tvoriacich nezávislé komponenty sme vytvorili skript v jazyku Python. Program na vstupe dostal dve výsledné databázy SCA a porovnal všetky komponenty dvoch vstupných databáz navzájom (kód viď. Dodatok B: elektronická príloha: Kody/compareICs.py). Ako výstup vypísal rôzne štatistické informácie o týchto dvojiciach. Takto sme získali tabuľku zhôd a nezhôd dvojíc komponentov a celkovú podobnosť dvojice určenú Jaccardovým indexom [17].



Obr. 3.14: Bodový graf - porovnanie skóre príslušností všetkých pozícií priradených do nejakého z nezávislých komponentov a ich kozervovanosti zo vstupu yarrowia_500.



Obr. 3.15: Bodový graf - porovnanie skóre príslušnosti do NK1 (os x) a konzervovanosti (os y) pre všetky pozície zo vstupu yarrowia_500. Zvislou prerušovanou čiarou sú oddelené tie pozície, ktoré dosiahli dostatočne vysoké skóre na priradenie do NK1. Pri týchto pozíciách vidíme mierne pozitívnu koreláciu - vo väčšine prípadov, čím vyššie skóre pozície dosiahli, tým viac boli konzervované, niektoré pozície však ostali na rovnakej hodnote konzervovanosti napriek zvýšenému skóre príslušnosti do NK1. Pozície sú odlíšené farbami podľa príslušnosti do sektorov, biele pozície neboli priradené do žiadneho sektora.

Navyše sme si nechali aj vypísať konkrétne pozície, v ktorých zhoda nastala a aj tie, ktoré boli navyše. Aby sme získali lepší prehľad o podobnosti, resp. rozdielnosti celkových získaných výsledkov, podobne sme porovnali aj všetky pozície zaradené do všetkých nezávislých komponentov spolu z jednej vstupnej množiny s takýmito pozíciami z druhej. Vytvorili sme potom prehľadnú tabuľku podobností výstupov. (viď. Tab. 3.3).

Naše očakávania boli do značnej miery splnené. Všimli sme si, že Jaccardov index množiny nahodny_vyber a kvasinkovými vstupmi bol najnižší, naopak index bol relatívne vysoký pri porovnaní pri oboch prípadoch vstupu podrodiny s jej dvojicou doplnenou o kvasinkové sekvencie. Avšak nebol stopercentný, a preto sme mohli prehlásiť, že pridanie kvasiniek do množín naozaj pozmenilo výsledky SCA. Keďže sme pracovali s obmedzeným množstvom dát, výsledky by sa mohli mierne odlišovať po zvolení inej náhodnej vzorky podrodiny, a tak by sme získali iné „nové pozície“. Niektoré by sa mohli stratiť - nezaradiť sa do komponentov vôbec alebo sa priradiť do nejakého z nezávislých komponentov vo výsledku z podrodinného vstupu.

1.vstup	NK/P	2.vstup	NK/P	Zhody	Jaccard Index
kvasinky_500	7/99	podobne_kvasinky	9/112	96	83,48
kvasinky_500	7/99	nahodny_vyber	5/91	80	72,73
yarrowia_500	8/104	podobne_yarrowia	8/106	99	89,19
yarrowia_500	8/104	nahodny_vyber	5/91	87	80,56
kvasinky_500	7/99	yarrowia_500	8/104	91	81,25

Tabuľka 3.3: Tabuľka porovnaní pozícií z jednotlivých vstupov zaradených do ľubovoľného z nezávislých komponentov. Stĺpce zľava: (1) názov prvého porovnávaného súboru; (2) počet neprázdnych nezávislých komponentov(NK)/počet pozícií priradených do nezávislých komponentov (P) vo výsledkoch pre (1); (3) názov druhého porovnávaného súboru; (4) počet neprázdnych nezávislých komponentov (NK)/počet pozícií priradených do nezávislých komponentov(P) vo výsledkoch pre (3); (5) počet pozícií, ktoré boli pri oboch vstupoch priradené do nejakého NK; (6) Jaccardov Index podobnosti množín všetkých pozícií zaradených do NK.

3.4 Výsledky GREMLIN analýzy

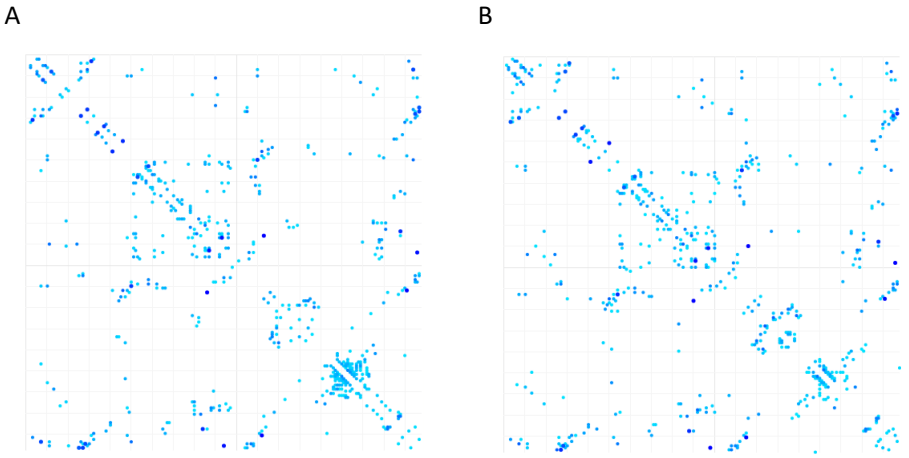
V tejto podkapitole sa venujeme jednotlivým výsledkom analýzy GREMLIN. Obrázky sme získali z platformy http://gremlin.bakerlab.org, kde bola celá analýza spúšťaná a vykonávaná. Aby sme pracovali s konzistentnými dátami pre prípadné porovnávania výsledkov, sme ako vstup použili zarovnania, v tak upravenej forme, ako ich upravila metóda SCA v jej druhom kroku (viď. Kap. 3.2.1). Keďže GREMLIN už ďalšie zostrihávania a úpravy nerobil, znamená to, že sme pracovali s rovnako zostrihanou referenciou a rovnakým zarovnaním ako v SCA. Konkrétne sme pracovali so súbormi kvasinky_500 a yarrowia_500. Referenčná sekvencia ostala podobne ako v SCA analýze PARP1_HUMAN, na ktorú sú všetky výsledky mapované.

3.4.1 Korelačná sila verzus štruktúry

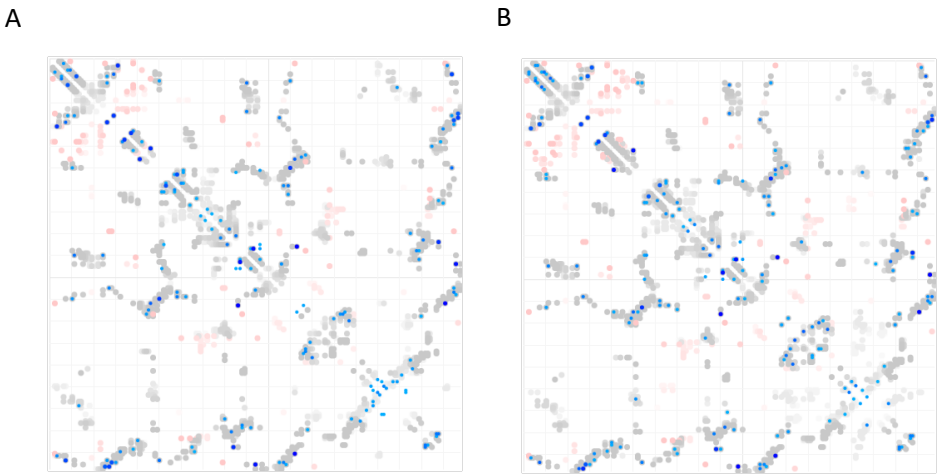
Na rozdiel od SCA metódy, korelačnú silu nám GREMLIN nevypísal pre všetky dvojice pozícií, ale len pre tie najlepšie. Spolu so skóre korelácie bola hneď daná aj pravdepodobnosť kontaktu dvojice pozícií. Vo všeobecnosti vieme povedať, že pre dvojice s vysokým skóre bola pravdepodobnosť kontaktu vysoká. Najvyššie korelácie boli znázornené pomocou bodového grafu (viď. Obr. 3.16). Ak vnímame vysoko korelované dvojice ako tie, čo majú skóre oveľa vyššie ako 1, môžeme pozorovať, že GREMLIN nenašiel príliš veľa takýchto dvojíc.

Ako sme spomínali, GREMLIN nájdené korelácie porovnáva s už existujúcimi homologickými štruktúrami pre pozorovaný proteín. Pre 10 vybraných štruktúr nám vykreslil bodový graf s prekrývajúcimi sa kontaktami v štruktúrach s predikovanými kontaktami, čo našiel GREMLIN vo svojej analýze (viď. Obr. 3.17). Pri oboch našich vstupoch sme však pozorovali veľmi nízke hodnoty $HH\Delta$ (v rozmedzí od 0,009 po maximálne 1,9 pre oba vstupy) a niekedy dokonca až záporné, konkrétne pre vstup yarrowia_500 nadobudla $HH\Delta$ s vybratou štruktúrou 1efyA hodnotu -0,028. Takéto nízke hodnoty znamenajú, že sekvencie v našom zarovnaní a predpočítanom zarovnaní pre referenčný proteín sú si veľmi podobné. Z toho vyplýva, že by GREMLIN pravdepodobne neprispel novými informáciami pri vytváraní novej terciárnej štruktúry. Pre referenčný ľudský proteín je to však očakávané, keďže má dobre stanovenú terciárnu štruktúru. Keď sme však analýzu spustili s referenčnou kvasinkovou sekvenciou, získali sme podobne nízke hodnoty $HH\Delta$, z čoho sme mohli následne predpokladať, že GREMLIN by nepomohol ani pri stanovovaní, resp. hľadaní novej 3D štruktúry pre kvasinkové sekvencie a ľudská štruktúra im je dostatočne blízka. Keďže sme pracovali so štruktúrou 1WOK aj v SCA, rozhodli sme sa využiť možnosť mapovania kontaktov aj na túto štruktúru (viď. Dodatok A: Obr. A1). Získali sme tak tabuľku s nájdeným skóre pre danú dvojicu a ich skutočnou vzdialenosťou v štruktúre. Tieto informácie

sme využili hlavne pre porovnávanie s výsledkami analýzy SCA, ktoré rozoberieme v ďalšej podkapitole. Kompletné výsledky získané v analýze GREMLIN prikladáme v elektronickej prílohe (viď. Dodatok B: elektronická príloha: GREMLIN_OUTPUT/).



Obr. 3.16: Bodový graf - korelácie medzi vybratými pozíciami. Spektrum veľkostí a modrej farby naznačujú silu korelácie medzi dvojicou pozícií: bledšia modrá a menšie body predstavuje slabšiu koreláciu a tmavšia modrá a veľké body zase silnú. Obe osi predstavujú pozície referenčnej sekvencie. Výsledky pre vstupy (A) kvasinky_500 a (B) yarrowia_500.



Obr. 3.17: Bodový graf - prekrytie významných korelujúcich dvojíc (skóre korelácie väčšie ako 1) s kontaktami nájdenými v desiatich vybraných štruktúrach. Modré body predstavujú korelácie získané GREMLIN-om a sivé a červené body sú kontakty v štruktúre. Môžeme si všimnúť, že je naozaj málo modrých bodov, ktoré by neboli prekryté so sivým, resp. červeným bodom z niektorej štruktúry. Obe osi predstavujú pozície referenčnej sekvencie. Výsledky pre vstupy (A) kvasinky_500 a (B) yarrowia_500.

3.5 Porovnanie výsledkov

Napriek tomu, že metódy SCA a GREMLIN nám poskytli rôznu škálu výsledkov, rozhodli sme sa porovnať niektoré údaje, ktoré počítali obe metódy. Primárne sme sa venovali 2 záležitostiam: (1) porovnanie miery koevolúcie dvojíc nájdenej metódou SCA a metódou GREMLIN a (2) hľadanie korelácií podľa GREMLIN-u v zaujímavých sektoroch nájdených SCA.

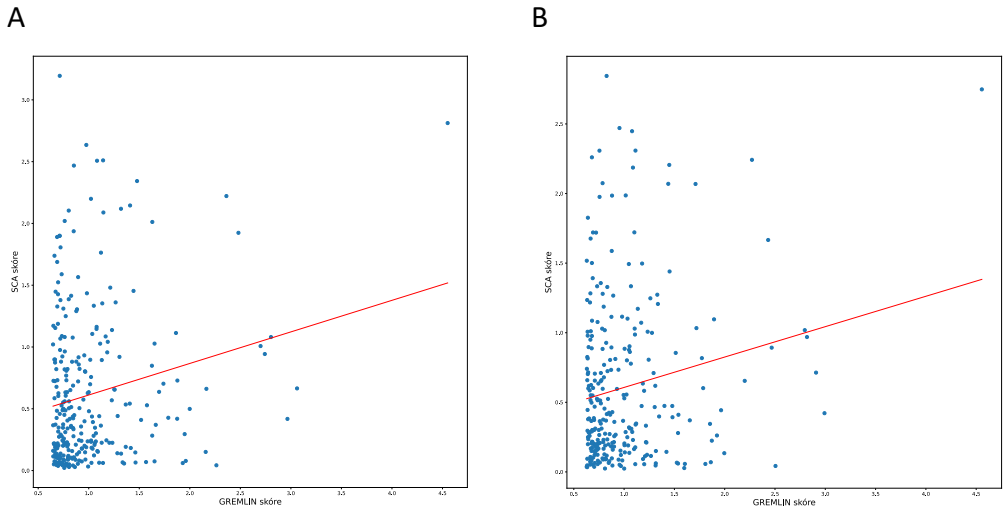
V prvom pozorovaní sme teda hľadali súvis medzi výškou hodnoty korelácie dvojíc pozícií nájdenej metódou SCA a GREMLIN. Keďže GREMLIN nám vrátil korelácie len pre dvojice pozícií s najvyššou hodnotou korelácie, aby sme pracovali s jednotnými údajmi, z výsledných SCA korelácií sme vybrali iba náležité pozície. Hodnoty z oboch metód sme zaznamenali v bodovom grae. Je zaujímavé, že vzťah medzi týmito hodnotami nie je takmer žiadny - ani pozitívny ani negatívny (viď. Obr. 3.18). Avšak pre dvojice s vysokým GREMLIN skóre majú relatívne vysoké skóre aj v SCA.

Naopak, mnohým dvojiciam pozícií, ktoré považuje SCA za vysoko korelované, GREMLIN priradil relatívne nízke skóre. Na základe našich poznatkov možno predpokladať, že korelácie medzi týmito dvojicami považuje GREMLIN za nepriame. Okrem toho to môžu byť vzdialenejšie pozície, pre ktoré GREMLIN nepripúšťa, že by boli korelované. Kód, ktorý vykonáva výber správnych pozícií z výslednej databázy SCA a vykresľuje porovnávacie bodové grafy, sme vytvorili v jazyku Python (kód viď. Dodatok B: elektronická príloha: Kody/SCA_GREM_corrCompare.py). Ako vstup sme tomuto skriptu poskytli databázu SCA a zoznam korelujúcich pozícií v textovom formáte, ktorý vytvoril GREMLIN.

Vstupný súbor	Pozícia 1	Pozícia 2	GREMLIN skóre korelácie	Vzdialenosť [Å]
kvasinky_500	891	924	4,725	3,34
	814	836	1,782	3,79
yarrowia_500	897	924	4,556	3,34
	942	946	1,002	5,39

Tabuľka 3.4: Pozície zo sektorov s neznámou funkciou, ktoré boli nájdené s významnou koreláciou aj v analýze GREMLIN. Vzdialenosť týchto pozícií je získaná z mapovania na štruktúru 1WOK. Môžeme si všimnúť, že pre jednu dvojicu pozícií je vzdialenosť miene vyššia ako 5 Å [Ångström].

V druhom pozorovaní sme sa zamerali na naše sektory nájdené v analýze SCA. Bližšie sme sa pozreli na tie sektory, ktoré neplnia katalytickú funkciu domény. Zaujímalo nás, či GREMLIN našiel aj v týchto sektoroch nejaké zaujímavé korelujúce dvojice.



Obr. 3.18: Grafy vyjadrujúce vzájomné vzťahy medzi koreláciami dvojíc pozícií nájdenými v GREMLIN analýze (os x) a SCA analýze (os y). Miera vzťahu medzi hodnotami korelácie pozícií nájdenými oboma analýzami je približne 0,2, čo predstavuje takmer žiadnu koreláciu medzi týmito hodnotami. Výsledky pre vstupy (A) kvasinky_500 a (B) yarrowia_500.

Podobne ako predošlé pozorovania, výsledky porovnaní neukázali veľkú zhodu. Pre obe vstupné zarovnania sme našli iba dve silno koevolvujúce dvojice pozícií. Naša úvaha bola, že pokiaľ nebudú tieto dvojice vo fyzickom kontakte, mohli by potenciálne predstavovať zaujímavé vzdialené koevolvujúce pozície. Ako sa však ukázalo, tieto pozície sú v relatívne tesnej blízkosti (vzdialenosť <5 Å) v terciárnej štruktúre 1WOK (viď. Tab. 3.4). Potvrdzuje to teda teóriu GREMLIN-u, že vysoko korelované dvojice, ktoré nájde, sú v terciárnej štruktúre vo fyzickom kontakte. Pre nás tým pádom nepredstavujú úplne výnimočné pozície, keďže je očakávateľné, že fyzicky blízke pozície budú koevolvovať a navzájom sa ovplyvňovať. Nás budú naopak zaujímať tie pozície, ktorým SCA analýza priradila vysoké hodnoty korelácie a GREMLIN ich nepovažuje za silno koevolvované, pretože my práve potrebujeme vzdialenejšie potenciálne koevolvujúce pozície, ktoré, ako sme spomínali, má GREMLIN tendenciu ignorovať. Pre nás teda pozície, nájdené analýzou SCA, naďalej ostávajú smerodajné pre návrhy nových biologických experimentov. V sumáre preto môžeme prehlásiť, že pre hľadanie vzdialených koevolvujúcich pozícií sa nám osvedčila viac metóda SCA. Správnosť jej výsledkov by bolo vhodné overiť spomínaným experimentom.

Záver

V našej práci sme najprv preštudovali princípy a fungovanie dvoch bioinformatických metód - SCA a GREMLIN. Zistili sme, že metódy analyzujú vstupné zarovnania odlišným spôsobom. SCA často používala maticové operácie z oblasti algebry, GREMLIN vo svojich výpočtoch využíval pravdepodobnostné modely a učiace sa algoritmy. Pre hľadanie proteínových sektorov bola SCA určite užitočnejšia, keďže zahŕňala kroky pre ich nájdenie. Vzhľadom na to, že GREMLIN mal tendenciu ignorovať vzdialené koevolvujúce dvojice pozícií, na ich nájdenie sme využili opäť metódu SCA. GREMLIN nám naopak umožnil porovnať jeho hodnoty koevolúcie a získané pravdepodobnosti kontaktu dvojíc pozícií s už existujúcimi štruktúrami.

Priebeh analýzy SCA sme dokázali ilustrovať pomocou obrázkov, ktoré vznikli priamo pri nami vykonanej analýze, čo nám pomohlo pri samotnom pochopení jej jednotlivých krokov. Spomínané nástroje sme aplikovali na katalytickú doménu proteínu PARP, ktorý bol objavený v kvasinkách. Obe analýzy sme spúšťali na viacerých vstupoch, ktoré sme vytvorili tak, aby sme výsledky analýz vedeli relevantne porovnávať. Jedným z pozorovaní bolo napríklad hľadanie vplyvu pridania kvasinkových sekvencií do vstupu podrodiny. Ďalej sme zisťovali, ako sa výsledok náhodnej vstupnej množiny z celej rodiny proteínu zmenil, ak sme ho zúžením na podrodinu prispôsobili kvasinkovým sekvenciám.

Pomocou nástroja SCA sme pre každý použitý vstup vyhodnotili konzervovanosť jednotlivých pozícií, odvodili sme koevolvujúce dvojice a získali nezávislé komponenty, tvorené vzájomne koevolvujúcimi pozíciami. Spojili sme ich do väčších celkov, ktoré už predstavovali hľadané proteínové sektory. V našej práci sme pre pozorovaný proteín PARP, našli pri každom použitom vstupe dva sektory. Za potenciálne zaujímavé sektory sme považovali hlavne také, ktoré v sebe nezahŕňali pozície plniace katalytickú funkciu domény PARP, keďže funkcia týchto sektorov nie je objasnená. Na druhej strane pozície, ktoré tvorili sektor s katalytickou funkciou, môžu predstavovať zaujímavých kandidátov na vzdialené spolupracujúce pozície s miestami viazania. Pomocou analýzy GREMLIN sme opäť získali koevolvujúce dvojice pozícií a predikciu kontaktov, ktoré sme porovnávali s kontaktami v už existujúcich štruktúrach. Na záver sme porovnali hodnoty koevolúcie dvojíc pozícií získanými oboma analýzami. Pre pozície z vybraných sektorov, získaných z SCA, sme sa pokúsili vyhľadať mieru ich vzájomných

korelácií pomocou analýzy GREMLIN.

Je na mieste rozmýšľať o vylepšení priebehu vykonanej analýzy. Jedným z návrhov na zlepšenie, resp. zautomatizovanie fungovania SCA analýzy, by bolo zhotovenie skriptu, ktorý by vytváral vstupné množiny pre samotnú analýzu. Zamedzilo by sa tak ručnej práci s dátovými množinami, ktorá je občas zdĺhavá a môžu pri nej vznikať chyby. Skript by ako vstup dostal sekvencie celej rodiny a vybrané, napríklad kvasinkové, sekvencie, ktorým by sme analýzu chceli prispôsobiť. Z rodiny by našiel podrodinu pre vybrané sekvencie podľa stanovených kritérií, odobral by určitý počet náhodných sekvencií z podorodiny a vytvoril podordinnú vstupnú množinu. Následne by k nej pridal vybrané sekvencie, čím by vytvoril vstupný súbor obohatený o napríklad kvasinkové sekvencie. V poslednom kroku by obe vytvorené množiny sekvencií zarovnal pomocou nástroja pre viacnásobné zarovnanie.

Bolo by taktiež zaujímavé ísť hlbšie do samotného kódu SCA a skúsiť zmeniť fungovanie tak, aby sekvenciám, ktoré my považujeme za dôležité (v našej analýze to boli kvasinkové sekvencie), pridal vyššie váhy ako ostatným. Ďalším zlepšením by mohlo byť mapovanie výsledkov na rôzne terciárne štruktúry v jednom behu programu, čo by sa následne vyhodnotilo podobne ako v analýze GREMLIN, aby sme získali čo najlepšie mapovanie. Keby sme sa chceli viac zamerať na pozície výnimočné len pre kvasinkové sekvencie, dal by sa spraviť program, ktorý by SCA analýzu spúšťal na náhodných podmnožinách podrodín, následne by k nim pridal kvasinkové sekvencie a opäť by spustil analýzu. Pomocou nami vytvoreného porovnávacieho skriptu by sa získali „nové pozície“ pre kvasinkové sekvencie a následne by sme našli najviac opakujúce sa „nové pozície“, ktoré by sme mohli považovať za skutočne výnimočné pre kvasinky. Treba však mať stále na pamäti, že obe analýzy sú založené iba na matematických výpočtoch a pracujú len s obmedzeným množstvom dát, a teda výsledky na nových podmnožinách, hoci tej istej rodiny proteínu, sa môžu mierne odlišovať. S cieľom lepšie vyhodnotiť výsledky SCA analýzy by mal nasledovať návrh biologického experimentu. Ten by spočíval v systematickej a cielenej zámene aminokyselín na významných pozíciách a následnom pozorovaní účinku týchto zmien. Ak by sme sa chceli pustiť do zámen dvojíc aminokyselín, postupovali by sme od najsilnejšie koevolvovaných pozícií. Overovanie spolupráce dvoch pozícií by mohlo byť založené na pozorovaní toho, či si proteín zachováva alebo naopak stráca funkciu po ich spoločnej alebo postupnej individuálnej zámene. Ďalším krokom by mohlo byť vyhodnotenie funkcií získaných nezávislých komponentov a následne sektorov, ktoré by mohli pomôcť vysvetliť a objasniť fungovanie proteínu PARP v kvasinkách.

Prínosom našej práce je to, že sme získali významné a zároveň neintuitívne pozície, pri ktorých vieme aj hodnoty ich vzájomných korelácií, a teda nemusíme spriahnuté dvojice aminokyselín proteínu hľadať experimentálne. V takom prípade by sme sa totiž museli pozrieť na príliš veľa dvojíc, s dvadsiatimi aminokyselinami pre každú pozíciu.

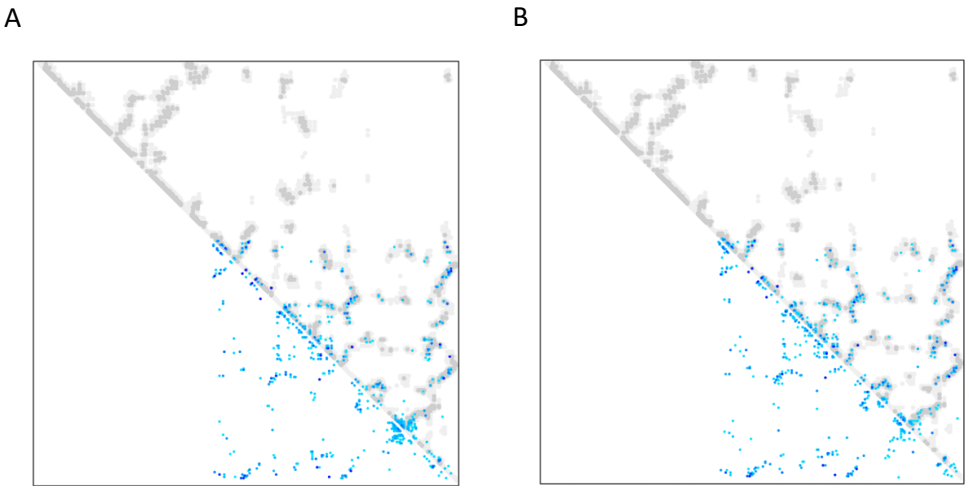
Slovník pojmov

Keďže je naša práca medziodborová a používa termíny aj z oblasti biológie, uvádzame krátky zoznam pojmov, ktoré v práci používame.

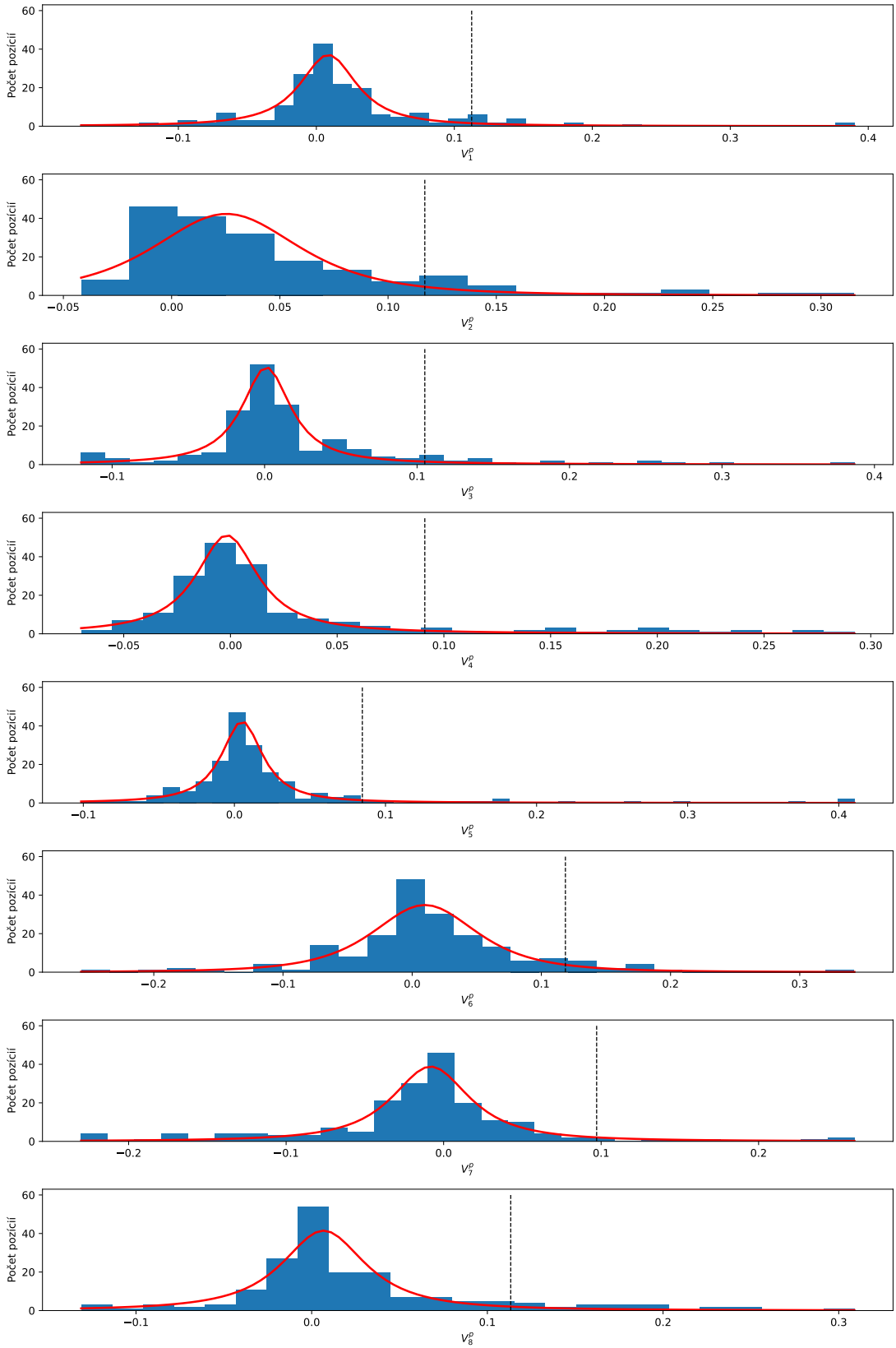
- biomakromolekuly - veľké molekuly, ktoré sú súčasťou živých organizmov (napríklad nukleové kyseliny, bielkoviny, polysacharidy)
- genóm - kompletná genetická informácia organizmu
- katalyzátor - látka urýchľujúca priebeh chemickej reakcie
- monomér proteínu - jedna molekula proteínu. Niektoré proteíny fungujú až po spojením viacerých monomérov - di-, tri- tetramér a podobne.
- mutácia - zmena genetickej informácie
- proteóm - súbor všetkých proteínov jedného organizmu
- taxonomická skupina - skupina príbuzných organizmov

Dodatok A: pridané obrázky

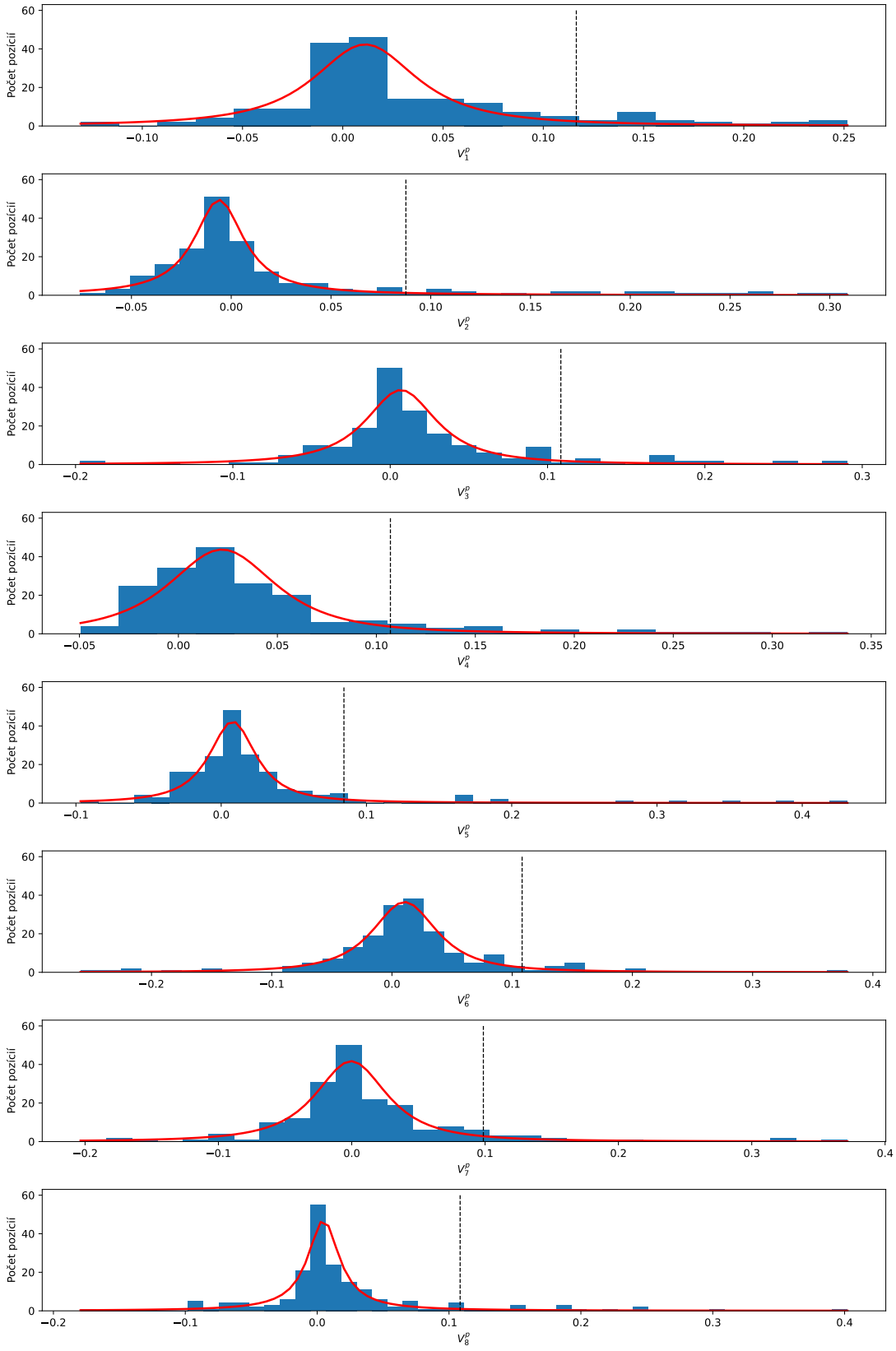
V tejto prílohe uvádzame niektoré obrázky, ktoré vznikli pri analýze SCA a GREMLIN.



Obr. A1: GREMLIN: Bodový graf - prekrytie významných korelujúcich dvojíc (skóre korelácie väčšie ako 1) s kontaktami nájdenými v štruktúre 1WOK. Môžeme vidieť, že všetky silno korelované dvojice (tmavomodré body) sú pokryté kontaktom aj v štruktúre 1WOK. Matica kontaktov štruktúry (sivé body) je vykreslená len nad diagonálou v pravej hornej polovici, čo je dôsledkom toho, že je matica kontaktov symetrická. Navyše vidíme, že matica so skóre korelujúcich dvojíc je len v dolnej ľavej polovici a prvé pozície štruktúry nie sú pokryté vôbec. Je to spôsobené tým, že štruktúra je vytvorená pre dlhšiu sekvenciu ako je naša pozorovaná doména. Osi X aj Y predstavujú pozície sekvencie štruktúry 1WOK.Výsledky pre vstupy (A) kvasinky_500 a (B) yarrowia_500.



Obr. A2: SCA: Mapovanie skóre príslušnosti pozícií do každého nezávislého komponentu na Studentovo rozdelenie. Na osiach x je hodnota skóre príslušnosti do daného komponentu pre každú pozíciu a na osiach y je počet pozícií s týmto skóre. Zvislá prerušovaná čiara predstavuje hraničnú minimálnu hodnotu pre pozície zaradené do daného NK a červená súvislá čiara zobrazuje priebeh Studentovho rozdelenia. Výsledky pre vstup kvasinky_500.fasta.



Obr. A3: SCA: Výsledky mapovania skóre príslušností do nezávislých komponentov pre vstup yarrowia_500.fasta. Popis grafu analogický ku popisu ku Obr. A2.

Dodatok B: Zoznam súborov elektronickej prílohy

V tejto prílohe uvádzame stručný, prehľadný zoznam elektronickej prílohy. Podrobnejšie informácie o obsahu jednotlivých priečinkov a súborov sa nachádzajú v jej samotných priečinkoch.

- SCA_OUTPUT: Výsledky výpočtov aj vizualizácií jednotlivých vstupov SCA analýzy
- GREMLIN_OUTPUT: Výsledky jednotlivých vstupov GREMLIN analýzy
- Vstupne_subory: Vstupné dátové množiny použité pre bioinformatické analýzy
- Kody: Skripty použité pre vizualizáciu a porovnávanie výsledkov:
  - printResultsSK.py
  - compareICs.py
  - SCA_GREM_corrCompare.py
- Tabulky: Zoznamy pozícií získaných zo sektorov, porovnávacie tabuľky:
  - PARP1.xlsx
  - porovnania_pozicii_vsetkych_NK.xlsx
  - SCA_vs_gremlin.xlsx
  - zaujimave_pozicie.xlsx
- Porovnanie: Kompletné výstupné súbory pre porovnanie nezávislých komponentov získaných z jednotlivých vstupov
- HMM_logos: Kompletné HMM logá pre súbory kvasinky_500, podobne_kvasinky, yarrowia_500 a podobne_yarrowia, originálne logo získané z databázy Pfam.

Literatúra

[1] Acland, A. a kol.: Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 2016, 44, s. D7-D19.

[2] Alberts, B. a kol.: Molecular Biology of the Cell. New York, Garland Science 2002.

[3] Anishchenko, I. a kol.: Origins of coevolution between residues distant in protein 3D structures. *Proceedings of the National Academy of Sciences of the United States of America*, 2017, 114, s. 9122-9127.

[4] Artemova, T.: The Organization of Protein Sectors. 2011. 5 s. Dostupné na internete ku dňu 10.5.2022: https://web.mit.edu/8.592/www/grades/projects/ Projects(2011)/TatianaArtemova.pdf.

[5] Bai, P.: Biology of Poly(ADP-Ribose) Polymerases: The Factotums of Cell Maintenance. *Molecular Cell*, 2015, 58, s. 947-958.

[6] Balakrishnan, S. a kol.: Learning Generative Models for Protein Fold Families. *Proteins*, 2011, 79, s. 1061-1078.

[7] Bell, A.J., Sejnowski, T.J.: An Information-Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Computation*, 1995, 7, s. 1129-1159.

[8] Burley, S.K. a kol.: RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Research*, 2021, 49, s. D437-D451.

[9] Camacho, C. a kol.: Biology of Poly(ADP-Ribose) Polymerases: The Factotums of Cell Maintenance. *BMC Bioinformatics*, 2009, 10, s. 421.

[10] Capra J. A. , Singh, M.: Predicting Funcionally Important Residues from Sequence Conservation. *Bioinformatics*, 2007, 23, s. 1875-1882.

[11] Dunn, S. D., Wahl, L. M., Gloor, G. B.: Mutual Information Without the Influence of Phylogeny or Entropy Dramatically Improves Residue Contact Prediction. *Bioinformatics*, 2008, 24, s. 333-340.

[12] Dahleh, M., Dahleh, M.A., Verghese, G.: Lectures on Dynamic Systems and Control. 2011. Dostupné na internete ku dňu 10.05.22: https://viterbi-web.usc.edu/ mihailo/courses/ee585/f17/mit-notes//mit-notes.pdf.

[13] Durbin, R. a kol.: Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge, The Press Syndicate Of The University Of Cambridge 1998. 365 s.

[14] Edgar, R.C.: MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 2004, 5, s. 113.

[15] faSomeRecords. 2020. Dostupné na internete ku dňu 10.5.2022: https://github.com/santiagosnchez/faSomeRecords.

[16] Finn, R. D., Clements, J., Eddy, S. R.: HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39, s. W29–W37.

[17] Fletcher, S., Islam, M.Z.: Comparing Sets of Patterns with the Jaccard Index.*Australasian Journal of Information Systems*, 2018, 22.

[18] github: GitHub. 2020. Dostupné na internete ku dňu 10.5.2022: https://github.com/.

[19] Gu, Z. a kol.: New perspectives on the plant PARP family: Arabidopsis PARP3 is inactive, and PARP1 exhibits predominant poly (ADP-ribose) polymerase activity in response to DNA damage. *BMC Plant Biology*, 2019, 19, s. 1-18.

[20] Halabi, N. a kol.: Protein Sectors: Evolutionary Units of Three-Dimensional Structure. *Cell*, 2009, 138, s. 774-786.

[21] Halabi, N. a kol.: Supplemental Data Theory Protein Sectors: Evolutionary Units of Three-Dimensional Structure. *Cell*, 2009, 138.

[22] Hunter, J.D.: Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 2007, 9, s. 90-95.

[23] Jones, D.T. a kol.: PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics (Oxford, England)*, 2012, 28, s. 184-190.

[24] Jumper, J. a kol.: Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021, 596, s. 583-589.

[25] Kamisetty, H., Ovchinnikov, S., Baker, D.: Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era *Proceedings of the National Academy of Sciences of the United States of America*, 2013, 110, s. 15674-15679.

[26] Ko, H. L., Ren, E. C.: Functional Aspects of PARP1 in DNA Repair and Transcription. *Biomolecules*, 2012, 2, s. 524–548

[27] Kubyschkin, V., Acevedo-Rocha, C. G., Budisa N.: On universal coding events in protein biogenesis. *Biosystems*, 2018, 164, s. 16-25.

[28] Madeira, F. a kol.: The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research*, 2019, 47, s. W636-W641.

[29] Mistry, J. a kol.: Pfam: The protein families database in 2021.*Nucleic Acids Research*, 2021, 49, s. D412-D419.

[30] Morales, J. a kol.: Review of Poly (ADP-ribose) Polymerase (PARP) Mechanisms of Action and Rationale for Targeting in Cancer and Other Diseases. *Critical Reviews In Eukaryotic Gene Expression*, 2014, 24, s. 15-28.

[31] Morcos, F. a kol.: Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America*, 2011, 108, s. E1293-E1301.

[32] National Center for Biotechnology Information (NCBI). Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. 1988. Dostupné na internete ku dňu 10.05.22: https://www.ncbi.nlm.nih.gov/.

[33] Nowak-Markwitz, E. a kol.: PARP Inhibitors: Review of Mechanisms of Action and BRCA1/2 Mutation Targeting. *Menopause Review,* 2016, 15, s. 215-219.

[34] Ornes, S.: Researches Turn to Deep Learning to Decode Protein Structures. *Proceedings of the National Academy of Sciences of the United States of America* 2022, 119.

[35] Perina, D. a kol.: Distribution of Protein Poly(ADP-Ribosyl) Action Systems Across All Domains of Life. *DNA Repair*, 2014, 23, s. 4-16.

[36] Quinlan, A.R., Hall, I.M.: BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 2010, 26, s. 841-842.

[37] Raman, A.S., White, K.I., Ranganathan, R.: Origins of Allostery and Evolvability in Proteins: A Case Study. *Cell*, 2016, 166, s. 468-480.

[38] Rivoire, O., Reynolds, K.A., Ranganathan, R.: Evolution-Based Functional Decomposition of Proteins. *PLoS Compxsdsational Biology*, 2016, 12.

[39] Rivoire, O., Reynolds, K.A., Ranganathan, R.: pySCA. 2019. Dostupné na internete ku dňu 10.5.2022: https://github.com/ranganathanlab/pySCA.

[40] Rivoire, O., Reynolds, K.A., Ranganathan, R.: Statistical Coupling Analysis in Python. 2019. Dostupné na internete ku dňu 10.5.2022: https://ranganathanlab.gitlab.io/pySCA/.

[41] Rivoire, O., Reynolds, K.A., Ranganathan, R.: Statistical coupling analysis: supplementary methods and codes. *PLoS Computational Biology*, 2016, 12.

[42] Rose, M. a kol.: PARP Inhibitors: Clinical Relevance, Mechanisms of Action and Tumor Resistance. *Frontiers in Cell and Developmental Biology*, 2020, 8, s. 564601.

[43] Schiewer, M. J. , Knudsen, K. E.: Transcriptional Roles of PARP1 in Cancer. *Molecular Cancer Research*, 2014, 12, s. 1059-1080.

[44] Schrödinger, LLC: The PyMOL Molecular Graphics System, Version 2.5. 2022. Dostupné na internete ku dňu 10.5.2022: http://www.pymol.org/pymol.

[45] Steinegger, M. a kol.: HH-suite3 for fast remote homology detection and deep protein annotation, *BMC Bioinformatics*, 2019, 20, s. 473.

[46] Teşileanu, T., Colwell, L.J., Leibler, S.: Protein Sectors: Statistical Coupling Analysis versus Conservation. *PLoS Computational Biology*, 2015, 11.

[47] The UniProt Consortium: UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 2021, 49, s. D480-D489.

[48] Yee, A. A. a kol.: NMR and X-ray Crystallography, complementary tools in structural proteomics of small proteins. *Journal of the American Chemical Society*, 2005, 127, S. 16512-16517.

[49] Zerihun, M.B. a kol.: pydca v1.0: a comprehensive software for direct coupling analysis of RNA and protein sequences. *Bioinformatics (Oxford, England)*, 2020, 36, s. 2264-2265.