UNIVERZITA KARLOVA V PRAZE
Přírodovědecká fakulta

Studijní program: Biologie
Studijní obor: Mikrobiologie

Bc. Albert Sokol

Nekovalentní interakce tryptofanu ve struktuře proteinu
Non-covalent interactions of tryptophan in protein structure

Diplomová práce

Vedoucí práce:
RNDr. Radovan Fišer, Ph.D

Praha, 2019

Prohlášení:

Prohlašuji, že jsem závěrečnou práci zpracoval samostatně a že jsem uvedl všechny použité informační zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

V Praze, 30. 7. 2019

Bc. Albert Sokol

Děkuji svému školiteli RNDr. Radovanu Fišerovi, Ph.D. za odborné konzultace a podporu.

Obsah

1 Abstrakt

Dokonalá znalost nekovalentních interakcí aminokyselin uvnitř proteinové struktury je esenciální pro úplné pochopení jeho konformace, stability a funkce. Mezi všemi aminokyselinami, které obvykle tvoří protein, se tryptofan vyjímá jednak svojí vzácností, ale také velikostí postranního řetězce tvořeného indolovou skupinou. Ta je schopna zajišťovat různé typy nepostradatelných interakcí uvnitř proteinu, mezi různými polypeptidovými řetězci, ale také třeba mezi proteinem a biologickou membránou. Navíc se jedná o nejčastěji využívaný přirozený bílkovinný fluorofor.

Ke studiu aminokyselinových interakcí se běžně využívají databáze vyřešených proteinových struktur, nad kterými se vytváří více či méně komplexní analýzy. Takto již byly nalezeny mnohé nekovalentní interakce, které mohou mezi tryptofanem a ostatními aminokyselinami nastávat. Většina těchto analýz se ale soustřeďuje na studium konkrétní interakce a nezabývá se prostředím tryptofanu jako celku, kde se všechny aminokyseliny vzájemně ovlivňují.

Pomocí nově vytvořených postupů jsou v této práci analyzovány profily výskytu jednotlivých aminokyselin okolo indolové skupiny tryptofanu a výsledky porovnány s dostupnou literaturou. Aminokyselina, která má největší preferenci k tryptofanu, se ukázala být opět tryptofan a tyto dvojice tryptofanů jsou podrobeny detailní analýze.

K závěrům mé práce mimo jiné patří zjištění, že arginin a lysin vykazují v literatuře popisovanou kationt-π interakci, ale zvýšeným výskytem se v jeho okolí neprojevují. Naproti tomu aniont-π interakci v kombinaci s tryptofanem jsem nepozoroval a domnívám se, že jde v literatuře o chybné výsledky.

Analýza tryptofanových párů ukázala, jak strukturovaný je interakční prostor okolo tryptofanu. Jedná se o různé vrstvy, kde vždy převažuje určitá orientace indolových skupin tryptofanů. Takovéto nenáhodné rozdělení se nachází až do vzdálenosti 10 Å mezi indoly. Proto pro všechny analýzy vlastností prostoru okolo tryptofanu (pravděpodobně i jiných aminokyselin) je potřeba postupovat pouze v určitých konkrétních směrech, a to i do větších vzdáleností a neomezovat se na průzkum průměrných vlastností celého okolního prostoru najednou.

Klíčová slova

Nekovalentní interakce, Tryptofan, PDB databáze, Struktury proteinů, Sekvenční homologie.

2 Abstract

A thorough knowledge of non-covalent amino acid interactions within a protein structure is essential for a complete understanding of its conformation, stability and function. Among all the amino acids that usually make up a protein, tryptophan is distinguished both by its rarity and size of its side chain formed by an indole group. It is able to provide various types of indispensable interactions within the protein and between different polypeptide chains, but also between the protein and a biological membrane. In addition, it is the most commonly used natural fluorophore.

Databases of solved protein structures are commonly used to study amino acid interactions and allow more or less complex analyzes of the issue. Thus many non-covalent interactions that may occur between tryptophan and other amino acids have been found. However, most of these analyzes focus on specific interactions and do not follow up the tryptophan's environment as a whole, where all amino acids interact.

Some newly developed methods have been used in this Thesis, specifically the occurrence profiles of the individual amino acids around the indole group of tryptophan and the results were compared with an available literature. The amino acid that has the greatest preference for tryptophan turned out to be tryptophan again, and these tryptophan pairs were subjected to more detailed analysis.

One of the conclusions of my work is the finding that arginine and lysine show a cation-π interaction described in the literature, but that they do not show an increased occurrence in its surroundings. On the other hand, I did not observe any anion-π interaction in combination with tryptophan and I believe that these are erroneous results in the literature.

Analysis of tryptophan pairs showed how structured the interaction space around tryptophan is. There are different layers, where the orientation of the indole groups of tryptophanes always prevail. Such a non-random distribution is located up to 10 Å distance between the indoles. Therefore, for all analyzes of the properties of the space around tryptophan (and probably other amino acids as well), it is necessary to proceed only in certain specific directions, even at greater distances and not to limit the research to the average properties of the surrounding at once.

Keywords

Non-covalent interactions, Tryptophan, PDB database, Protein structures, Sequence homology.

3 Seznam zkratek

PDB - Protein Data Bank

LV - Levenshteinova vzdálenost

UniProt - Universal Protein Resource

SQL - Structured Query Language

NMR - Nuclear magnetic resonance

BLAST - Basic Local Alignment Search Tool

DSSP - Define Secondary Structure of Proteins

4 Úvod

Nekovalentní interakce mezi aminokyselinami jsou esenciální pro strukturu, stabilitu a funkci proteinu. Tyto interakce jsou součástí mnoha výzkumů, které se zaměřují především na aromatické aminokyseliny tryptofan, tyrosin, fenylalanin a histidin. Tyto aromáty se díky své struktuře mohou účastnit mnoha různých nekovalentních interakcí od velmi známých π-π interakcí, kationt-π interakcí, všudypřítomných CH-π interakcí až po málo známé aniont-π interakce. Tryptofan jsem zvolil z důvodu nejnižšího výskytu v primární sekvenci proteinů, a protože je součástí výzkumu v mnoha různých odvětvích. Jedná se například o nejčastěji využívaný přirozený bílkovinný fluorofor.

V posledních letech exponenciálně stoupá počet vyřešených proteinových struktur v PDB databázi a díky tomu mohou vznikat jejich komplexnější analýzy. Většina výzkumů se snaží hledat a analyzovat v těchto strukturách konkrétní molekulové interakce a jen několik z nich se věnuje širšímu průzkumu. Dle mého názoru je tento přístup chybný, jelikož v proteinových strukturách lze najít i konfigurace aminokyselin, které by nastávat neměly. Tím může vzniknout dojem, že nalezená konkrétní konfigurace je nějak zásadní. Tryptofan je ale v proteinu přímo obalen aminokyselinami, které spolu interagují, takže se při studiu nelze soustředit pouze na jednu z nich.

Vzhledem k mým zkušenostem se zpracováním velkého objemu dat jsem se rozhodl pro komplexní analýzu interakcí tryptofanu v těchto vyřešených strukturách. Zaměřil jsem se především na jeho aminokyselinové prostorové okolí, a specificky pak na jeho interakce s jinými tryptofany. Při procházení dostupné literatury jsem postupně nabýval dojmu, že mnohé práce přistupují k analýze chybně (především v nahlížení na π-π interakce), a také nedostatečně názorně. Proto jsem se raději rozhodl vytvořit vlastní postupy pro analýzu aminokyselinových interakcí místo toho, abych přebíral již publikované algoritmy.

PDB databáze bohužel obsahuje obrovské množství homologních proteinů (případně struktury se zcela identickou sekvencí aminokyselin), což může značně ovlivnit výsledky. Autoři obdobných výzkumů využívají již vytvořené programy, které jim poskytnou velmi „ořezaný“ dataset unikátních proteinů, nad kterými poté vytvoří vlastní analýzu. Já sám byl ale přesvědčen, že pokud chci plně pochopit problematiku interakcí vyskytujících se v PDB databázi, je potřeba, abych se zároveň co nejlépe seznámil s obsaženou homologií a vymyslel vlastní postupy k jejímu odstranění s ohledem ke zkoumanému tryptofanu.

Jelikož jednotlivé postupy a analýzy většinou vycházely z průběžných výsledků mé práce, musel jsem tomu přizpůsobit i její strukturu. Jednotlivé grafy bylo potřeba rovnou

komentovat, abych vysvětlil svůj další postup. Proto kapitoly Výsledky a Diskuze jsou spojeny do jedné. Práce rovněž neobsahuje žádná experimentální data, jelikož by se týkala pouze několika málo proteinů a já měl v úmyslu analyzovat co největší datasety, aby byly závěry obecně platné. Přesto doufám, že moje výsledky a jejich grafická prezentace poskytnou čtenáři nový pohled na poměrně složitou problematiku.

5 Cíle práce

Prvotním cílem této práce je pro jednotlivé specializované analýzy vytvořit co nejlepší, a zároveň dostatečně početný dataset proteinových řetězců, jednotlivých tryptofanů a jejich párů. Při tomto postupu se podrobně seznámit s PDB souborem obsahující informace o proteinové struktuře, analyzovat výskyt homologních proteinů v PDB databázi a vyvinout postupy pro odstraňování homologie s ohledem na tryptofan.

Druhá část práce se zabývá prostorovým okolím tryptofanu. V té chci zjistit, jestli existují skutečné tendence jednotlivých aminokyselin vyskytovat se v blízkosti indolu. Pokud ano, jsou pak nalezené preference popsatelné jednoduchými pravidly? Jaká jsou tato pravidla a jakou roli hrají konkrétní chemické skupiny v postranních řetězcích? Nemohou sekvenčně blízké aminokyseliny narušovat výsledky analýz? A jaký vliv mají sekundární struktury proteinů?

Pro nalezení odpovědí je potřeba vytvořit hodnotící postup pro porovnání aminokyselin a jejich snahy být v určité vzdálenosti od tryptofanu. Dále určit jaký počet aminokyselin v blízkém okolí každého tryptofanu má smysl zkoumat, aby se jednalo o reprezentativní vzorek.

Dají se najít jednoznačné kationt-π, aniont-π, CH-π, prolin-π a další interakce? I přesto, že jsou hledány všechny najednou ve stejném datasetu? Nejedná se jen o artefakty? Existují rozdíly mezi “velkými” a “malými” postranními řetězci v preferenci k tryptofanu?

Třetí část práce analyzuje páry tryptofanů. V ní chci zjistit, do jaké vzdálenosti se tryptofany ovlivňují. To znamená, v jaké nejdelší vzdálenosti se dá prokázat nenáhodnost uspořádání dvou tryptofanů? A čím je to umožněno? S tím souvisí snaha nalézt veličinu, která by charakterizovala (ne)náhodnost uspořádání tryptofanů v prostoru.

Dále bych rád zjistil, jaké jsou skutečně preferované orientace a pozice tryptofanových párů při zohlednění vlastností prostoru a pravděpodobnosti. Jaký je rozdíl v uspořádání u plochy a u hrany indolové skupiny tryptofanu? Dopouští se autoři chyby, když analyzují interakce aromátů v prostoru okolo tryptofanu jako celku?

6 Přehled literatury

6.1 Vlastnosti tryptofanu

Tryptofan patří mezi čtyři aromatické aminokyseliny vyskytující se v primární sekvenci proteinů (tryptofan, tyrosin, fenylalanin a histidin). Jeho aromaticita je dána postranním řetězcem obsahující planární cyklickou část - indol, která splňuje potřebný počet π-elektronů odpovídající Hückelovu pravidlu 4n+2. Tryptofan se vyznačuje největší plochou povrchu postranního řetězce ze všech aminokyselin (Chothia, 1976) a zároveň nejvzácnějším výskytem v proteinech.

Tryptofan má v proteinech mnoho specifických funkcí, proto je důležité se jím zabývat. Jedná se o stabilizační vlastnosti, jako například u α-helixů (Bhattacharyya et al., 2002; Situ et al., 2018), β-vlásenek (Diana et al., 2018; McCaslin et al., 2019; Santiveri a Jimenez, 2010), tryptofanových zipů (Cochran et al., 2001; Fesinmeyer et al., 2004) nebo termostabilních proteinů (Kannan a Vishveshwara, 2000). Má vliv na funkci a sbalování proteinu (Raimondi et al., 2011; Waters, 2004), stejně tak na funkci například bakteriálních toxinů (Gerhard et al., 2005; Kunthic et al., 2011; Padilla et al., 2006). Navíc má centrální roli v přenosu elektronů v proteinech (Stubbe a van der Donk, 1998; Warren et al., 2012). Všechny tyto vlastnosti jsou dány především tím, že se účastní mnoha různých nekovalentních interakcí.

6.2 Metody výzkumu

Obecně jsem zaznamenal dva hlavní směry výzkumu interakcí mezi molekulami. Za prvé pomocí složitých výpočetních metod (ab initio), při kterých se počítá energie jednotlivých molekulových konformací. Například se jedná o studium π-π interakcí v pracích o aromatických klastrech (Sun a Bernstein, 1996) a benzenových dimerech (Jaffe a Smith, 1996; Lee et al., 2019). Benzen v tomto případě byl použit jako nejjednodušší reprezentativní aromatická molekula. Nicméně se ukázalo, že to není úplně vhodný model, a proto se například pro studium π-π interakcí fenylalaninu v proteinech začal využívat i toluen (Chipot et al., 1996; Ishikawa et al., 1996). Ab initio výpočty se například používají i u dalších interakcí jako CH-π interakce (Liu et al., 2015; Tsuzuki et al., 2000).

Pro studium interakcí mezi aminokyselinami v proteinech se s rostoucím množstvím vyřešených proteinových struktur využívá spíše statistický přístup a někdy i v kombinaci s energetickými funkcemi (de Freitas a Schapira, 2017; Kumar et al., 2018; Kumar a Balaji, 2014; Lanzarotti et al., 2011; Lucas et al., 2016; Ninkovic et al., 2014).

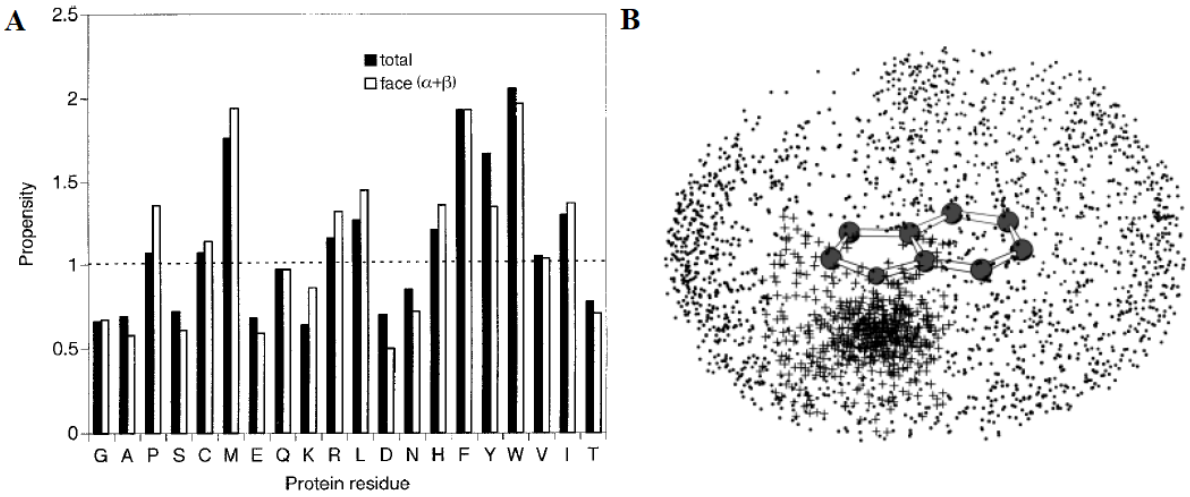
Bohužel energetický přístup je pro mě komplikovaný, jelikož s ním nemám praktickou zkušenost. Zato mám bohatou praxi s velkými objemy dat, a proto jsem se rozhodl pro studium tryptofanu v proteinech jen za pomoci dostupných strukturních dat z PDB databáze (Rose et al., 2017). Všechny další kapitoly budu tedy věnovat pouze článkům, které využívají srovnatelný přístup.

6.3 Vlastnosti prostředí v okolí tryptofanu

I přes popisovanou důležitost tryptofanu ve strukturní biologii jsem našel pouze jednu komplexní práci, která se zabývá jeho bezprostředním okolím, to znamená, jaké aminokyseliny se u něj preferenčně vyskytují. Většina prací se zabývá nějakou konkrétní interakcí, kterou pak autoři cíleně vyhledávají v proteinech, jak bude popsáno v dalších kapitolách. Práce je z roku 2000 (Samanta et al., 2000), v této době bylo relativně málo vyřešených proteinových struktur a autoři jich použili pouze 180. Jeden z parametrů, kterým filtrovali struktury, byla nutnost výskytu alespoň jednoho tryptofanu (já jsem například tento parametr ve své práci navýšil na výskyt alespoň dvou tryptofanů, a to z důvodu, aby každý tryptofan měl v analýze možnost se vyskytovat u jiného). Z těchto proteinů bylo získáno 719 tryptofanů. Autoři rozdělili tryptofan na několik úseků (obě plochy indolu a okraj), a na nich pak vykonali statistickou analýzu. Jako partnerské aminokyseliny byly použity všechny aminokyseliny do 4 Å od jakéhokoli atomu indolu daného tryptofanu.

Nyní víceméně přepíši část jejich závěrů, které jsou důležité pro mojí práci. Výsledky jsou rozděleny na několik bodů. 1) Malé (Gly, Ala), negativně nabité (Asp, Glu) a polární (Ser, Thr) aminokyseliny se vyhýbají indolovému kruhu. 2) U dvou strukturně podobných aminokyselin má ta polárnější menší preferenci k indolu. 3) Arginin je preferovanější, než lysin. 4) Valin je k tryptofanu neutrální, zatímco podobné aminokyseliny s delšími řetězci (Leu, Ile) jsou preferované. 5) Aromatické aminokyseliny mají velkou preferenci, pokud jsou posuzovány společně. 6) Prolin má celkovou preferenci jen k ploše indolu. 7) Výpočty ukázaly, že některé aminokyseliny mají silnější preferenci ke konkrétní ploše indolu (Met, Phe, Ile, Arg, Trp).

Na obrázku 6.1 (A) jsou zobrazeny výsledky autorů pro jednotlivé aminokyseliny a jejich preference k tryptofanu. Na obrázku 6.1 (B) je zobrazení pozic kyslíkových atomů z okolních aminokyselin nalezených v blízkosti indolu.



Obrázek 6.1: (A) Preference aminokyselin být u tryptofanu (černé sloupce, prahová hodnota 1), včetně rozdělení jestli jsou spíše u plochy indolu (bílé sloupce). (B) Zobrazení kyslíkových atomů okolo indolu (tečky). Ty, které tvoří vodíkovou vazbu, jsou zobrazeny křížkem. (Převzato z Samanta et al., 2000)

Tato práce mě velmi zaujala, už jen proto, že je ojedinělá. Bohužel v ní postrádám prostorové rozložení jednotlivých aminokyselin okolo tryptofanu a vynechání problematiky sekvenčně blízkých aminokyselin, které z definice budou nalézány prostorově blízko zkoumaného tryptofanového zbytku.

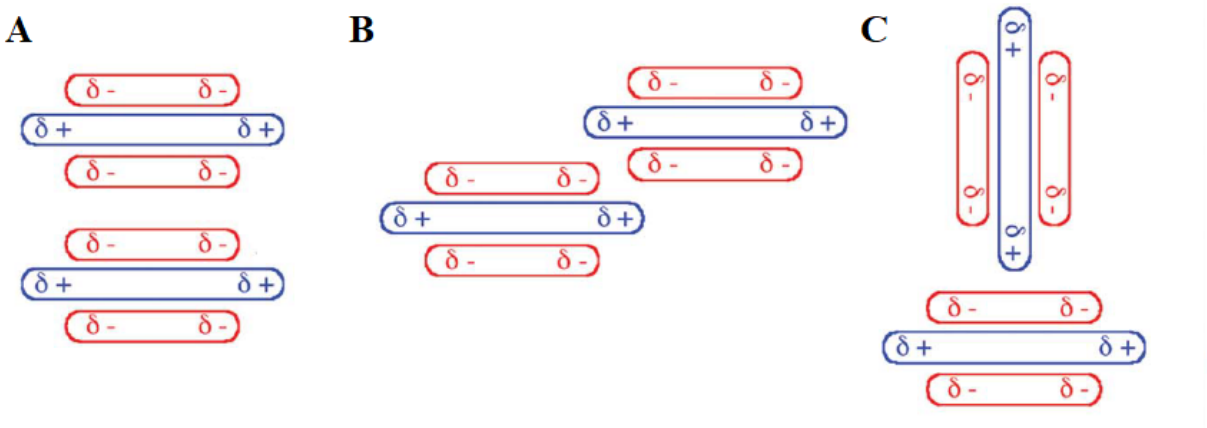
Značnou část své práce věnuji podobné analýze, ale postupy jsem zvolil vlastní.

6.4 Interakce tryptofanu

6.4.1 Aromatické interakce tryptofanu

Jako aromatické interakce jsou označovány interakce mezi aromáty, na kterých se účastní jejich delokalizované π elektrony. Aromatické interakce zastávají v proteinech důležité funkční a stabilizační role. Účastní se mezi-proteinových interakcí (Ma et al., 2003), interakcí proteinu s ligandem (Cotte et al., 2000), přispívají stabilitě proteinu (Kannan a Vishveshwara, 2000) a napomáhají samotnému sbalování proteinu do nativní konformace (Serrano et al., 1991). Není tedy překvapivé, že jsou intenzivně zkoumány v mnoha pracích (Bhattacharyya et al., 2002; Budyak et al., 2013; Burley a Petsko, 1985; Chourasia et al., 2011; McGaughey et al., 1998; Zhuang et al., 2019).

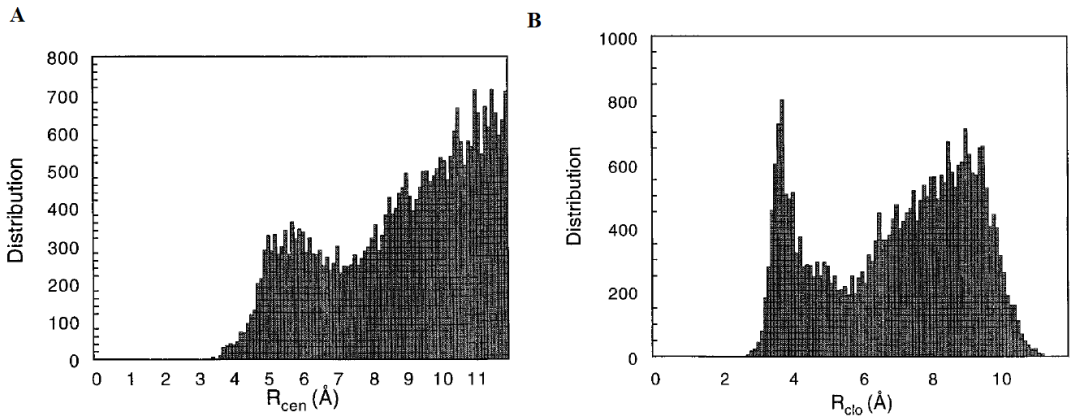
Nejčastěji je studovaná vzájemná orientace aromatických částí molekul. Na obrázku 6.2 jsou znázorněny obecné orientace, které aromatické aminokyseliny vůči sobě zaujímají. Zkoumány jsou také celé aromatické klastry (Lanzarotti et al., 2011). Ve své práci se budu zabývat také interakcemi párů tryptofanů, proto nyní popíši studie, které je rovněž zkoumaly.



Obrázek 6.2: (A) Paralelní „stacking“ orientace, neboli „face-to-face“. (B) Orientace „off-centered“. (C) „Edge-to-face“ orientace, neboli „T-shaped“. Červeně a modře jsou znázorněny oblasti různého parciálního náboje v okolí aromátu. Všechny molekuly jsou zde schematicky zobrazeny z bočního pohledu. (Převzato z Martinez a Iverson, 2012)

Jedna z prvních prací, která se zabývala interakcí aromatických aminokyselin pomocí vyřešených proteinových struktur, byla publikována v roce 1985. Autoři analyzovali 580 aminokyselinových párů (pouze Phe, Trp a Tyr), které byly k sobě blíž než 7 Å. Data pocházela z 35 proteinových struktur. Výsledek ukázal, že 60 % těchto aminokyselin spolu interagují, a to v úhlu okolo 90° a 80 % z nich tvoří interakční síť tří a více. Autoři předpokládali, že tyto interakce mohou přispívat ke stabilizaci přirozené konformace proteinu (Burley a Petsko, 1985). Přesvědčení, že interagující aromatické aminokyseliny sdílí v proteinech nejčastěji T-shaped (kolmou) konfiguraci, sdíleli ve svých publikacích postupně i další autoři (Blundell et al., 1986; Hunter a Sanders, 1990; Thornton et al., 1988).

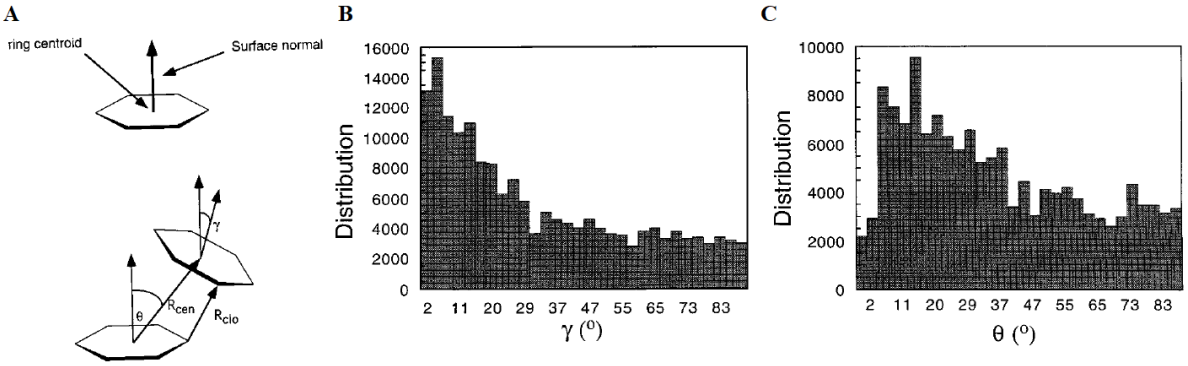
McGaughey et al. v roce 1998 publikovali článek s komplexnější analýzou (McGaughey et al., 1998). Analyzovali kombinované páry všech čtyř aromatických aminokyselin z 505 nehomologních proteinů. Výsledkem bylo 30 444 aminokyselinových párů, které měly vzdálenosti centra kruhů (u Trp menšího pyrrolového kruhu) menší než 12 Å (R_{cen}) a jejich výsledek je na obrázku 6.3 (A). Výsledek je bimodální s minimem okolo 7,5 Å. Druhý graf na obrázku 6.3 (B) zobrazuje distribuci vzdáleností mezi nejbližšími atomy (R_{clo}) aromatických cyklů v aminokyselinovém páru. Výsledek je opět bimodální s minimem mezi 4,5-5 Å. Autoři interpretují levou část od tohoto minima jako místo, kde probíhá interakce mezi kruhy, zatímco vpravo od této oblasti je kvůli tepelnému pohybu molekul jakákoli informace o interakci ztracena.



Obrázek 6.3: (A) Distribuce vzdáleností centra cyklů mezi aromatickými aminokyselinami. (B) Distribuce vzdáleností dvou nejbližších atomů dvou aromatických zbytků. Pokles v pravé části grafu je artefakt způsobený omezením vzdálenosti center na 12 Å (Převzato z McGaughey et al., 1998)

Dále autoři analyzují pouze páry, které mají R_{cen} menší než 7,5 Å nebo R_{clo} menší než 4,5 Å, což odpovídalo 1 682 párům aromatických aminokyselin. Mezi těmito aromáty vypočítali dva úhly, jak je zobrazeno na obrázku 6.4 (A). Úhel Θ, který popisuje vzájemnou prostorovou pozici, a druhý úhel γ, který popisuje jejich vzájemné natočení (úhel mezi normálami aromatických kruhů).

Jedna z nejdůležitějších věcí, na kterou podle mě autoři McGaughey et al. poukázali je, že při studiu četnosti úhlů mezi dvěma molekulami je nutné pozorované histogramy normalizovat na pravděpodobnost, s jakou bychom danou orientaci nalézali v případě velkého množství náhodně rozmístěných aminokyselin. Tato pravděpodobnost naprosto není stejná pro jednotlivé zkoumané úhly. Naopak, je nejpravděpodobnější, že dvě náhodně rozmístěné molekuly nalezneme v kolmém uspořádání a prakticky nikdy ne v paralelním (McGaughey et al., 1998). Tato nutnost je mnoha autory opomíjena, a proto na to budu dále v kapitole upozorňovat. Výsledky normalizovaných úhlů jsou na obrázcích 6.4 (B) a 6.4 (C).

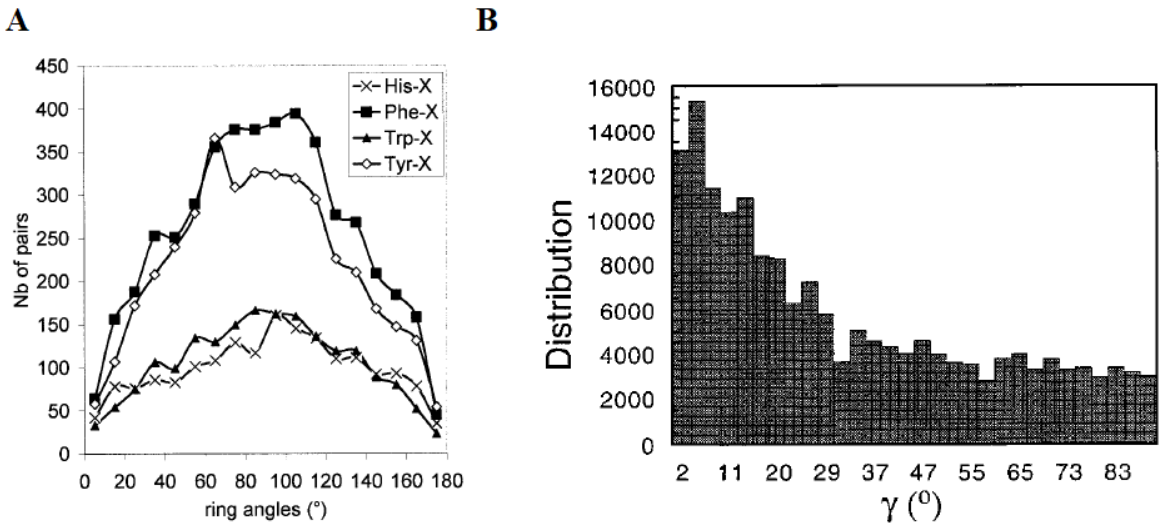


Obrázek 6.4: (A) Zobrazení parametrů použitých pro analýzu (převzato z McGaughey et al., 1998). (B) Normalizované četnosti úhlu γ. (C) Normalizované četnosti úhlu Θ (Převzato z McGaughey et al., 1998)

Autoři navíc upozorňují na problém klastrů, kdy 3 a více kruhů u sebe se chová jinak, než pár, proto analyzovali zvlášť izolované páry a izolované trimery. Výsledek jejich analýzy je, že preferovaná orientace aromatických kruhů je paralelní off-centered (především z důvodu normalizace na pravděpodobnost výskytu), což bylo v rozporu s předchozími pracemi. Je určitě dobré zmínit, že tato práce byla jíž 730 krát citována a dá se tedy pokládat za věrohodnou a zásadní.

V roce 2002 vyšla zajímavá práce, která analyzovala 593 nehomologních proteinů (Thomas et al., 2002). Opět byly vytvořené kombinované páry aromatických aminokyselin se vzdáleností menší než 5,5 Å mezi dvěma nejbližšími atomy postranního řetězce. Četnosti výsledných úhlů mezi aminokyselinami jsou na obrázku 6.5 (A). Autoři výsledky nenormalizovali, a proto se značně liší od výsledků z publikace McGaughey et al. na obrázku 6.5 (B), které byly získány v podstatě totožně.

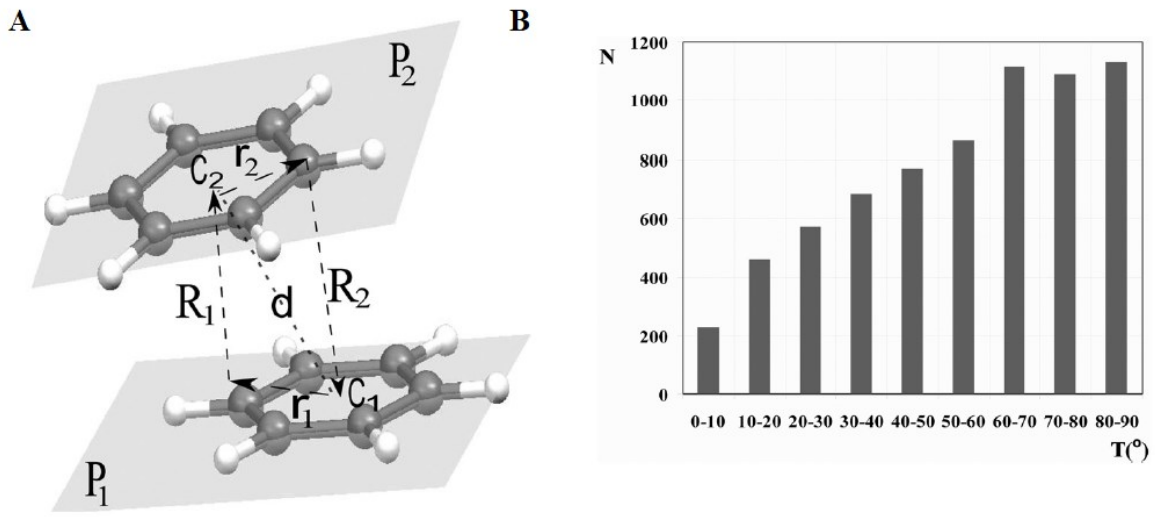
V této publikaci výsledky ukazují na preferenci kolmé orientace aromatických aminokyselin stejně jako některé předchozí analýzy (Burley a Petsko, 1985; Singh a Thornton, 1985). Kromě toho si autoři také uvědomili možný vliv sekundárních struktur na výsledky. Proto rozdělili páry aminokyselin na „v sekvenci blízké“ a „v sekvenci vzdálené“ (vzdálenější než 5 aminokyselin) a nad těmito skupinami provedli další analýzu. Jejich výsledek je, že aromatické aminokyseliny blízké v sekvenci stabilizují lokální struktury a vzdálené stabilizují terciální strukturu. Nicméně z důvodu malé četnosti blízkých párů v sekundárních strukturách prý podporuje hypotézu, že sekundární struktury se formují před párováním aromatických aminokyselin (Thomas et al., 2002).



Obrázek 6.5: (A) Výsledné četnosti úhlů normál mezi páry aromatických aminokyselin z článku Thomas et al. (Thomas et al., 2002). (B) Normalizované četnosti úhlů normál mezi páry aromatických aminokyselin z článku McGaughey et al. (McGaughey et al., 1998).

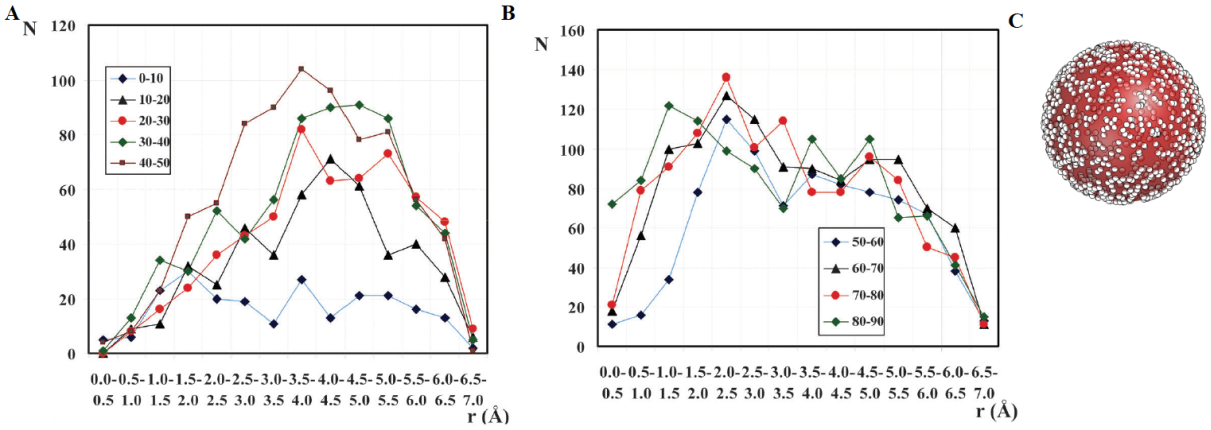
Další práce, která stojí za zmínku, je ohledně preferovaného horizontálního posunu v aromatických interakcích Ninkovic et al. (2014). Jako studovanou aminokyselinu zvolili fenylalanin, a to z důvodu porovnání s prací ohledně benzenových dimerů, kde byla identifikována paralelní (stacking) interakce s velkým horizontálním posunem (3,5-5,0 Å) (Lee et al., 2007). Tato interakce u benzenových dimerů byla stanovena jako energeticky výhodnější než face-to-face orientace.

Ninkovic et al. použili 6 919 fenylalaninových párů z blíže neurčeného počtu proteinů. Pár určili tak, že druhý z fenylalaninů se musel vyskytovat v elipsoidu prvního ($r = 7{,}0$ Å a $R = 6{,}0$ Å, geometrické parametry jsou na obrázku 6.6 (A)). Výsledek úhlů normál na obrázku 6.6 (B) odpovídá předchozím pracím, které kolmou orientaci považují za preferenční. Nicméně opět neřešili normalizaci, a tento výsledek je tedy prakticky nic neříkající.



Obrázek 6.6: (A) Geometrické parametry použité pro popis interakcí. (B) Úhly normál mezi fenylalaninovými kruhy. (Převzato z Ninkovic et al., 2014)

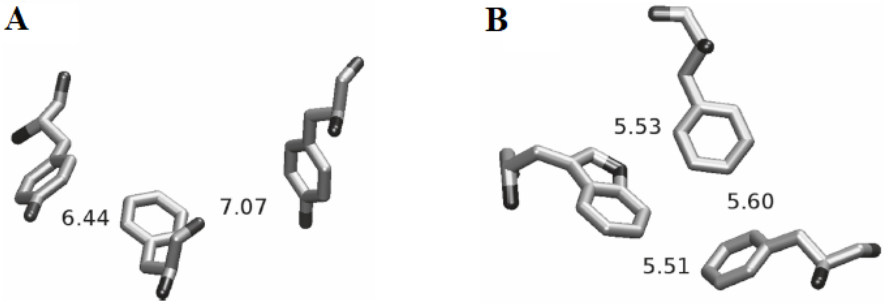
Samotné výsledky horizontálního posunu jsou na obrázku 6.7 (A, B), kde je distribuce horizontálního posunu pro různé úhly normál. Jak píší autoři, pro úhly 0-10° je mírný posun pro hodnoty nad 3 Å. Pro úhly 10-50° má víc 50 % párů posun větší než 3,5 Å. A pro úhly nad 50° jsou četnější menší posuny. Z těchto hodnot usuzují, že „stacking" interakce mezi fenylalaniny v proteinech vykazuje energeticky silný horizontální posun mezi 3,5-5 Å.



Obrázek 6.7: (A, B) Četnosti horizontálního posunu pro různé vzdálenosti a různé úhly normál (převzato z Ninkovic et al., 2014). Barevně jsou rozlišeny různé úhly normál fenylových skupin (viz legenda). Osa x ukazuje intervaly vzdáleností mezi aminokyselinami (čili posun), které jsou trochu nešikovně umístěny pod sebou. Obrázek A ukazuje fenylové skupiny odkloněné o 0-50°. (C) Náhodně rozmístěné body na povrchu koule.

Na této práci chci demonstrovat, jak někteří autoři postupují v analýze naprosto chybně. Poklesy v pravé části grafů jsou dány ořezem prostoru na elipsoid okolo fenylalaninu a je to stejný artefakt, jaký je vidět na obrázku 6.3 (B). Z toho vyplívá, že přeci nelze studovat tvar nějaké křivky (pozice jejího maxima), když ji sami vpravo ořezávají. Dále levá část grafu je také zavádějící. Je tam opět problém náhodného rozdělení v prostoru, jak bylo uvedeno výše. Když se rozmístí kolem těžiště fenylalaninu pravidelně (nebo náhodně) jiná těžiště fenylalaninu, tak jich přesně s posunem 0 Å (tedy dole a nahoře) bude jen pár, řádově jednotky až nula (viz obrázek 6,7 (C)). Zatímco s velkým posunem jich bude o hodně víc, jelikož mají tu možnost. Autoři sice použili elipsoid, ale problém zůstává. Takže ve výsledku graf měl být normalizován a pravá část neměla být ořezána. Z grafů na obrázku 6.7 (A, B) nelze z těchto důvodu vyčíst nic o nějaké preferenci.

Jak už jsem zmínil, někteří autoři se věnují i jakýmsi klastrům aromatických aminokyselin. Ve své práci jsem na ně také narazil, ale bohužel mi nezbyl prostor se jimi detailněji zabývat. Rád bych tedy zmínil například práci Lanzarotti et al. (Lanzarotti et al., 2011), která se jako první věnovala těmto strukturám analyzováním proteinových struktur. Jejich závěr je, že u téměř 50 % studovaných proteinů se nacházejí trimery, tetramery a i početnější klastry aromatických molekul. Pro ukázku uvádím podobu dvou nejčastějších konformací trimerů na obrázku 6.8.



Obrázek 6.8: Prostorové uspořádání skupin aromatických aminokyselin. (A) Konformace trimerů typu „Ladder“. (B) Konformace trimerů typu „Symmetric“. (Převzato z Lanzarotti et al., 2011)

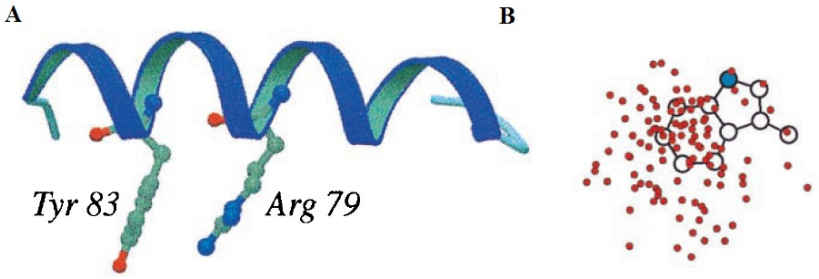
6.4.2 Kationt-π interakce

Termín kationt-π interakce označuje nekovalentní interakci mezi pozitivně nabitou skupinou (iontem) a molekulou s π elektrony. Tyto interakce se uplatňují například ve vázání ligandů (Kumar et al., 2018) v katalytické aktivitě proteinů (Tu et al., 2017) nebo struktuře a funkci proteinu (Peter et al., 2014).

V roce 1981 bylo zjištěno, že iont draslíku se váže energeticky výhodněji na izolovanou molekulu benzenu spíše než na molekulu vody (Sunner et al., 1981). Toto zjištění vedlo k dalším studiím kationtů a π systémů (Amicangelo a Armentrout, 2000; Ma a Dougherty, 1997) i pracím zabývající se významem těchto interakcí v proteinech (Burley a Petsko, 1986; Gallivan a Dougherty, 1999). Díky těmto a dalším pracím bylo prokázáno, že kationt-π interakce je silná nekovalentní vazba přispívající k sekundární struktuře proteinu. Tyto interakce zahrnují v proteinech kationtové postranní řetězce aminokyselin lysinu a argininu a aromatické aminokyseliny (Dougherty, 2007, 2013; Gallivan a Dougherty, 1999; Mahadevi a Sastry, 2013). Navíc dřívější i nedávné práce odhalují, že arginin je v těchto interakcích významnější než lysin (Flocco a Mowbray, 1994; Gallivan a Dougherty, 1999; Kumar et al., 2018).

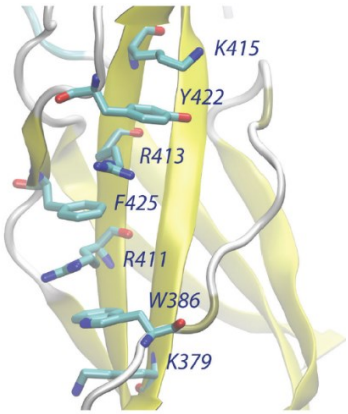
Statistický výzkum tohoto jevu pomocí PDB databáze byl proveden v roce 1999 (Gallivan a Dougherty, 1999). U 593 proteinů analyzovali kationt-π interakce pomocí vypočtené vazebné energie. Do výzkumu zahrnuli 3 aromatické aminokyseliny (Phe, Tyr a Trp) a pozitivně nabité skupiny (Lys, Arg). První zajímavé zjištění bylo, že 70 % argininů se nacházelo blízko nějakého aromatického postranního řetězce. Dále, že v kationt-π interakcích se častěji nachází arginin než lysin. Toto se snaží vysvětlit tím, že postranní řetězec argininu je větší než u lysinu, a navíc je méně rozpustný ve vodě. To mu umožňuje lépe interagovat s aromátem pomocí van der Waalsových sil. Navíc odkazují na práci

Mitchell a spol. (Mitchell et al., 1994), kde bylo naznačeno, že arginin může vytvářet vodíkové vazby, zatímco je vázán na aromatický kruh. Toto lysin při vazbě na aromatický kruh nedokáže. Jako nejvíce překvapivou věc uvádí, že 26 % tryptofanů zahrnují alespoň jednu kationt-π interakci a docházejí k výsledku, že tryptofan má největší schopnost tvořit tyto interakce ze všech studovaných aromatických aminokyselin. V neposlední řadě ukazují, že tyto interakce jsou využívány ke správné orientaci postranních řetězců ve sbaleném proteinu, jak je ukázáno na obrázku 6.9 (A). Výsledek pro interakce Lys-Trp je na obrázku 6.9 (B), kde je vidět, že dusík lysinu se nachází především okolo šestičlenného kruhu indolu.



Obrázek 6.9: (A) Ukázka silné kationt-π interakce v helixu, která orientuje postranní řetězce zúčastněných aminokyselin. (B) Kationt-π interakce dvojic Lys-Trp v prostoru. Červené tečky označují aminoskupinu na konci lysinového řetězce. (Převzato z Gallivan a Dougherty, 1999)

V roce 2017 vyšla zajímavá statistická analýza PDB databáze od autorů Silvana Pinheiro a spol. (Pinheiro et al., 2017). Ve 21 000 unikátních proteinových strukturách hledali málo prozkoumané komplexy kationt-π-kationt interakcí, jejichž příklad je na obrázku 6.10. Analýza ukázala, že tento strukturní motiv je vysoce konzervovaný a je běžný u 7 % proteinů, které obsahují alespoň jednu kationt-π interakci. To naznačuje nějakou strukturní nebo funkční roli.

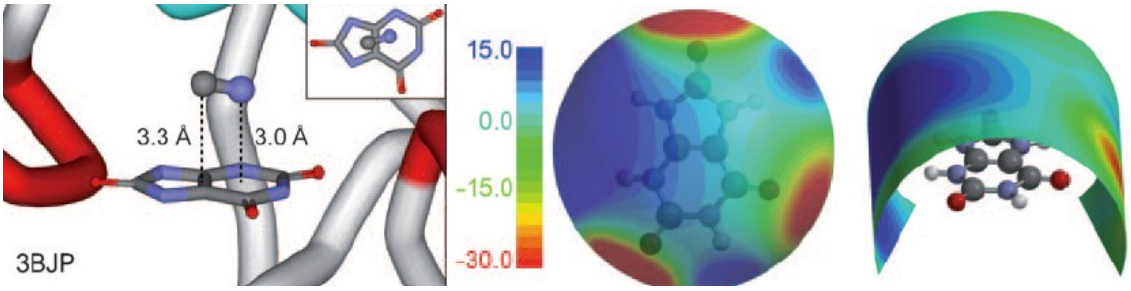


Obrázek 6.10: Mnohonásobné kationt-π-kationt interakce nalezené v lidském receptoru pro růstový hormon (PDB: 1A22) (převzato z Pinheiro et al., 2017).

6.4.3 Aniont-π interakce

Termín aniont-π interakce označuje nekovalentní interakci mezi negativně nabitou skupinou (iontem) a molekulou s π elektrony. Aniont-π interakce byly zkoumány v různých odvětvích studia proteinů. Například v enzymech se jedná o aniont-π katalýzu (Cotelle et al., 2016; Zhao et al., 2018), stabilizace struktur (Pucci a Rooman, 2016; Smith et al., 2017) nebo návrh léčiv (Ellenbarger et al., 2018).

Název aniont-π interakce vznikl v roce 2002, kdy na derivátech benzenu jako hexafluorbenzen a 1,3,5-trinitrobenzene bylo demonstrováno, že se jedná o energeticky preferovanou nekovalentní interakci (Quinonero et al., 2002a; Quinonero et al., 2002b). Jejich biologická důležitost byla prokázána až v roce 2011 na příkladu inhibice enzymatické aktivity urát oxidázy, která je součástí metabolizmu kyseliny močové (Estarellas et al., 2011b). Tato inhibice je reverzibilní a je způsobena kyanidovým aniontem CN^- v aktivním místě (Colloc'h et al., 2008; Conley a Priest, 1980) (viz obr. 6.11)

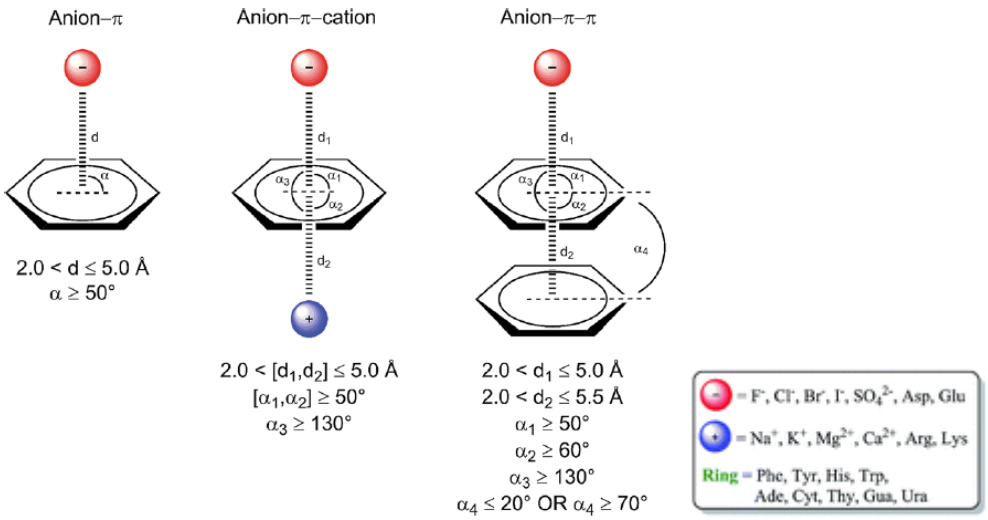


Obrázek 6.11: Aniont π interakce mezi CN^- a kyselinou močovou. Vpravo je její elektrostatický potenciál (převzato z Estarellas et al., 2011b)

V roce 2011 byla také publikována první práce, která se zabývala analýzou PDB databáze, kdy byly hledány a analyzovány kontakty mezi aromatickými aminokyselinami (Trp, Phe, Tyr a His) a anionty (Cl^-, Br^-, F^-, NO_3^-, ClO_4^- a PO_4^{n-}). Autoři nakonec konstatovali, že nalezené kontakty nebyly statisticky významné (Robertazzi et al., 2011). Následovalo více dalších studií, které se tímto tématem zabývaly (Chakravarty et al., 2012; Estarellas et al., 2011a; Jenkins et al., 2013).

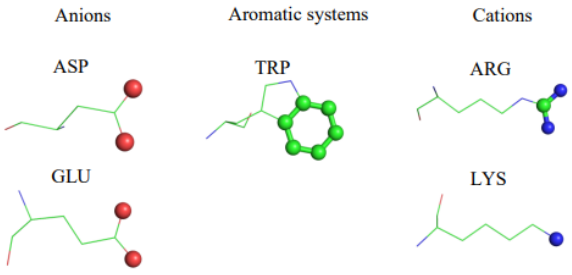
Pro nás je stěžejní práce, kterou v roce 2016 publikoval Xavier Lukas a spol. (Lucas et al., 2016). Ti provedli rozsáhlou studii kontaktů aniontů s aromatickými molekulami (včetně tryptofanu) vyskytujících se v proteinech, protein-DNA a protein-RNA komplexech. Mezi anionty započítali i glutamát s aspartátem, které mají nízké hodnoty pK_a, a tudíž jsou vždy ionizované (Pace et al., 2009). Navíc se zaměřili i na možné kooperativní interakce triád

jako aniont-π-kationt a aniont-π-π, u kterých předpokládají, že hrají stěžejní roli v aniont-π interakcích (viz obr. 6.12).



Obrázek 6.12: Studované aniont-π interakce v článku Xaviera Lukase a spol. a jejich vzájemné uhly a vzdálenosti (převzato z Lucas et al., 2016). Pod každým obrázkem je popis vzdáleností a úhlů, které danou interakci charakterizují. Proč se rozhodli pro tyto mezní hodnoty, není popsáno.

Jak je patrné na obrázku 6.12, autoři se zabývali různými anionty, kationty i aromatickými molekulami. Dále se budu věnovat pouze částem jejich práce, které se týkaly glutamátu a aspartátu (anionty reprezentované atomy kyslíku v postranních řetězcích), tryptofanem (reprezentovaný větším indolovým kruhem) a argininem a lysinem (za kationty jsou pokládány koncové atomy dusíku v postranním řetězci). Detaily jsou na obrázku 6.13.



Obrázek 6.13: Části aminokyselin, mezi kterými byly zkoumány interakce (převzato z Lucas et al., 2016).

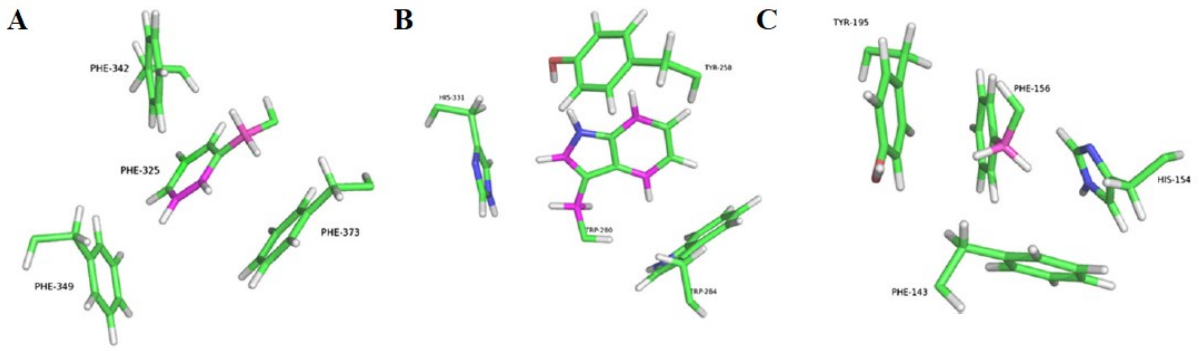
Analýzu provedli nad 38 027 unikátními proteinovými strukturami. Mimo jiné nalezli v proteinech tisíce anion-π interakcí obsahující tryptofany s větší preferencí ke glutamátu. Dále statisticky významné výskyty triád s kationtem (Glu-Trp-Arg a Glu-Trp-Lys) a triád s aromátem (Glu-Trp-His a Glu-Trp-Trp).

6.4.4 CH-π interakce

CH-π interakce je třída nekovalentních interakcí patřící pod nejslabší vodíkové můstky, které nastávají mezi slabou kyselinou (CH) a slabou zásadou (π skupina) (Nishio, 2012). Analýzy proteinových struktur ukázaly, že tyto interakce se vyskytují ve velkých počtech a přispívají ke správné konformaci proteinu (Umezawa et al., 1999). Často se vyskytují v interakci s aromatickými aminokyselinami (Brandl et al., 2001) a identifikují substrát (Balaji, 2011; Spiwok et al., 2004; Umezawa a Nishio, 2005). Například analýza 1 154 proteinových řetězců ukázala, že tři čtvrtiny tryptofanů, polovina fenylalaninů a tyrosinů se účastní těchto interakcí (Brandl et al., 2001).

Rozsáhlá analýza CH-π interakcí z roku 2014 mimo jiné odhalila, že s aromatickým kruhem interaguje mnoho různých CH skupin a v těchto interakcích se nejčastěji vyskytují tryptofan, fenylalanin, prolin a metionin (Kumar a Balaji, 2014). Autoři uvádějí, že i přesto, že leucin, alanin, glycin a valin jsou v proteinu nejčastější aminokyseliny, tak v CH-π interakcí nejsou moc zastoupené.

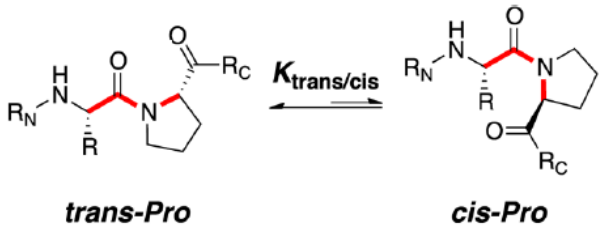
Závěrem bych rád upozornil, že i T-shape orientace aromatických aminokyselin je druh CH-π interakce (Kadam et al., 2013). A stejně tak se CH-π interakce vyskytují i v aromatických klastrech, jak je vidět na obrázku 6.14.



Obrázek 6.14: Popsané mnohonásobné CH-π interakce z článku Kadam et al. (Kumar a Balaji, 2014). (A) Sarcosine oxidáza. (B) Alfa-amyláza. (C) Těžký řetězec myozinu 2.

6.4.5 Prolin-π interakce

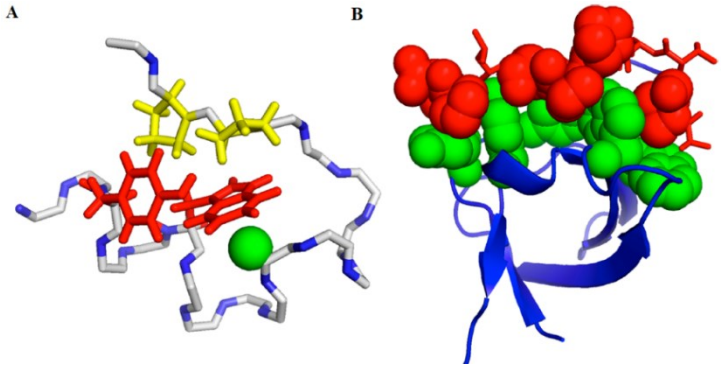
Díky své speciální struktuře má aminokyselina prolin v proteinech unikátní funkci. Její konformační omezení má vliv na strukturu celého proteinu tvorbou cis a trans konfigurací peptidové vazby (Macarthur a Thornton, 1991), které jsou měněny například pomocí peptidyl-prolyl cis/trans isomeráz (Fischer, 1994). Tyto dvě konfigurace jsou ukázány na obrázku 6.15.



Obrázek 6.15: Cis a trans konfigurace peptidové vazby v prolinu (převzato z Zondlo, 2013).

Mnohé práce ukazují, že prolin ve spojení s aromatickou aminokyselinou má silnou preferenci pro cis konformaci, což může ukazovat na preferovanou interakci (Bhattacharyya a Chakrabarti, 2003; Brandl et al., 2001; Stewart et al., 1990; Wu a Raleigh, 1998). Například se uvádí, že 80 % prolinu u tryptofanu zaujímá právě cis konformaci (Wu a Raleigh, 1998). Některé práce naznačují, že tato preference je způsobená CH-π interakcí, kde parciální pozitivně nabitý vodík prolinu interaguje s π oblastí aromatické aminokyseliny (Brandl et al., 2001; Kumar a Balaji, 2014; Umezawa et al., 1999). Ze všech aromatických aminokyselin má největší preferenci k prolinu tryptofan (Zondlo, 2013).

Interakce prolinu s aromatickými aminokyselinami je nejvíce znát v malých proteinech, kde se v konformaci proteinu méně uplatňuje hydrofóbní efekt (Zondlo, 2013). Příklad je uveden na Andersonovým trp-cage miniproteinu (Neidigh et al., 2002) na obrázku 6.16 (A) a interakci SH3 domény bohaté na aminokyseliny s jejím ligandem bohatým na proliny (Musacchio et al., 1994) na obrázku 6.16 (B).



Obrázek 6.16: (A) Ukázka interakce stabilizující trp-cage miniprotein, kde žluté proliny interagují červeným tryptofanem a tyrosinem. Tryptofan dále interaguje CH interakcí se zeleným glycinem (převzato z Neidigh et al., 2002). (B) Interakce SH3 domény bohaté na aminokyseliny (zeleně) s jejím ligandem bohatým na proliny (červeně) (převzato z Musacchio et al., 1994).

6.4.6 Sulfur-π interakce

Pro úplnost se krátce zmíním i o sulfur-π interakci, i když mi na ní bohužel v práci nezbyl prostor. Jedna z prvních prací hledající tuto nekovalentní vazbu v proteinech byla v roce 1985 na 36 proteinech (Reid et al.). Práce analyzovala geometrii cysteinu a methioninu vůči fenylalaninu, tyrosinu a tryptofanu. Už v ní bylo poznamenáno, že se jedná o běžně se vyskytující motiv, a to především v hydrofóbním jádru. Valley et al. provedli rozsáhlý výzkum týkající se methionin-π interakce a zjistili, že třetina proteinů obsahuje alespoň jednu tuto energeticky silnou nekovalentní vazbu (Valley et al., 2012). Rozsáhlá práce z roku 2017 poskytuje komplexní pohled na tyto interakce (Forbes et al., 2017). Další práce, které se zabývají konkrétními proteiny, jsou například ohledně D2 receptoru pro dopamin (Daeffler et al., 2012; Sencanski et al., 2015) nebo methionin-aromatické interakce u oxidoreduktáz (Weber a Warren, 2018).

Touto interakcí bych rád skončil, jelikož žádné další jsem již ve vztahu k tryptofanu nenašel. Četnosti výskytů jednotlivých interakcí budou pravděpodobně záviset na tom, jak je daná interakce silná, ale taky na tom, jak jsou běžné v proteinech funkční skupiny účastnící se dané interakce. A to by se mělo ve statistických pracích zohledňovat.

6.5 Použité programy

Značná část této práce je věnována procházení PDB databáze a vytvoření ideálního datasetu proteinů pro účel analýzy. Téměř všechny články věnující se analýze interakcí v proteinech, které jsem studoval, vycházeli z nějakého již vytvořeného datasetu použitého jinde nebo využili nějaký software pro výběr vhodných proteinových struktur. Uvádím tedy pro představu jejich stručný výpis.

Například McGaughey et al. (1998), Samanta et al. (2000) a další autoři využili algoritmus publikovaný v roce 1994, který se snaží vypořádat s velkou redundancí v PDB databázi velmi podobně jako já (Hobohm a Sander, 1994). Algoritmus tedy například nejdříve odstraňuje 100 % shodné řetězce proteinů a ponechává ten s lepší kvalitou (určené jejich definicí). Vynechává struktury řešené s horším rozlišením než 3,5 Å a ty s větším výskytem neznámých (netypických) aminokyselin. Ve finálním datasetu nakonec není žádná dvojice řetězců, která by měla větší sekvenční podobnost než 25 %. Oproti tomu Ninkovic et al. (2014) použili již novější nástroj PDBSELECT, poskytující v roce 2009 kolem 4500 unikátních proteinů (Griep a Hobohm, 2010). Další nástroj poskytující vhodné struktury proteinů byl například UniProt (Universal Protein Resource) (Wu et al., 2006), použitý autory Lanzarotti et al. (2011). Nicméně autoři uvádějí, že nástroj neposkytuje dostatečně odlišné proteiny a museli si dataset dofiltrovat.

Někteří autoři použili externí nástroje i na analýzu, například Thomas et al. (2002) využili pro svojí práci analytický nástroj Pex (Thomas et al., 2001). Ten obsahuje informace například o sekundárních strukturách, dihedrálních úhlech a vodíkových vazbách.

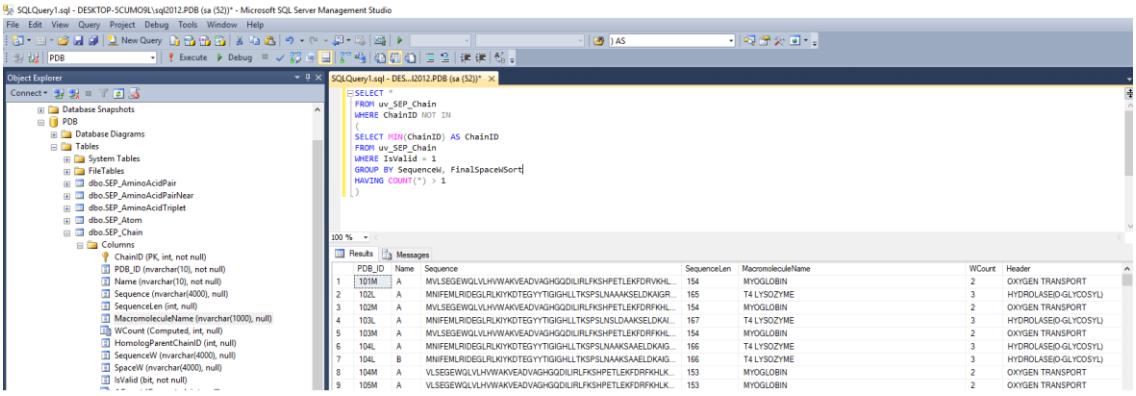
Při čtení této práce může vzniknout otázka, proč jsem také nepoužil nějaké externí nástroje. Domníval jsem se, že bych se tím ochudil o spoustu zajímavých informací a hlavně o zkušenosti, které mi detailní práce s PDB soubory poskytla. Všechny data k analýzám jsem také vytvářel od počátku, abych měl o všem detailní přehled. Vycházel jsem pouze ze sekvence a pozic atomů uložených v PDB souborech.

7 Materiál a metody

7.1 Zdroj dat a vývojové prostředí

Pro získání proteinových struktur jsem použil databázi PDB (Protein Data Bank)[1]. Je to rozsáhlý volně přístupný archiv experimentálně vyřešených biologických makromolekul včetně proteinů (Rose et al., 2017). Tyto makromolekuly jsou v ní uloženy v takzvaných PDB souborech, což jsou formátované textové soubory, obsahující veškeré podrobné informace včetně 3D struktury. Popis těchto souborů je uveden v kapitole 7.2. Poslední stažení všech proteinových struktur PDB databáze jsem provedl dne 28.10.2018, takže v této práci nejsou obsaženy struktury novějšího data. Jednalo se o 144 960 PDB souborů ve 27,2 GB dat.

Z těchto textových souborů jsem vytvořil databázi v relačním a analytickém databázovém systému Microsoft SQL[2]. Pojem relační označuje databázi založenou na vzájemně propojených tabulkách, z nichž každá obsahuje položky jednoho konkrétního typu. V mém případě se jedná například o tabulky proteinů, proteinových řetězců, ale i jednotlivých tryptofanů nebo atomů. Tyto tabulky pak obsahují sloupce, které definují vlastnosti konkrétního typu, například u řetězce se jedná o název, sekvenci, počty jednotlivých aminokyselin apod. Pro práci v Microsoft SQL systému se využívá prostředí Microsoft SQL Server Management Studio[3], jehož grafická podoba je na obrázku 7.1. Díky němu lze jednoduše vytvářet a upravovat tabulky a zároveň umožňuje práci s daty pomocí skriptovacího jazyku SQL. Jazyk umožňuje různé logické a aritmetické operace, seskupování dat, zobrazování požadovaných záznamů a mnoho dalšího.



Obrázek 7.1: Ukázka prostředí Microsoft SQL Server Management Studio. Vlevo seznam tabulek. Vpravo nahoře je skript v jazyku SQL a jeho výsledek pod ním.

[1] http://rcsb.org
[2] https://docs.microsoft.com/cs-cz/sql/sql-server/sql-server-technical-documentation
[3] https://docs.microsoft.com/cs-cz/sql/ssms/sql-server-management-studio-ssms

Samotná databáze ale nestačí. Pro složitější výpočty a tvorbu grafů jsem využíval objektově orientovaný programovací jazyk C# a vývojovou platformu .NET Framework[4] ve vývojovém prostředí Microsoft Visual Studio[5]. Pomocí těchto prostředků jsem si vytvářel pomocné webové stránky, které byly napojeny na SQL databázi. To mi umožňovalo tvorbu grafů, rychlé filtrování a zobrazování dat.

Pro grafické zobrazování proteinů a jeho částí jsem využil open source software PyMOL[6]. Tento velmi užitečný nástroj využívám i pro různá 3D zobrazení analyzovaných dat.

Pro kvantifikaci některých dat pomocí Gaussovy funkce jsem použil program fityk[7].

7.2 PDB soubor

PDB soubor se běžně používá pro ukládání souřadnic strukturních modelů získaných např. pomocí NMR a krystalografie. Obsahuje velké množství informací o konkrétním proteinu v podobě formátovaného textu[8]. Každý soubor má v názvu jedinečný čtyřmístný kód složený s číslic a písmen, který je specifický pro daný protein. Text se skládá ze základních částí popisujících protein, jeho řetězce a souřadnice jednotlivých atomů. Dále většinou obsahuje i obrovské množství dodatečných informací, včetně různých variant modelů, výskytů sekundárních struktur a mnoho poznámek od autorů (takzvaná metadata). Z PDB databáze lze ručně stáhnout jednotlivé soubory, ale v případě potřeby i kompletní databázi v komprimovaném stavu pomocí FTP (File Transfer Protocol).

Pro tuto práci jsou nejdůležitější oblasti v PDB souboru SEQRES a ATOM, které jsou vyznačeny na ukázce na obrázku 7.2. SEQRES obsahuje sekvenci aminokyselin jednotlivých proteinových řetězců daného proteinu, čili primární strukturu. V této práci používám slovo řetězec jako odkaz na tuto sekvenci převedenou na jednopísmenné názvy aminokyselin. Pokud PDB soubor obsahuje více řetězců, tak to zpravidla odpovídá situaci, kde reálný protein je tvořen více podjednotkami. Oblast ATOM je složena ze souřadnic jednotlivých atomů aminokyselin.

[4] https://docs.microsoft.com/cs-cz/dotnet/csharp/getting-started/introduction-to-the-csharp-language-and-the-net-framework
[5] https://visualstudio.microsoft.com/cs/
[6] https://pymol.org/2/
[7] https://fityk.nieto.pl/
[8] https://www.cgl.ucsf.edu/chimera/docs/UsersGuide/tutorials/pdbintro.html

```
TITLE     2 HYDROXYL GROUP WITHIN THE CORE OF A PROTEIN DETERMINED FROM ALA TO
TITLE     3 SER AND VAL TO THR SUBSTITUTIONS IN T4 LYSOZYME
COMPND     MOL_ID: 1;
COMPND    2 MOLECULE: T4 LYSOZYME;
COMPND    3 CHAIN: A;
COMPND    4 EC: 3.2.1.17;
COMPND    5 ENGINEERED: YES
SOURCE     MOL_ID: 1;
SOURCE    2 ORGANISM_SCIENTIFIC: ENTEROBACTERIA PHAGE T4;
SOURCE    3 ORGANISM_TAXID: 10665;
SOURCE    4 EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID;
SOURCE    5 EXPRESSION_SYSTEM_PLASMID: M13
KEYWDS     HYDROLASE(O-GLYCOSYL)
EXPDTA     X-RAY DIFFRACTION
AUTHOR     M.BLABER,B.W.MATTHEWS
REVDAT    3   29-NOV-17 118L     1       HELIX
REVDAT    2   24-FEB-09 118L     1       VERSN
REVDAT    1   31-OCT-93 118L     0
JRNL        AUTH   M.BLABER,J.D.LINDSTROM,N.GASSNER,J.XU,D.W.HEINZ,B.W.MATTHEWS
REMARK    2
REMARK    2 RESOLUTION.    3.20 ANGSTROMS.
SEQRES    1 A  301  MET VAL ALA PHE LYS GLY VAL TRP THR GLN ALA PHE TRP
SEQRES    2 A  301  LYS ALA VAL THR ALA GLU PHE LEU ALA MET LEU ILE PHE
SEQRES    3 A  301  VAL LEU LEU SER VAL GLY SER THR ILE ASN TRP GLY GLY
SEQRES    4 A  301  SER GLU ASN PRO LEU PRO VAL ASP MET VAL LEU ILE SER
SEQRES    5 A  301  LEU CYS PHE GLY LEU SER ILE ALA THR MET VAL GLN CYS
SEQRES    6 A  301  PHE GLY HIS ILE SER GLY GLY HIS ILE ASN PRO ALA VAL
SEQRES    7 A  301  THR VAL ALA MET VAL CYS THR ARG LYS ILE SER ILE ALA
SEQRES    8 A  301  LYS SER VAL PHE TYR ILE THR ALA GLN CYS LEU GLY ALA
SEQRES    9 A  301  ILE ILE GLY ALA GLY ILE LEU TYR LEU VAL THR PRO PRO
SEQRES    1 B  626  SER ASN ALA MET LYS LYS LEU ARG ASP ASP PHE SER GLU
SEQRES    2 B  626  ASP SER ASP SER ASP ILE PRO GLU LYS PHE THR PRO LYS
SEQRES    3 B  626  THR ASP LEU PHE ASP TYR THR ARG ARG GLU GLU MET ILE
SEQRES    4 B  626  PRO MET ARG ASP GLY VAL LYS LEU ASN THR ILE ILE LEU
ATOM      1  N   THR A  31      23.941  58.504 115.359  1.00 75.67           N
ATOM      2  CA  THR A  31      23.999  58.145 116.802  1.00 75.67           C
ATOM      3  C   THR A  31      22.623  58.400 117.412  1.00 75.67           C
ATOM      4  O   THR A  31      21.804  57.491 117.465  1.00 75.67           O
ATOM      5  CB  THR A  31      25.084  58.992 117.560  1.00 61.77           C
ATOM      6  OG1 THR A  31      26.323  58.948 116.836  1.00 61.77           O
ATOM      7  CG2 THR A  31      25.330  58.440 118.967  1.00 61.77           C
ATOM      8  N   GLN A  32      22.383  59.641 117.837  1.00 55.42           N
ATOM      9  CA  GLN A  32      21.132  60.086 118.473  1.00 55.42           C
ATOM     10  C   GLN A  32      20.006  59.071 118.666  1.00 55.42           C
ATOM     11  O   GLN A  32      19.415  59.006 119.741  1.00 55.42           O
ATOM     12  CB  GLN A  32      20.592  61.358 117.771  1.00 95.99           C
ATOM     13  CG  GLN A  32      21.185  62.688 118.329  1.00 95.99           C
ATOM     14  CD  GLN A  32      20.782  63.921 117.536  1.00 95.99           C
ATOM     15  OE1 GLN A  32      19.596  64.193 117.347  1.00 95.99           O
ATOM     16  NE2 GLN A  32      21.774  64.679 117.076  1.00 95.99           N
```

Obrázek 7.2: Ukázka důležitých částí PDB souboru. Červená oblast obsahuje primární strukturu dvou řetězců A a B (ve třetím sloupci). Zelená oblast souřadnice atomů a modrá ukazuje rozlišení, pokud je protein řešen rentgenovou krystalografií.

Další důležitá informace je, jakou metodou byla struktura konkrétního proteinu získána. Ze všech mnou stažených proteinů bylo 90 % řešeno pomocí rentgenové krystalografie, 7 % NMR spektroskopií a zbytek pomocí méně známých metod. Problém nastává ve chvíli, kdy potřebuji porovnat dva proteiny a určit, který z nich je v lepší kvalitě. Tato situace se v mé práci vyskytuje často, především při odstraňování homologie.

V rentgenové krystalografii je určující, v jakém rozlišení je protein vyřešen (udáváno v Angstromech, v PDB souboru za RESOLUTION). Čím menší je míra nejistoty umístění atomu v prostoru, tím je pochopitelně vyšší (lepší) rozlišení modelu (Smyth a Martin, 2000). Čím lepší je rozlišení, tím menší je udané číslo. U modelů získaných pomocí NMR jsem podobné jednoznačné kritérium kvality celého modelu nedohledal. Použitelnost této metody je navíc limitovaná velikostí proteinů (Marion, 2013). Dostupná literatura uvádí, že obě metody se navzájem doplňují, ale ne, která je obecně kvalitnější (Snyder et al., 2005). Nakonec jsem se rozhodl při výběru reprezentativních zástupců homologních proteinů upřednostňovat rentgenovou krystalografii, a to z prostého důvodu, že je v PDB databázi nejčastější. To znamená, že pokud potřebuji z nějaké skupiny modelů proteinů vybrat ten nejlépe reprezentativní, tak zvolím ten, který byl získán pomocí krystalografie, a to s nejlepším rozlišením. V případě, že není u žádného proteinu rozlišení dostupné, tak zvolím ten, který je nejnovějšího data. Vzhledem k tomu, že analýzu řeším na úrovni jednotlivých řetězců, tak je na tento postup následně odkazováno v dalších částech práce, jako ponechání toho nejlepšího řetězce v rámci nějaké skupiny homologních řetězců.

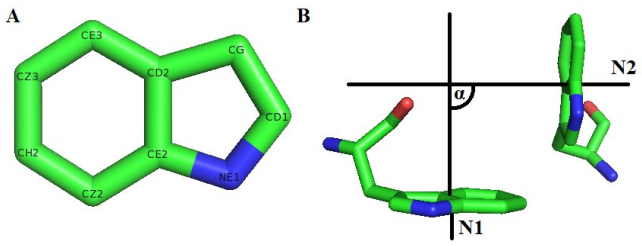
7.3 Levenshteinova vzdálenost

Na mnoha místech této práce je potřeba řešit podobnost mezi dvěma primárními strukturami proteinových řetězců nebo jejich částí. Tento problém se vyskytuje především při odstraňování homologie. Typicky používaný nástroj v bioinformatice pro srovnávání primárních struktur je BLAST (Altschul et al., 1990). Je to algoritmus, který využívá statisticky signifikantní úseky pro rychlé vyhledávání dlouhých řetězců ve velkých databázích. Pro moje účely se ale ukázal jako nevhodný, protože jsem potřeboval jednoduchý nástroj na srovnávání menšího počtu krátkých sekvencí, který jsem nalezl v podobě Levenshteinovy vzdálenosti (Navarro, 2001). Tuto vzdálenost zavedl v roce 1965 Vladimir Levenshtein a je charakterizovaná jako minimální počet operací typu substituce, inserce a delece, aby po jejich provedení byly dvě porovnávané písmenné sekvence totožné. Vzhledem ke složitým matematickým formulacím při přesné definici této veličiny odkazuji v tomto případě spíše na wikipedii[9].

[9] https://en.wikipedia.org/wiki/Levenshtein_distance

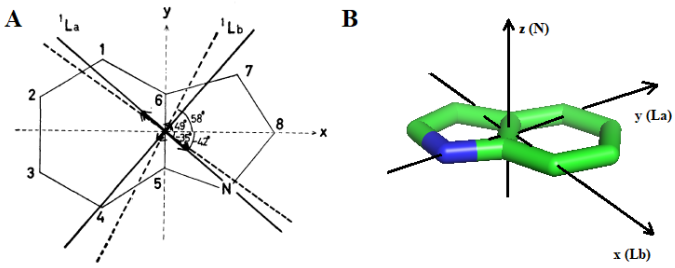
7.4 Molekula indolu v analýze

Jelikož se tato práce zabývá tryptofanem, proberu v této kapitole přístup k dané molekule v rámci analýzy. Do analýz jsem bral v úvahu pouze indolovou část tryptofanu bez vodíků, jak je zobrazena na obrázku 7.3 (A). Centrum indolu, které označuji jako těžiště, jsem stanovil mezi atomy CD2 a CE2. Od tohoto bodu jsou následně dohledávány nejbližší atomy postranních řetězců okolních aminokyselin nebo je využíván pro výpočet vzdáleností dvou indolů v analýze tryptofanových párů. Další údaj používaný v analýzách je úhel normál mezi indoly. Normála je dána jako kolmice na plochu danou atomy CD1, CE3 a CZ2. Příklad pravého úhlu mezi normálami je na obrázku 7.3 (B).



Obrázek 7.3: (A) Indolová skupina tryptofanu s pojmenovanými atomy. (B) Ukázka pravého úhlu mezi normálami dvou indolů.

Pro porovnávací analýzy bylo potřeba větší množství různých indolů překrýt přes sebe, aby bylo možno porovnat jejich okolí. Z tohoto důvodu jsem je vkládal těžištěm do počátku souřadnicového systému. Normálu určenou výše jsem proložil osou z. Osy x a y jsem stanovil na základě tranzičních dipólů indolu La a Lb (Yamamoto a Tanaka, 1972), které jsou zobrazeny na obrázku 7.4 (A). Osu y jsem tedy vedl od dusíku NE1 k těžišti indolu, což odpovídá dipólu La. Osa x podle dipólu Lb vede od těžiště indolu k atomu CZ2. Vložení je zobrazeno na obrázku 7.4 (B). Samotné dipóly La a Lb jsou zásadní pro spektroskopické vlastnosti indolu, ale já se jimi v této práci nebudu dále zabývat.

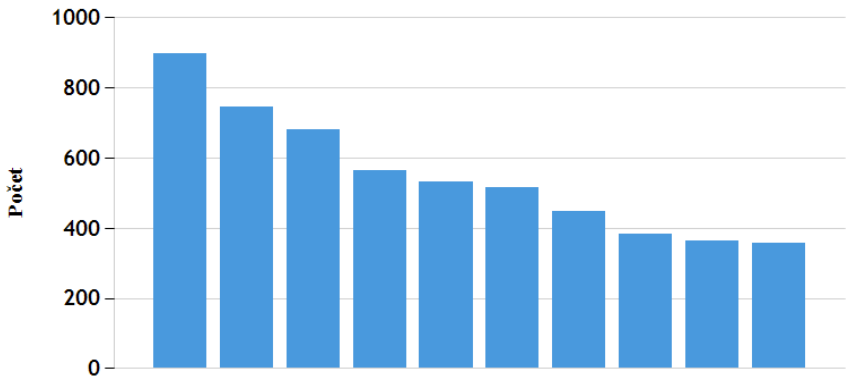


Obrázek 7.4: (A) Tranziční dipóly indolu La a Lb (převzato z Yamamoto a Tanaka, 1972). (B) Indol a jeho těžiště umístěné v počátku souřadnicového systému na základě tranzičních dipólů, jak jsem je definoval já ve své práci.

7.5 Tvorba analyzovatelného datasetu

7.5.1 Úvod

Pro získání dat jsem použil PDB databázi, která bohužel obsahuje velké množství homologních proteinů, což je znázorněno na grafu deset nejčastějších názvů stažených makromolekul (viz obr. 7.5). Na tomto grafu je patrné, že například β-2-mikroglobulin nebo Lysozym je zkoumán velmi často. Navíc pro stejné nebo podobné proteiny se vyskytují odlišné názvy, takže počty prakticky shodných modelů mohou být reálně daleko větší. Tyto homologní proteiny mohou značně narušovat výsledky statistických analýz. Například pokud dojde k objevení, že určitý počet proteinů obsahuje jistou tryptofanovou interakci, tak tento výsledek nebude relevantní, pokud se bude jednat o samé lysozymy. Z tohoto důvodu je nutné takovéto homology v databázi identifikovat a nejlépe odstranit před jakoukoliv analýzou. Autoři, kteří se zabývali podobnými analýzami jako já, použili různé postupy (téměř vždy externí) na odstranění homologie, jak jsem popsal v kapitole 6.5. V této kapitole jsem uvedl i důvody, proč jsem se rozhodl pro vlastní postup.



Název	Počet
BETA-2-MICROGLOBULIN	897
UNCHARACTERIZED PROTEIN	745
LYSOZYME C	681
CARBONIC ANHYDRASE 2	563
ENDOTHIAPEPSIN	532
LYSOZYME	516
BETA-LACTAMASE	446
MYOGLOBIN	383
BETA-SECRETASE 1	362
DIHYDROFOLATE REDUCTASE	358

Obrázek 7.5: Deset nejčastějších názvů makromolekul v PDB databázi z celkového počtu 141 058 PDB souborů obsahujících proteiny.

Nejjednodušší způsob odstranění homologů by tedy pravděpodobně bylo určit procento maximální přípustné podobnosti primární struktury řetězců. Pro cíle mé práce mi ale tento způsob přišel příliš radikální. Jelikož se zabývám především tryptofany, mohl bych tímto způsobem ztratit některá zajímavá data. Například v PDB databázi existují proteiny, které jsou sice homologní, ale liší se počtem tryptofanů. Pokud bych odstraňoval homologie pouze na úrovni podobnosti primární struktury (např. s použitím nástroje BLAST), mohl bych o tento teoreticky důležitý rozdíl v uspořádání tryptofanových zbytků přijít.

Především jsem musel rozhodnout, zda analyzovat interakce na úrovni celých proteinů nebo na úrovni jednotlivých proteinových řetězců. Problém analýzy celých proteinů je, že se některé skládají z opakujících se stejných nebo podobných podjednotek (například virové kapsidy) a jakákoliv jejich vnitřní aminokyselinová interakce by pak byla v analýzách nadhodnocena. Navíc jsem nedokázal přijít na způsob, jakým bych hromadně porovnával mezi sebou proteiny s různým počtem řetězců. Z těchto důvodů jsem se tedy rozhodl pro analýzu na úrovni řetězců, jejíž nevýhodou může být ztráta interakcí tryptofanů mezi řetězci v rámci proteinu.

7.5.2 Získání řetězců

Základem je získání vzorku ze všech dostupných řetězců v PDB databázi, který by nebyl zatížen homologií a zároveň by obsahoval všechny možné tryptofanové interakce. Z PDB databáze jsem dne 28.10.2018 získal 144 960 PDB souborů, které obsahovaly 393 041 proteinových řetězců a vše pro mě důležité jsem naimportoval do SQL databáze. Vzhledem k tomu, že se zabývám především interakcemi tryptofanů v rámci řetězců, odstranil jsem všechny řetězce, které mají méně než dva tryptofany. Tím mi zůstalo 241 923 řetězců.

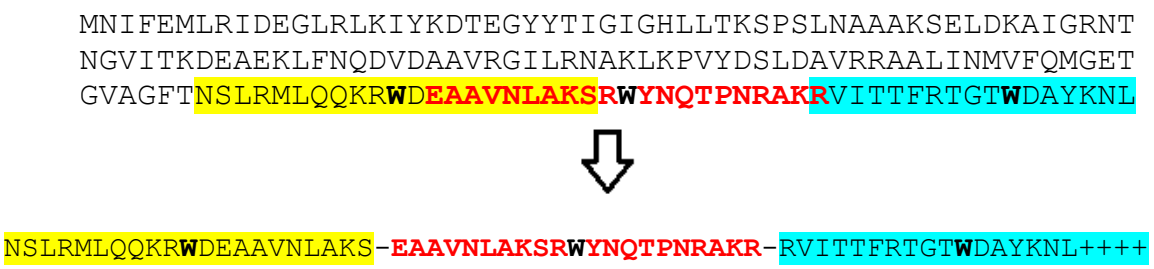
Jak jsem již zmínil, PDB databáze obsahuje obrovské množství stejných nebo velmi podobných proteinů, tudíž i opakujících se řetězců, které je nutné odstranit. Odstranění nadbytečných řetězců se shodnou primární strukturou je v podstatě snadné, stačí je podle ní seskupit a ponechat ten nejlepší řetězec (viz Kapitola 7.2). Takto jich zůstalo jen 63 629. Toto čtyřnásobné snížení počtu je vcelku překvapivé, protože to ukazuje, v jak velkém zastoupení jsou v PDB databázi uloženy totožné proteiny a řetězce.

7.5.3 Odstranění homologie

Další postup vyžadoval zamyšlení, jak odstranit ze vzorku homologní řetězce (tedy s velmi vysokou podobností, ale již ne dokonale shodné). Vyšel jsem z předpokladu, že jelikož se zabývám tryptofany, tak je pro mé analýzy nejdůležitější jejich sekvenční okolí v řetězci a jeho zbylé části je možné ignorovat. Z tohoto důvodu mě napadlo vytvořit pro každý řetězec odvozenou sekvenci aminokyselin, která by byla složená výhradně z úseků blízko každého tryptofanu a tato odvozená sekvence by se následně použila při odstranění homologie. Nicméně tato sekvence nemá dostatek informací o prostorovém okolí tryptofanu, jelikož v prostoru blízko něj, se mohou vyskytovat aminokyseliny v primární sekvenci

poměrně vzdálené. Proto jsem před další analýzou pro každý řetězec vytvořil dvě odvozené sekvence.

První odvozená sekvence, kterou označuji jako „lineární“, obsahuje tryptofan a dvacet aminokyselin v jeho sekvenčním okolí (-10 až +10). Pro každý tryptofan takto vznikne sekvence dlouhá 21 aminokyselin. Pokud je v řetězci více tryptofanů, pak se takovéto sekvenční bloky spojí pomlčkou. Jako příklad uvádím sekvenci řetězce A T4 lysozymu 102L na obrázku 7.6.

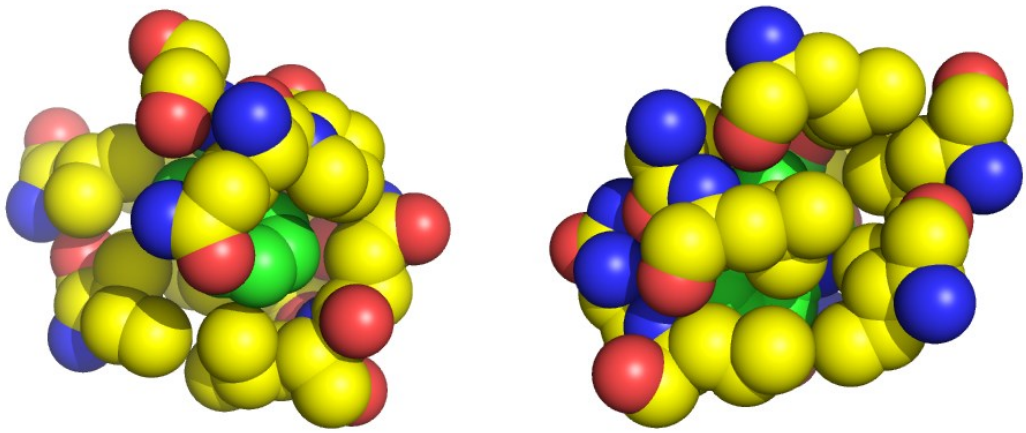


Obrázek 7.6: Řetězec A T4 lysozymu 102L a jeho převedení na okolní sekvenci složené z okolí dvaceti aminokyselin každého tryptofanu. Jednotlivé úseky jsou oddělené pomlčkou. Pokud se tryptofan nachází na okraji řetězce, tak jsou místo aminokyselin doplněny znaménka plus. Tryptofany jsou zvýrazněny tučně. Jednotlivé úseky jsou barevně zvýrazněny a mohou se překrývat.

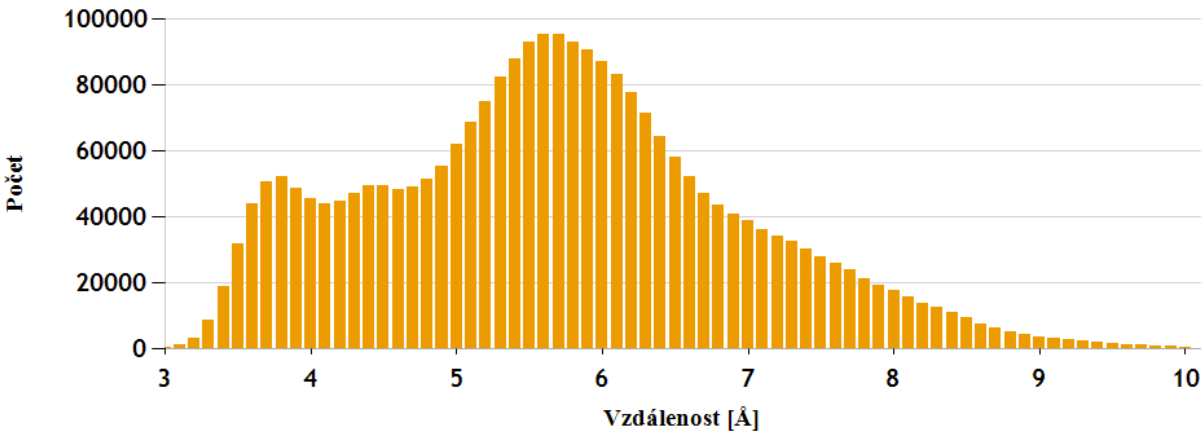
Druhá odvozená sekvence, kterou označuji jako „prostorovou“, je složená z aminokyselin v trojrozměrném prostoru okolo tryptofanu. Abych získal informaci o jeho prostoru, importoval jsem souřadnice prvních deseti nejbližších aminokyselin (dle vzdálenosti od těžiště tryptofanu definované v kapitole 7.4) a vzdálenost jsem omezil do 10 Å. Toto omezení je především z důvodu výskytu filamentárních proteinů, kdy by poslední aminokyseliny (devátá, desátá) byly zbytečně vzdálené (a tedy irelevantní při analýze interakcí). Nejbližší aminokyseliny jsem dohledal pouze na základě pozic atomů postranního řetězce, a to z důvodu zajištění specifity dané aminokyseliny. Kdybych použil i atomy v peptidové vazbě, tak by se vyskytovaly případy, že postranní řetězec nejbližší aminokyseliny je ve skutečnosti na opačné straně od indolu. Nejbližší by byla peptidová vazba, který by mohla patřit jakékoliv aminokyselině. Výjimku jsem udělal pouze u glycinu, u kterého jsem při importu využil právě pouze jeho Cα z peptidové vazby. Při dohledávání aminokyselin jsem navíc u postranních řetězců nezohledňoval atomy vodíků (nejsou k dispozici ve všech modelech). Během importu se vynořily konkrétní problémy, kterými jsou PDB soubory zatíženy. Z aktuálního vzorku například 1 365 řetězců obsahovalo v definici PDB souboru pro každou aminokyselinu jen její Cα souřadnici. Navíc 14 722 řetězců obsahovalo neúplné části v souřadnicích atomů v okolí tryptofanů nebo zcela chyběly

souřadnice samotného tryptofanu. Jednotlivá čísla udávám, aby si čtenář vytvořil obrázek o tom, v jak různém stavu mohou PDB soubory být, a že vyřešený protein v PDB souboru nemusí být úplný. Tyto všechny problematické modely jsem ze svého vzorku odstranil, takže mi zůstalo 47 542 řetězců.

Dále bylo potřeba rozhodnout kolik aminokyselin okolo tryptofanu by mohlo dostatečně určit specifitu jeho prostorového okolí. Tryptofan je v podstatě úplně obalený aminokyselinami, jak je vidět na příkladu na obrázku 7.7. Doufal jsem, že mi pomohou rozhodnout různá zobrazení importovaných aminokyselin v grafech. První graf, který se nabízel, byl histogram vzdáleností všech importovaných aminokyselin od těžiště tryptofanu zobrazených na obrázku 7.8.

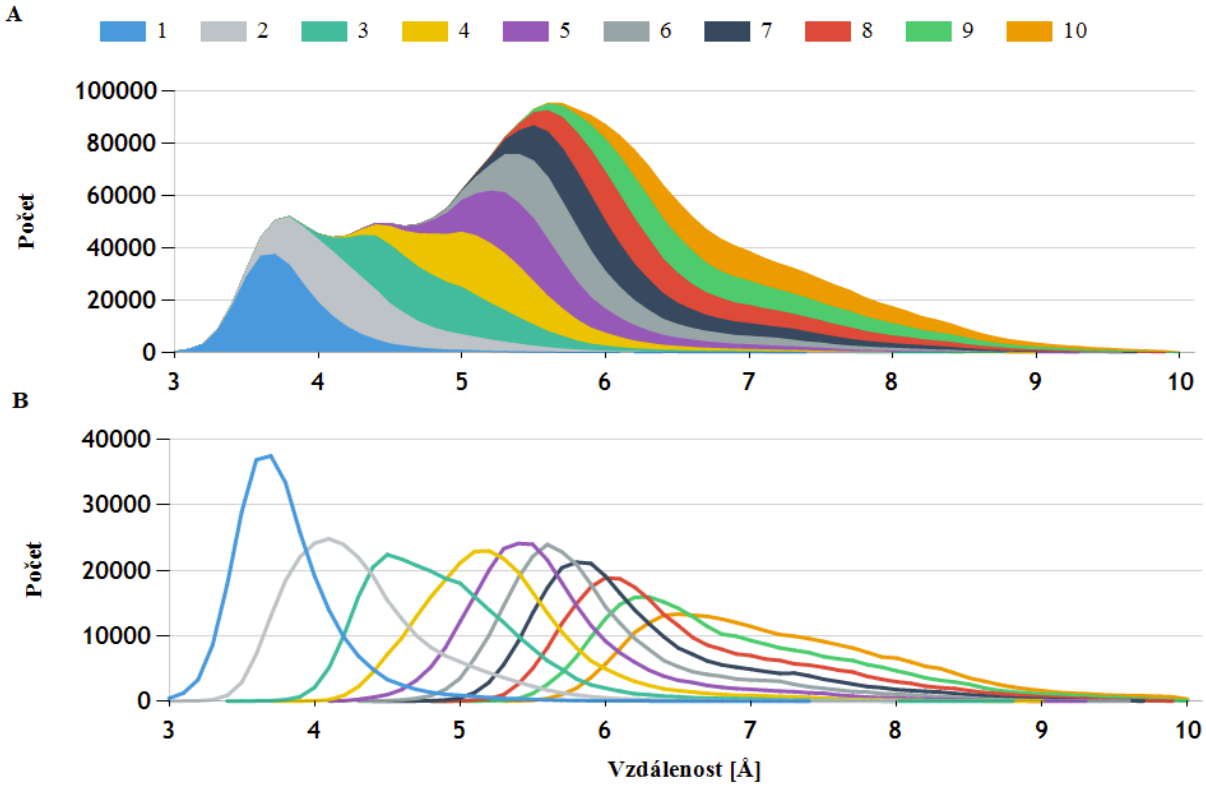


Obrázek 7.7: Ukázka, že tryptofan (zelený) interaguje s mnoha aminokyselinami. Zobrazeno deset nejbližších aminokyselin od těžiště indolové skupiny tryptofanu. (PDB: 4PH1)



Obrázek 7.8: Histogram vzdáleností postranních řetězců aminokyselin od těžiště tryptofanu (indolu). Pro výpočet bylo vždy použito pouze deset nejbližších aminokyselin ke každému tryptofanu. Celkově se jedná o 2 625 570 aminokyselin.

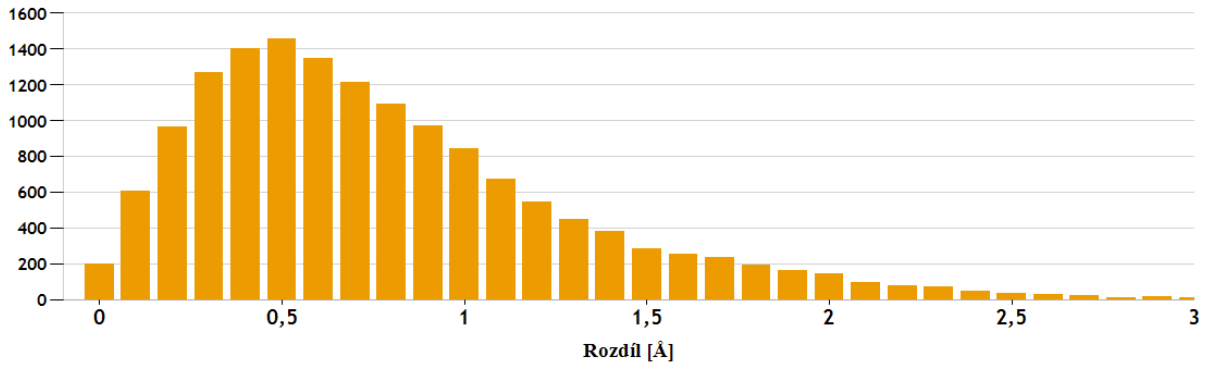
Pro lepší pochopení jsem na obrázku 7.9 graf rozdělil na deset částí, které reflektují jejich pořadí od tryptofanu.



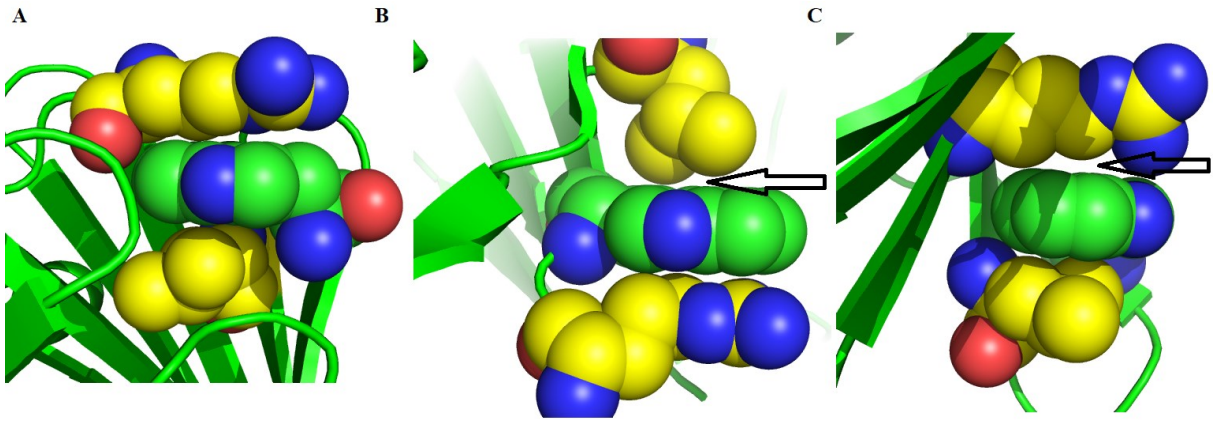
Obrázek 7.9: Kumulovaný (A) a nekumulovaný (B) histogram vzdáleností aminokyselin k tryptofanu zobrazující data z obrázku 7.7 rozpadlá podle pořadí vzdáleností jednotlivých aminokyselin. Plocha pod každou křivkou v grafu B je shodná (celkem jde vždy o desetinu z celkových 2 625 570 aminokyselin). Legenda: pořadí aminokyseliny (1 = nejbližší...).

Je zřejmé, že první pozice se velmi liší od ostatních a bude pravděpodobně vysoce specifická, protože je jednak nejblíže těžiště, a také zaujímá samostatně značnou část prostoru. Druhá pozice je určitě také důležitá, protože se pravděpodobně často vyskytuje u druhého povrchu indolové skupiny tryptofanu. Otázka je, proč se druhá pozice od první tak liší, když by z principu mohly být podobné (pokud by první dvě aminokyseliny interagovaly každá s jedním povrchem indolu). Křivka je nižší z důvodu většího rozmístění do prostoru, ale obecně není důvod pro posunutí doprava o přibližně 0,5 Å. Něco napovědět by mohl histogram rozdílů mezi první a druhou pozicí pro vzdálenost první pozice do 3,4 Å (oblast, kde se data nepřekrývají) na obrázku 7.10. Je vidět, že nejčastější rozdíl vzdáleností mezi blízkých prvních a druhých aminokyselin je okolo 0,5 Å, což prakticky odpovídá grafu 7.9 (A) pro první a druhé aminokyseliny. Nicméně při procházení PDB souborů s tímto rozdílem, jsem nenašel žádné zákonitosti, které by tento jev vysvětlovaly. Možné vysvětlení by byla například nějaká preference nejbližší aminokyseliny k tryptofanu, která by následně

při tvoření modelu zapříčinila oddálení od druhé aminokyseliny na opačné straně indolu. Bohužel ani to jsem nezaznamenal a uvádím příklad na dvojici Arg-Leu na obrázku 7.11, u které by se dala nějaká preference očekávat. Zdá se, že aminokyseliny nejsou často těsně u sebe a nabízí se otázka, jestli to nějak neovlivňuje prováděné analýzy. Musím se tedy zatím smířit s tím, že se pravděpodobně jedná o artefakt.



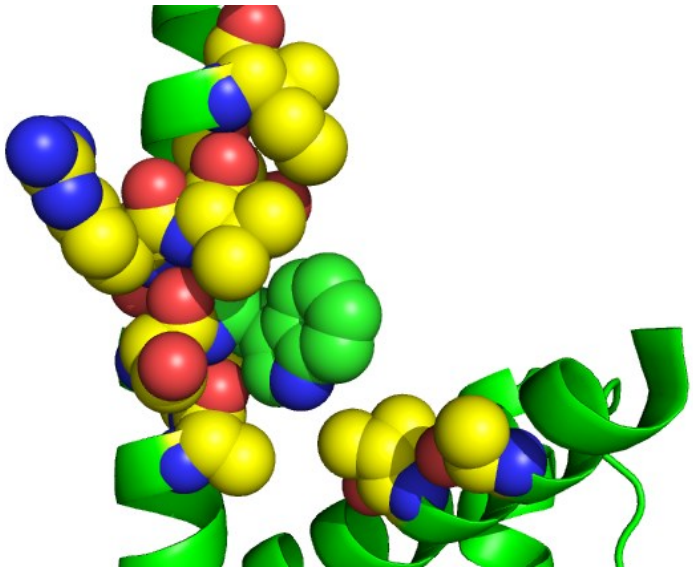
Obrázek 7.10: Histogram rozdílů vzdáleností mezi aminokyselinami na první a druhé pozici od těžiště indolu. Výběr je omezen na vzdálenost první pozice do 3,4 Å. Z definice lze očekávat průměrný rozdíl větší než nula, ale k maximu okolo 0,5 Å není důvod. Předpokládal jsem monotónně klesající závislost.



Obrázek 7.11: Obrázek ukazuje tryptofan a dvě nejbližší aminokyseliny, vždy arginin a leucin. A) Arginin i leucin jsou od těžiště tryptofanu stejně vzdáleny (PDB: 4EDE). B) Leucin je od těžiště tryptofanu o 0,5 Å dále než arginin (PDB: 1G0X). C) Arginin je od těžiště tryptofanu o 0,5 Å dále než leucin (PDB: 1II8).

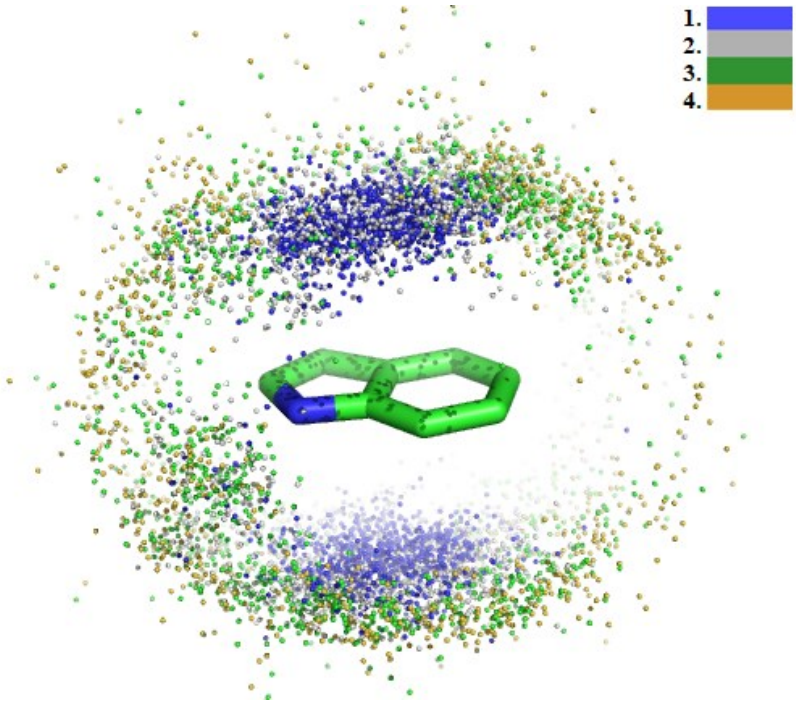
Třetí a čtvrtá pozice na obrázku 7.9 vypadají relativně podobně a od páté dochází k poklesu četnosti výskytu. Tento pokles je dán pouze čím dál větší „teoreticky možnou rozvolněností“ proteinových struktur dále od tryptofanu (tedy třeba na okraji proteinu). Zajímavé také je, že první pozice u jednoho proteinu může být dále, než například čtvrtá pozice u jiného. Tento úkaz je dán případy, kdy tryptofan vyčnívá do prostoru na povrchu

proteinu jako je na obrázku 7.12, popřípadě interaguje s jiným řetězcem, který ale nyní není brán v potaz.

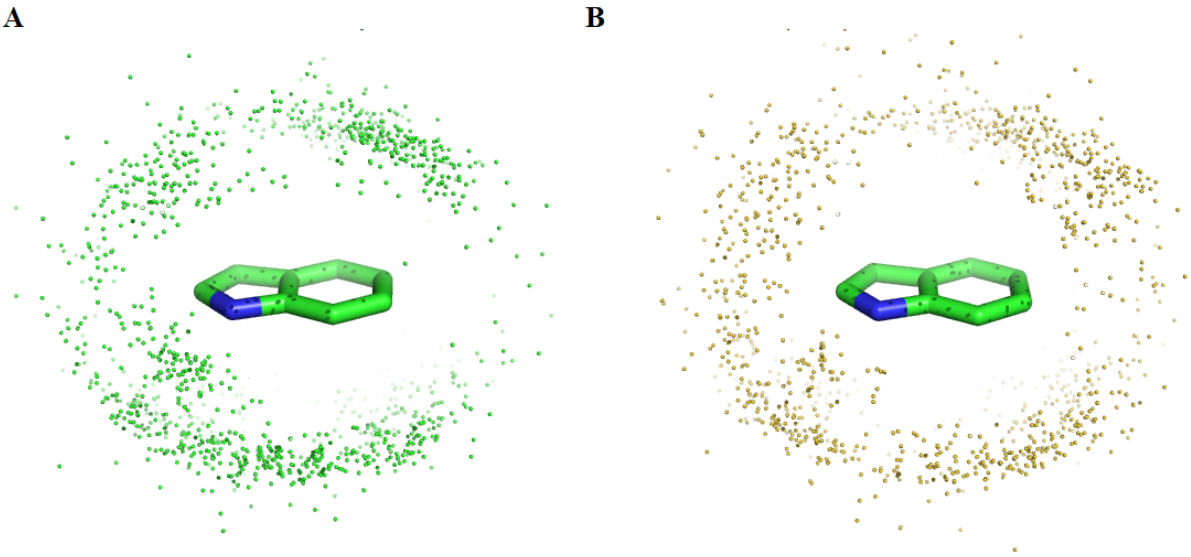


Obrázek 7.12: Ukázka tryptofanu, který vyčnívá do prostoru (zelený) a jeho 10 nejbližších aminokyselin (žlutě) (PDB: 1MJT).

Pro rozhodnutí kolik aminokyselin budu považovat jako dostatečně vypovídající o prostoru okolo tryptofanu, jsem vytvořil ještě jednu analýzu. Každý tryptofan (indol) jsem umístil stejně do počátku souřadnicového systému, jak je pospáno v kapitole 7.4. Díky tomu jsem mohl zobrazit atom nejbližších aminokyselin, podle kterého byly dohledány, tak aby odpovídaly relativní pozici vůči indolu. Nejprve jsem pro představu zobrazil náhodný výběr 10 000 aminokyselin na prvních čtyřech pozicích na obrázku 7.13. Není překvapivé, že první dvě pozice se vyskytují především nad a pod rovinou indolu. Na následujících obrázcích 7.14 (A) a 7.14 (B) jsou zobrazeny samostatně pozice třetí a čtvrté aminokyseliny v pořadí. Jak už vypovídal graf na obrázku 7.9, tyto pozice se značně překrývají, ale čtvrtá pozice je pochopitelně většinou dále. Nejdůležitější je ale informace, že již třetí pozice se vyskytuje okolo celého tryptofanu, a tím jí můžu brát jako doplňující k prvním dvěma pozicím. Čtvrtou můžu při hledání strukturních homologů vynechat, protože se od třetí liší pouze v tom, že je v průměru dále od těžiště.

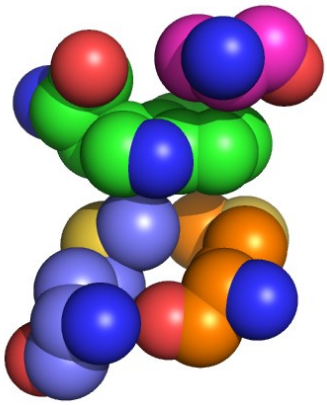


Obrázek 7.13: Indol tryptofanu a výskyt čtyř nejbližších aminokyselin v prostoru. Vždy je zobrazen pouze nejbližší atom z postranního řetězce okolních aminokyselin. Barva značí pořadí blízkých aminokyselin (viz legenda). Zobrazeno je náhodných 10 000 aminokyselin.



Obrázek 7.14: A) Indol tryptofanu a výskyt třetí nejbližší aminokyseliny v prostoru. B) Indol tryptofanu a výskyt čtvrté nejbližší aminokyseliny v prostoru. V obou případech je zobrazeno náhodných 2 500 aminokyselin.

A **B**



MNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAAKSE
LDKAIGRNTNGVITKDEAEKLFNQDVDAAVRGILRNAKLKPVYD
SLDAVRRAALIN**M**VFQ**M**GETGVAGFTNSLRMLQQKRWDEAA
VNLAKSR**W**YNQTPNR**A**KRVITTFRTGTWDAYKNL

Obrázek 7.15: Ukázka aminokyselin prostorově blízkých k tryptofanu. A) Světle zelený tryptofan na pozici 138 řetězce A T4 lysozymu (PDB: 102L). K němu zobrazené tři nejbližší aminokyseliny. První metionin (modrá), druhý alanin (růžová) a třetí metionin (oranžová). B) Sekvence řetězce s barevně zvýrazněnými příslušnými aminokyselinami.

Pro představu je tato trojice ukázána na příkladu T4 lysozymu na obrázku 7.15. Je nutné zdůraznit, že takto definované prostorové okolí není pro analýzy nijak stěžejní, jde pouze o pomocný nástroj využívaný při odstraňování homologie. Vzhledem k tomu, že nejbližší aminokyseliny se v tomto postupu nacházejí u těžiště tryptofanu, jsou tím pádem upřednostňovány basické aminokyseliny, které se mohou nacházet v oblasti záporného náboje nad a pod rovinou indolu. Interakce aminokyselin na okrajích indolu jsou z tohoto důvodu v tomto postupu častěji upozaděny.

Druhá odvozená sekvence, tedy „prostorová“, obsahuje z důvodů pospaných výše tryptofan a 3 nejbližší aminokyseliny z jeho okolního prostoru. Pro každý tryptofan takto vznikne sekvence dlouhá 4 aminokyseliny. Pokud je v řetězci více tryptofanů, pak se takovéto sekvenční bloky spojí pomlčkou. Během vytváření této sekvence se ale vyskytl problém, že u homologních proteinů může být z důvodu nepřesnosti vytvořeného modelu u nějakého tryptofanu prohozena v prostoru třetí a čtvrtá nejbližší aminokyselina (například WMA**MQ** a WMA**QM**). V případě, že bych se zaměřil pouze na první tři, tak by se toto prohození jevilo jako odlišný protein. Pokud se tedy takovéto tryptofany vyskytly, vybral jsem náhodně jednu z trojic (například WMA**M**) a přidělil jsem jí všem tryptofanům v dané skupině jako prostorovou sekvenci, tím je zajištěno, že se s nimi bude zacházet jako se stejnými. Dále mohou být ze stejného důvodu prohozené i aminokyseliny třeba na první a druhé pozici. To jsem vyřešil postupem, že jsem seřadil aminokyseliny v těchto trojicích abecedně.

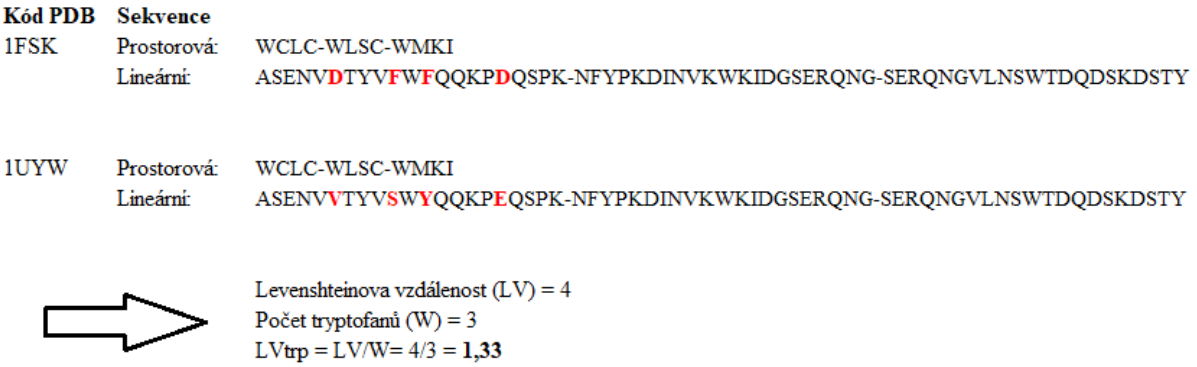
Nyní jsem měl tedy k dispozici dvě odvozené sekvence od každého řetězce, které by měly obsahovat pro mě důležité informace. Jedná se o jednu lineární a jednu prostorovou

sekvenci. Tyto dvě sekvence nyní použiji pro identifikaci a následné odstraňování homologie řetězců v aktuálním vzorku. Jelikož považuji prostorovou sekvenci za specifickou vůči tryptofanu a jakékoliv změny v ní mohou být důležité, rozdělil jsem aktuální vzorek 47 542 řetězců do 5 037 skupin právě na základě prostorové sekvence. Ukázka nejpočetnějších skupin je na obrázku 7.16.

Prostorová sekvence	**Počet**
WCLL-WCLS	181
WKLR-WAMM-WAMV	157
WGKL-WGLV	92
WIIM-WKRY-WAVY-WALV-WCIN	80
WKLR-WAMQ-WMTV	69
WCLV-WCLS	64
WKSW-WFQW	62
WFKK-WLPY	53
WCLV-WCLS-WLPP	49
WCLL-WCLS-WMNS	49
WIIM-WKRY-WAVY-WELV-WCIN	47
WALS-WKPT-WCLS-WLLP	46
WESV-WANY	46
WPVW-WPPP-WILR-WFFL-WFRT	42

Obrázek 7.16: Výskyt nejpočetnějších prostorových sekvencí nalezených ve 47 542 řetězcích.

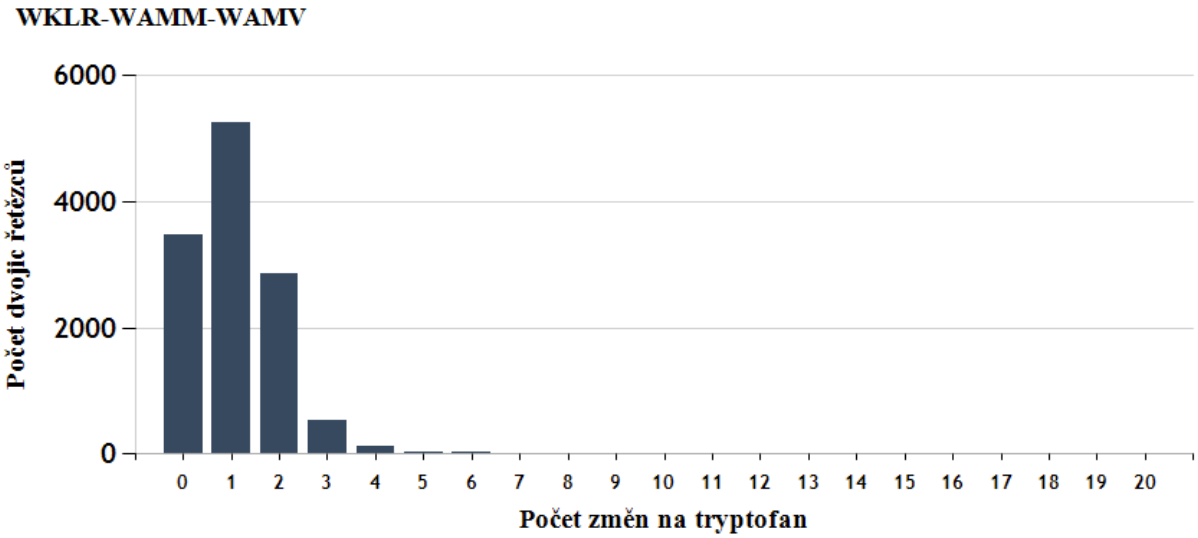
Homologii řetězců jsem dále řešil pouze uvnitř těchto samostatných skupin. Pro zjištění podobnosti řetězců v rámci skupin jsem pro každou dvojici řetězců v dané skupině vypočítal Levenshteinovu vzdálenost (LV) mezi jejich lineárními sekvencemi, což vždy dalo celé číslo odpovídající počtu záměn, insercí a delecí aminokyselin (viz kapitola 7.3). Pro lepší výpovědní hodnotu jsem každý výsledek normalizoval na základě počtu tryptofanů, který je charakteristický pro danou skupinu (řetězce ve skupině mají stejný počet tryptofanů). Ukázkový postup je znázorněný na obrázku 7.17. Tyto normalizované Levenshteinovy vzdálenosti (vzniklé z dvojic řetězců) jsem následně zobrazil do grafů, a to způsobem, že jsem hodnoty zaokrouhlil vždy na celá čísla nahoru, takže například hodnota 1,33 byla zobrazena jako 2. Toto zaokrouhlené číslo dále označuji jako počet změn na tryptofan (LVtrp). Vzhledem k tomu, že každý tryptofan má v sekvenci okolo sebe deset aminokyselin na každou stranu, může být maximální normalizovaná Levenshteinova vzdálenost (LVtrp) 20.



Obrázek 7.17: Výpočet normalizované Levenshteinovy vzdálenosti dvou řetězců se stejnou prostorovou sekvencí a různou sekvencí lineární. Červeně jsou zvýrazněny odlišnosti v lineární sekvenci.

Na obrázku 7.18 je zobrazen výsledný graf pro druhou nejpočetnější skupinu WKLR-WAMM-WAMV. Skupina má 157 řetězců, což při porovnávání každého řetězce s každým dá 12 246 dvojic ($(157^2 – 157) / 2$). Na grafu je patrné, že mezi dvojicemi řetězců vycházejí pouze nízké LVtrp, což vypovídá o značné homologii skupiny. Neexistují totiž dvě lineární sekvence v rámci skupiny, které by se od sebe lišily o více než šest změn na tryptofan, což při třech tryptofanech v řetězci odpovídá osmnácti změnám na lineární sekvenci. Jiný příklad uvedu na obrázku 7.19, kde je zobrazena skupina WCCL-WCLS, která obsahuje 40 řetězců. Z profilu grafu lze vyčíst, že z větší části se ve skupině objevují podobné lineární sekvence. Ale také je zde skupina řetězců, která se úplně liší. Z toho se dá usuzovat, že prostorová sekvence WCCL-WCLS je sdílená u více nehomologních proteinů.

Čím má skupina více tryptofanů, tím musí být skupina podobnější, a to z důvodu, že čím je více trojic v prostorové sekvenci, tím je méně pravděpodobné, že se budou v rámci různých řetězců shodovat. Například čtvrtá nejčastější skupina WIIM-WKRY-WAVY-WALV-WCIN složená z 80 lysozymů zobrazená na obrázku 7.20, se skládá v podstatě z totožných lineárních sekvencí, kdy je maximálně jedna změna na tryptofan.

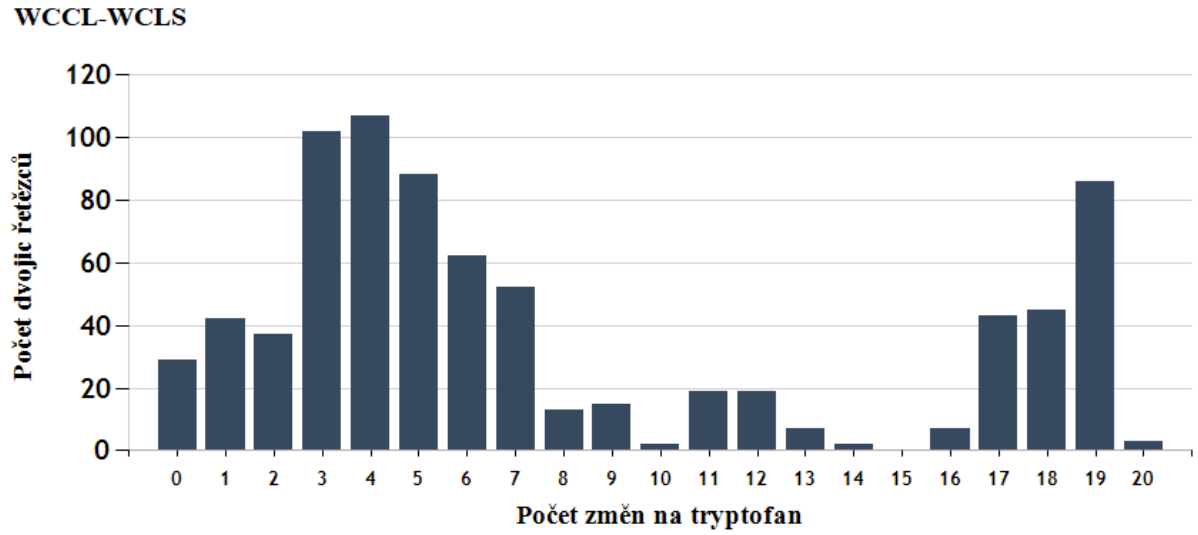


Kód PDB	Lineární sekvence
158L	NALAMLQQKRWDEAAVNLAKS-EAAVNLAKSRWYNQTPNRAKR-RVITTFRTGTWDAYKNL++++
195L	NSLRMLQQKRWDELAVNLAKS-ELAVNLAKSRWYNQTPNRAKR-RVITTFRTGTWDAYKNL++++

LV = 4 / 3 = 1,33

LVtrp = 2

Obrázek 7.18: Zobrazeny počty normalizovaných Levenshteinových vzdáleností pro skupinu WKLR-WAMM-WAMV o 181 řetězcích. Skupina má tři tryptofany, takže 0 odpovídá dvojicím řetězců, které mají stejnou lineární sekvenci. Druhý sloupec je u dvojic, které mají v lineární sekvenci jednu, dvě nebo tři změny (díky zaokrouhlení). Třetí sloupec odpovídá čtyř, pěti, nebo šesti změnám atd. Pod grafem ukázka dvou lineárních sekvencí s hodnotou LVtrp 2.

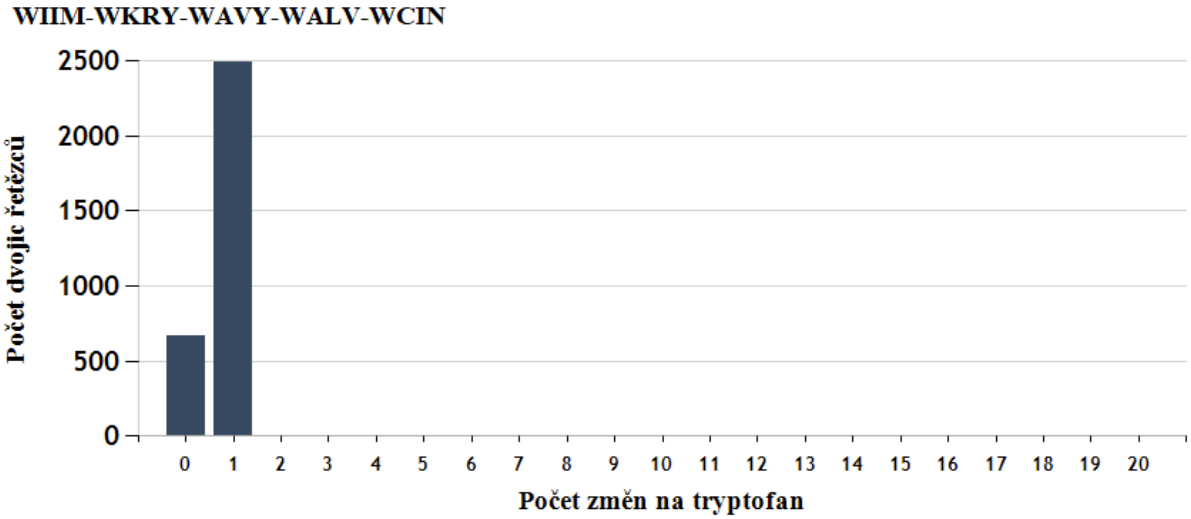


Kód PDB	Molekula	Lineární sekvence
5JW4	MEDI8852 LIGHT CHAIN	TSQSLSSYTHWYQQKPGKAPK-NFYPREAKVQWKVDNALQSGN
3MJG	BETA-TYPE PLATELET-DERIVED GROWTH FACTOR RECEPTOR	LTCSGSAPVVWERMSQEPPQE-VIGNEVVNFEWTYPRKESGRL

LV = 37 / 2 = 18,50

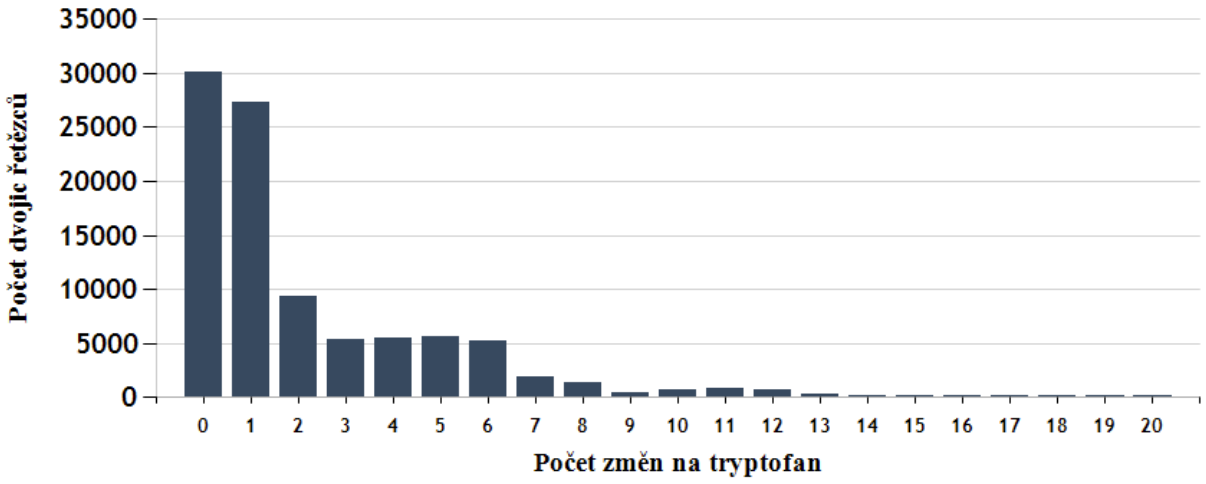
LVtrp = 19

Obrázek 7.19: Zobrazeny počty normalizovaných Levenshteinových vzdáleností pro skupinu WCCL-WCLS o 40 řetězcích. Pod grafem je ukázka dvou naprosto rozdílných řetězců.

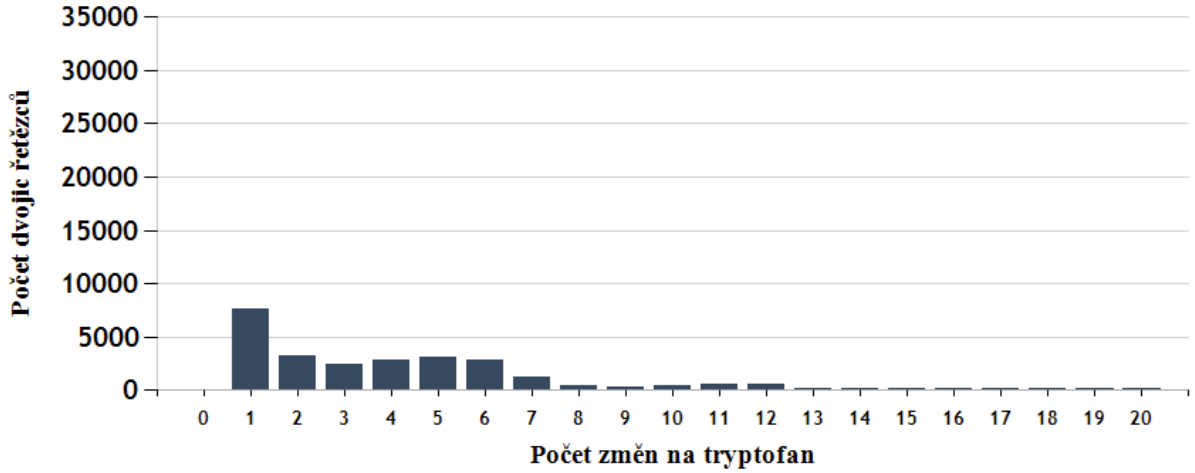


Obrázek 7.20: Zobrazeny počty normalizovaných Levenshteinových vzdáleností pro skupinu WIIM-WKRY-WAVY-WALV-WCIN o 80 řetězcích. Řetězce obsahují vždy pět tryptofanů a je vidět, že se téměř neliší.

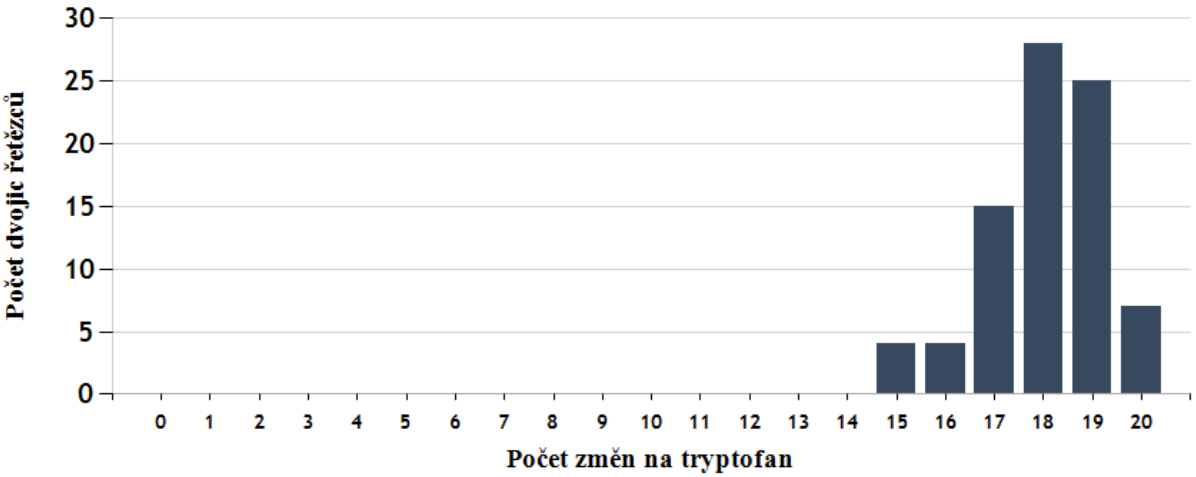
Pro snazší práci jsem výsledky všech skupin pro jednotlivé sloupce (LVtrp) sečetl a zobrazil v jednom grafu na obrázku 7.21. Profil dat vypovídá o obrovském výskytu stejných nebo podobných lineárních sekvencí. Nyní jsem tedy mohl začít jednotlivé řetězce odstraňovat, a to tak, že jsem postupoval po jednotlivých sloupcích zleva. Nejprve jsem jednoduše vzal každou dvojici řetězců v prvním sloupci (dvojice se stejnou lineární sekvencí) a odstranil ten řetězec s horším rozlišením. Takovým odstraněním jsem zároveň odstranil i všechny další dvojice řetězců ve všech dalších sloupcích vpravo, kde se vyskytoval daný řetězec. Tento proces byl opakován celkem 30 000krát (viz počet v prvním sloupci na grafu 8.21). Výsledný graf je na obrázku 7.22. Na grafu je patrné, že první sloupec je nulový z důvodu neexistence stejných dvojic. U ostatních sloupců došlo ke značnému celkovému poklesu. Jen tímto krokem bylo odstraněno 8 004 řetězců obsažených v 69 198 dvojicích řetězců. Otázka nyní byla, jak velké množství změn mám při takovémto odstraňování homologie tolerovat. Postupně jsem tedy stejným způsobem odstraňoval jednotlivé sloupce zleva. Tímto jsem došel až ke sloupci obsahující patnáct změn na tryptofan. Od tohoto sloupce jsem již pozoroval rozdílné proteiny, takže jsem se rozhodl pro tuto hranici rozdílnosti a dál už jsem dvojice neodstraňoval. Finální graf je na obrázku 7.23. Díky tomuto postupu jsem z celkového počtu 47 542 řetězců odstranil dalších 12 741 na výsledných 34 801. Pro tuto chvíli ve vzorku řetězců stále přetrvává určitá homologie, ale ta bude odstraněna v rámci konkrétních analýz.



Obrázek 7.21: Zobrazeno suma hodnot ze všech skupin, což dává celkových 95 522 dvojic řetězců.



Obrázek 7.22: Zobrazen graf z obrázku 7.21 po odstranění prvního sloupce, tedy stejných řetězců. Vysvětlení viz text.

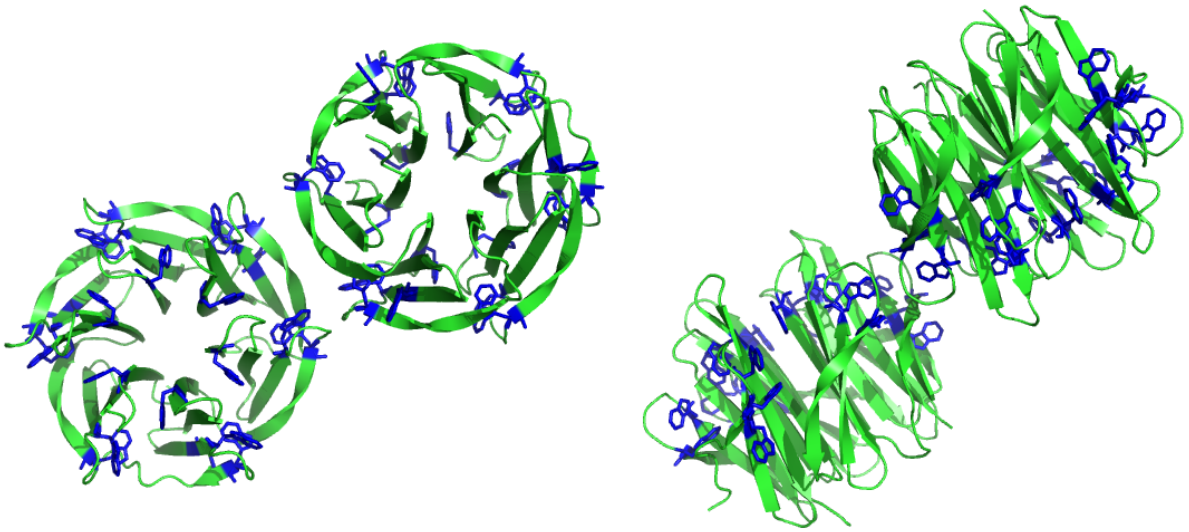


Obrázek 7.23: Profil podobnosti řetězců ve finálním datasetu, který obsahuje 34 801 řetězců.

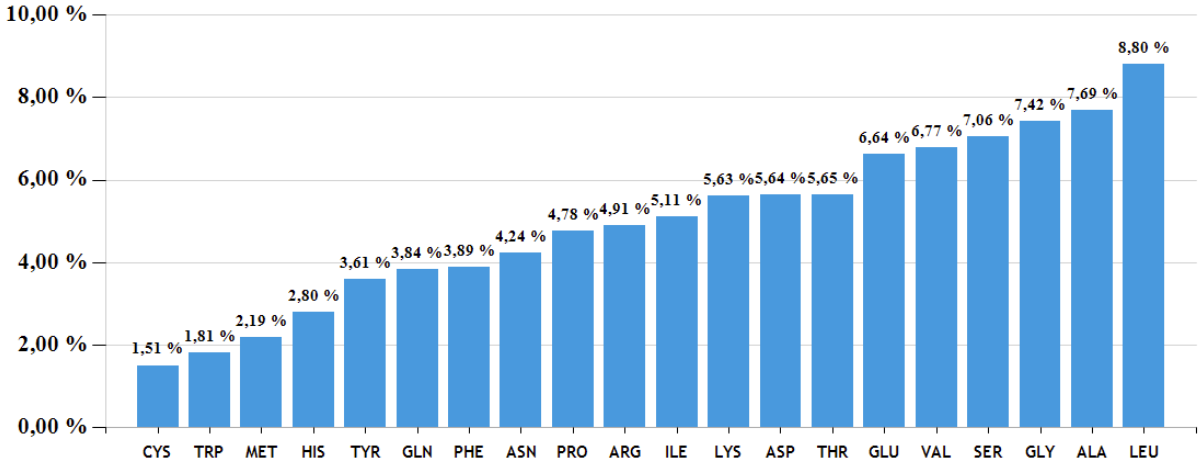
7.5.4 Velký počet tryptofanů

Během práce s řetězci jsem narazil na problém, že některé obsahují relativně vysoký počet tryptofanů jako je například Fucose-binding lectin protein (PDB: 4CSD) na obrázku 7.24. U takových proteinů může docházet k pochybnostem, jak moc jsou jednotlivé tryptofany specifické. Konkrétně jejich vzájemné interakce a okolí. Pokud se v proteinu nacházejí pouze dva tryptofany a jsou zrovna blízko sebe, lze se domnívat, že spolu interagují a jsou důležité. Naproti tomu pokud jsou v proteinu u sebe dva tryptofany, ale zároveň ve zbytku proteinu se nachází mnoho dalších, mohlo by to podle mě svědčit spíše o tom, že tyto tryptofany jsou důležité pro celkovou funkčnost proteinu, než že by byly zajímavé jednotlivě. Pro představu jsem proto sestrojil graf na obrázku 7.25, který zobrazuje průměrný procentuální výskyt

jednotlivých aminokyselin v aktuálním vzorku řetězců. Je vidět, že tryptofan má průměrný výskyt 1,81 %, naproti tomu ve Fucose-binding lectin proteinu je to 7,72 %, což je více, než obsah druhé nejpočetnější aminokyseliny alaninu. Rozhodl jsem se tedy, že se ve své analýze nebudu zabývat řetězci, které mají více než trojnásobný výskyt tryptofanů oproti výskytu průměrnému. Jedná se o 215 řetězců.



Obrázek 7.24: Fucose-binding lectin protein (PDB: 4CSD) a jeho modře zobrazené tryptofany.



Obrázek 7.25: Průměrný procentuální výskyt jednotlivých aminokyselin v 34 811 řetězcích. Řetězce mají vždy minimálně dva tryptofany.

Finální dataset pro tuto práci je tedy 34 586 řetězců z 32 823 proteinů. Pro všechny analýzy v dalších kapitolách je jako výchozí seznam řetězců použit tento vzorek. Cílem bylo získat množinu řetězců, která by byla reprezentativní a zároveň co nejvíce robustní. Nicméně stále v sobě obsahují určitou úroveň homologie (například pokud se dva homologní řetězce

liší počtem tryptofanů). Tato homologie je následně řešena v konkrétních analýzách. Postup odstraňování homologie v této kapitole je relativně složitý a značně náročný na vysvětlení, proto na obrázku 7.26 předkládám seznam kroků, které zde byly popsány. Tento princip odstraňování homologie pomocí prostorové a lineární sekvence je využíván v následujících kapitolách.

Získání reprezentativního vzorku řetězců

Krok		**Počet zbývajících řetězců**
1.	Získání řetězců z PDB databáze	393 041
2.	Odstranění všech řetězců s méně než dvěma tryptofany	241 923
3.	Odstranění stejných řetězců	63 629
4.	Vytvoření odvozené lineární sekvence ke každému řetězci	
5.	Vytvoření odvozené prostorové sekvence ke každému řetězci	
6.	Rozdělení řetězců na základě prostorové sekvence do skupin	
7.	Odstranění homologie na základě lineární sekvence v rámci vytvořených skupin	34 801
8.	Odstranění řetezců s vysokým počtem tryptofanů	**34 586**

Obrázek 7.26: Seznam kroků popisující získání reprezentativního vzorku 34 586 řetězců.

7.6 Analýzu prostorového okolí

7.6.1 Úvod

První a zdánlivě nejjednodušší analýza spočívá ve zjištění, jaké aminokyseliny se vyskytují okolo jednotlivých tryptofanů. Ze vzorku řetězců z kapitoly 7.4 jsem vybral všechny jejich tryptofany, kterých bylo 183 750. Vzhledem k tomu, že postup získání vzorku řetězců umožnil zachování určité homologie, je potřeba se s ní dodatečně vypořádat na úrovni jednotlivých tryptofanů. Tato homologie je ukázána na obrázku 7.27 v případě dvou lysozymů T4, kde jeden z nich (PDB:146L) má o jeden tryptofan navíc. V analýzách by následně byl problém, že informace ze zbylých tří tryptofanů, které jsou homologní, by byla nadhodnocena.

Kód PDB	**Prostorová sekvence**
146L	WKLR-WAMQ-WALM-WAMV
152L	WKLR-WAMQ-WMVY

Lineární sekvence

NSLRMMQQKRWDELAVNMAKS-ELAVNMAKSRWYNQTPNRAKR-PNRAKRIITTWRTGTWDAYKN-RIITTWRTGTWDAYKNL++++
NSLRMLQQKRWDEAAVNLAKS-EAAVNLAKSRWYNQCPNRAKR-RVITTFRTGTWDAYKNC++++

Obrázek 7.27: Ukázka zachovalé homologie u dvou lysozymů T4. PDB: 146L má o jeden tryptofan navíc. Zbylé úseky jsou velmi podobné.

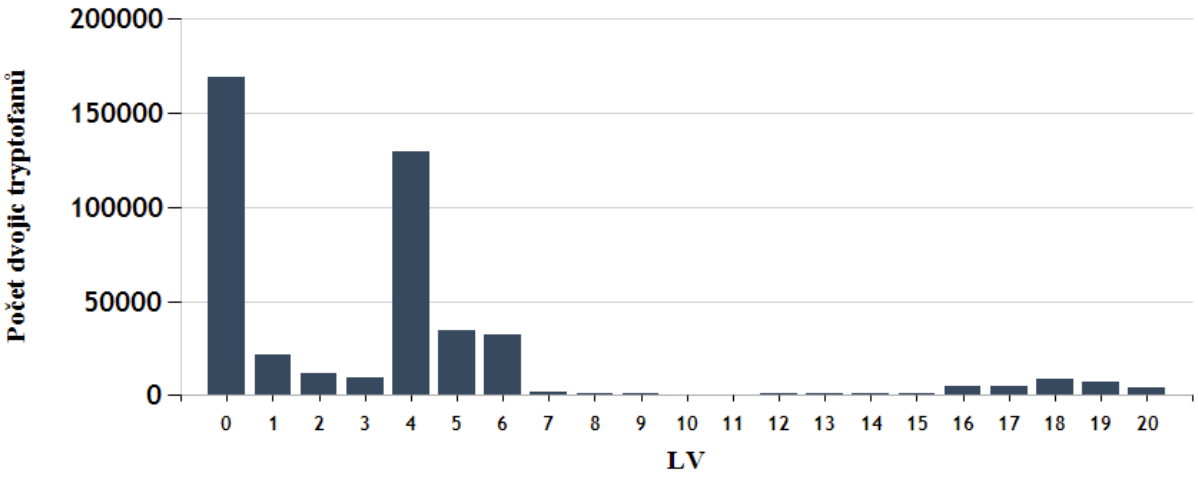
7.6.2 Získání datasetu tryptofanů

Pro odstraňování homologie jsem využil stejný princip jako při odstraňování homologie řetězců. Ke každému tryptofanu byla přiřazena jeho odpovídající část prostorové sekvence (abecedně seřazené první tři nejbližší aminokyseliny v prostoru) a lineární sekvence (deset aminokyselin na každou stranu od tryptofanu v sekvenci řetězce), které byly definovány v kapitole 7.5.3. Pro analýzu prostoru je stěžejní prostorová sekvence a jakékoliv rozdíly v ní mohou být důležité, proto jsem tryptofany podle ní rozdělil do 1 546 skupin a homologii řešil jen na základě lineární sekvence v rámci těchto skupin. Nejpočetnější skupiny jsou zobrazeny na obrázku 7.28. Již z tohoto obrázku by se mohlo mylně vyvozovat, že cystein se u tryptofanu vyskytuje preferenčně, protože se vyskytuje hned v prvních třech nejpočetnějších skupinách. Po kontrole jsem ale zjistil, že se jedná v drtivé většině protilátky (s vysokým obsahem cysteinu), a tudíž značně podobné proteiny. Tento příklad ukazuje, jak je odstraňování homologie důležité, a jak moc může ovlivnit analýzu.

Prostorová sekvence	Počet
WCSV	941
WCLL	913
WCLS	797
WILP	773
WLPV	710
WCCL	681
WELR	674
WFLL	671

Obrázek 7.28: Četnosti nejpočetnějších prostorových sekvencí vyskytující se v okolí 183 750 tryptofanů.

Stejně jako v kapitole 7.5.3 jsem vypočítal mezi jednotlivými lineárními sekvencemi tryptofanů v daných skupinách Levenshteinovu vzdálenost (nyní nebylo potřeba normalizovat, jelikož byl vždy jeden tryptofan) a tyto dvojice jsou pro skupinu WCSV zobrazeny v grafu na obrázku 7.29.



Obrázek 7.29: Zobrazeny četnosti Levenshteinových vzdáleností pro skupinu WCSV o počtu 941 tryptofanů.

Na rozdíl od porovnávání řetězců, kde jejich dlouhé prostorové sekvence neumožňovaly příliš velkou podobnost nehomologních proteinů, tak v případě tryptofanů je situace jiná. Spočívá v tom, že prostorová sekvence jednoho tryptofanu je složená z pouhých tří nejbližších aminokyselin, takže se může snadno vyskytovat u nepříbuzných proteinů. Z tohoto důvodu bych se nyní těmto grafům rád věnoval podrobněji. Rozložení dat na grafu 7.29 mě zaujalo a pokusím se vysvětlit, proč vypadá zrovna takto. Ve skupině WCSV je 941 tryptofanů, což dá dohromady 442 270 porovnávaných dvojic ($(941^2 - 941) / 2$), které jsou zobrazeny v grafu**.** Hlavní zvláštností je největší počet dat ve sloupcích 0 a 4. To by mělo svědčit o obrovské homologii v této skupině WCSV a zřetelně se to dá vypozorovat

po seskupení dle lineární sekvence na obrázku 7.30. Tryptofanů s lineární sekvencí DYFPEPVTVSWNSGALTSGVH je 533 z 941 Je zřejmé, že díky kombinacím těchto 533 tryptofanů je obrovské množství dvojic ve sloupci 0 (bez rozdílu). Sloupec 4 je početný především proto, že mezi první (533) a druhou (229) nejpočetnější lineární sekvencí jsou přesně čtyři změny**.** Je vidět, že dokud nebudou odstraněny stejné lineární sekvence, je jakákoliv analýza zbytečná.

Lineární sekvence	**Počet tryptofanů**
DYFPEPVTVSWNSGALTSGVH	533
GYFPEPVTVTWNSGSLSSGVH	229
GYFPESVTVTWNSGSLSSSVH	38
GYFPEPVTLTWNSGSLSSGVH	30
GYLPEPVTVTWNSGTLTNGVR	11
GYFPEPVTVKWNYGALSSGVR	9
DYFPEPVTVSWNSGSLTSGVH	8
GYFPEPVTVTWNSGALSSGVH	6
DYFPQPVTVSWNSGALTSGVH	5

Obrázek 7.30: Zobrazeny nejčastější lineární sekvence ve skupině WCSV.

První část homologních tryptofanů jsem identifikoval na základě stejné prostorové i lineární sekvence, kterých byla necelá třetina původního počtu (52 142 tryptofanů). Při rozhodnutí, který tryptofan ve dvojici ponechám, jsem opět postupoval podle lepšího rozlišení řešeného proteinu. Zbylé tryptofany (131 608) jsem seskupil podle prostorové sekvence a výsledek je na obrázku 7.31.

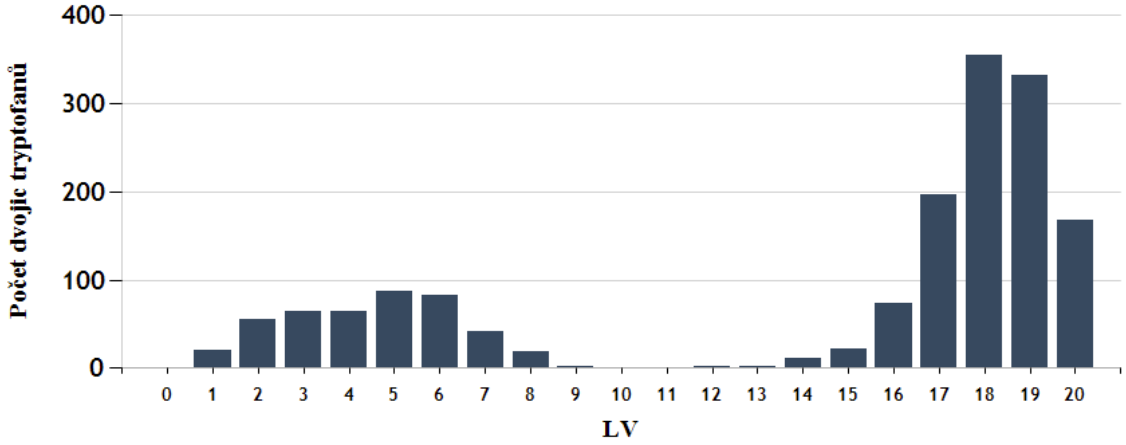
Prostorová sekvence	**Počet**
WCLL	688
WFLV	493
WFIL	486
WILV	482
WELR	479
WFLL	473
WLLV	467
WLPV	466

Obrázek 7.31: Počty nejpočetnějších prostorových sekvencí vyskytující se v okolí 131 608 tryptofanů.

Je vidět, že horní příčky zaujímají jiné trojice aminokyselin než na obrázku 7.28. Dříve nejpočetnější skupina WCSV (nyní nezobrazena) má nyní jen 57 tryptofanů z původních 941. Opět jsem pro tuto skupinu zobrazil graf Levenshteinových vzdáleností na obrázku 7.32. Kromě obrovského poklesu množství dat je patrné, že velká část dvojic si již

není podobná, protože se vyskytuje v pravé části grafu**.** Ovšem levá část grafu udává, že se ve skupině stále nějaká homologie vyskytuje.

Nyní bych chtěl vysvětlit, proč graf 7.32 od hodnoty 1 roste, následně dosáhne vrcholu v hodnotě 5, a dále pak klesá. Dalo by se to mylně pokládat za nějaký trend, ale opět to má na svědomí homologie. Do dvanácti rozdílů se jedná o páry, které jsou tvořené z 99 % homologními tryptofany z protilátek, a to vytvoří levou stranu grafu. Pravá strana je dána nepříbuznými tryptofany, kdy maximální počet dvacet rozdílů je snížený pravděpodobně proto, že i u naprosto odlišných tryptofanů je šance, že se objeví na stejných místech od tryptofanu stejné aminokyseliny.



Obrázek 7.32: Zobrazeny počty Levenshteinových vzdáleností pro skupinu WCSV po odstranění stejných tryptofanů. 57 tryptofanů.

Dále jsem pokračoval stejně jako u řetězců. Postupně jsem odstraňoval zleva horší tryptofany z jednotlivých dvojic a pozoroval rozdíly. Bohužel dvojic bylo tolik, že na rozdíl od řetězců jsem nemohl rozhodnout na základě pozorování rozdílných lineárních sekvencí, zda se ještě jedná o příbuzné proteiny. Stanovil jsem tedy maximální hranici 25 % podobnosti (měřeno pomocí LV), což odpovídá maximálně pěti stejným aminokyselinám na lineární sekvenci (která má dvacet aminokyselin). Ve výsledku mi zůstalo 92 290 tryptofanů.

Odstraňování homologie se zachováním všech změn v tryptofanech v rámci prostorové sekvence může mít jeden problém. Pokud například jsou dva tryptofany získané z homologních proteinů, které se liší v prostorové sekvenci na základě jedné aminokyseliny, pak budou ponechány oba. To ale znamená, že zbylé dvě aminokyseliny budou v analýze nadhodnocené, protože se jedná o homologní proteiny. Tento problém, jsem ale nedokázal vyřešit tak, abych zároveň nepřišel o danou mutaci. Nicméně předpokládám, že tím analýza nebude zatížena, protože by to měla vykompenzovat značná velikost studovaného vzorku tryptofanů.

7.7 Analýza párů tryptofanů

7.7.1 Úvod

Existuje mnoho článků, které se zabývají vzájemnou orientací arenů. Z důvodu rozsahu problematiky bych se rád zaměřil pouze na jednu konkrétní aminokyselinu, a to tryptofan. Tryptofan jsem zvolil z důvodů výsledků v kapitole 8.1.3, a také protože se již tryptofanem zabývám, tak je logické k němu zvolit jako interakčního partnera do páru další tryptofan.

7.7.2 Získání dvojic tryptofanů

U dvojic tryptofanů mě zajímala především jejich orientace a vzájemná pozice, proto jsem homologii odstraňoval jiným způsobem. Na začátku jsem vzal všechny řetězce, které byly vytvořeny v kapitole 7.5 a z každého řetězce jsem vytvořil páry tryptofanů. To znamená, že pokud měl řetězec tři tryptofany, tak vzniklo šest dvojic (1-2,1-3,2-3,2-1,3-1,3-2). Zdánlivá redundance (opakování stejného páru) je ponechána schválně, abych získal všechny možné orientace. Pouze je potřeba s tím v analýze počítat. Takto jsem získal 797 084 párů.

Opět je potřeba se vypořádat se zbývající homologií v řetězcích. V této části již ale nebudu do detailu rozebírat postup a zobrazovat grafy, jelikož doufám, že ho čtenář pochopil v předešlých kapitolách.

Prostorové sekvence mě v této analýze nezajímaly, nicméně porovnávat všechny lineární sekvence párů navzájem by nebylo časově možné. Zkusil jsem tedy porovnávání lineárních sekvencí v několika různých množinách. Největší pravděpodobnost homologie je u stejných názvů molekul (například Lysozyme), proto jsem všechny páry tryptofanů rozdělil do skupin podle těchto názvů. Z párů jsem v rámci skupin opět vytvořil dvojice, mezi kterými jsem vypočítal LVtrp (viz kapitola 7.5.3). Pokud vyšla LVtrp do 20, tak dvojici vyhodnotil jako homologní.

U dvojic jsem ale nemohl porovnávat lineární sekvence jako celek, protože by se například mohlo stát, že první tryptofan je stejný a druhý jiný. V tom případě by LVtrp vyšel do 20, ale o homologii by se určitě nejednalo. Porovnával jsem tedy jednotlivé tryptofany v dvojici zvlášť, vždy první s prvním a druhý s druhým. A pokud měly oba LV do 10, tak jsem je také bral jako homologní. Problém jsem zobrazil na obrázku 7.33.

Kód PDB	Lineární sekvence
6D0O	PEIAGYGTDEWTDWSWKSLRI-EIDTTRAVASWQKEVAENLAK
6D1O	PRIAGATSDEWTDWVWYTTEV-EMDTTRPVAAWMKEVEANIAR
	LV = 18
4LXC	DEVMKQDGHVWVGYTGNSGQR-DYVKAGQIIGWSGSTGYSTAP
5LEO	DEVMKQDGHVWVGYTGNSGQR-GQRIYLPVRTWNKSTNTLGVL
	LV = 18

Obrázek 7.33: Zobrazení, proč je nutné porovnávat tryptofany v párech zvlášť. I přes stejnou hodnotu LV je první pár vyhodnocen jako homologní a druhý ne.

Po odstranění homologů přes skupiny podle názvů molekuly, jsem stejný postup uplatnil podle prostorové sekvence a podle stejné sekvenční vzdálenosti mezi tryptofany ve dvojici (je zjevně velká pravděpodobnost, že jsou dva proteiny homologní, pokud mají vzdálenost mezi dvěma tryptofany shodnou). Takto zůstalo 585 443 párů tryptofanů. Nakonec jsem je ale stejně musel porovnat navzájem. Bohužel při čísle 585 443 by to stále nebylo výpočetně možné, takže jsem zkusil vyloučit ty, které s největší pravděpodobností homology nejsou. To znamená, že jsem je porovnával všechny navzájem, ale pokud se zkoumaná dvojice lišila v sekvenci mezi tryptofany o víc než padesát aminokyselin, tak už jsem je rovnou považoval za nehomologní. Stejně jsem postupoval, pokud se lišily ve vzdálenosti mezi těžišti o víc než 5 Å. Pokud je vzdálenost větší, tak i když se může jednat o určitou homologii, vzdálenost je natolik rozdílná, že považuji za vhodné oba páry v dvojici v analýze ponechat. Tímto se mi podařilo odstranit dalších 26 471 na výsledných 558 962 párů.

8 Výsledky a diskuze

8.1 Prostorové okolí tryptofanů

8.1.1 Úvod

První a zdánlivě nejjednodušší analýza zkoumá, které aminokyseliny se vyskytují okolo jednotlivých tryptofanů. V kapitole 7.6.2 bylo získáno 92 290 unikátních tryptofanů, se kterými bude nyní pracováno. Před samotnou analýzou bych rád objasnil jeden z cílů práce, jestli sekvenčně blízké aminokyseliny mohou ovlivňovat výsledky analýz. Především v kombinaci se sekundárními strukturami.

8.1.2 Sekundární struktury

Některé články analyzují interakce aromatických aminokyselin v rámci sekundárních struktur. Například aromatické interakce uvnitř a v okolí α-helixů (Bhattacharyya et al., 2002), jejich přispění ke stabilitě α-helixu (Butterfield et al., 2002) nebo jejich role ve stabilizaci β-listů (Budyak et al., 2013). Již tyto práce mohou napovědět, že sekundární struktury mohou ovlivňovat některé analýzy, které se zabývají prostorem v proteinu.

Informaci, jestli se daná aminokyselina vyskytuje v definované sekundární struktuře, jsem získal z PDB databáze[10]. Ta používá pro jejich identifikaci algoritmus DSSP, který rozlišuje sedm druhů sekundárních struktur (Kabsch a Sander, 1983). Jedná se o otočku s vodíkovým můstkem (T), z nich je složen α-helix (H), 3-helix (G) a 5 helix (I). Dále izolovaný „β-bridge“ (B, složený z jednoho vodíkového můstku a dvou aminokyselin), který vytváří β-vlákna (E) tvořící β-skládané listy. Poslední je ohyb (S)[11]. Využil jsem již získaných deset prostorově nejbližších aminokyselin okolo každého tryptofanu a u každé jsem zjistil její výskyt v sekundárních strukturách. Bohužel opět kvůli nepřesnostem v PDB souborech se mi podařilo určit sekundární struktury pouze u 378 390 aminokyselin. Z toho 305 985 se vyskytovaly v nějaké výše definované sekundární struktuře. 72 % byly rovnoměrně rozděleny v α-helixu nebo β-listu, zbytek se skládal převážně z ohybů a otoček. Z důvodu zjednodušení a malého výskytu ostatních sekundárních struktur jsem analyzoval pouze tři skupiny α-helixy (H), β-vlákna (E) a výskyt „bez sekundárních struktur“ (budu označovat jako „0“). V následujících obrázcích je vždy α-helix zobrazený červeně, β-vlákna modře a šedivě je zobrazena aminokyselina bez sekundární struktury.

[10] https://www.rcsb.org/pdb/static.do?p=download/http/index.html
[11] https://www.rcsb.org/pdb/static.do?p=help/ssHelp.html

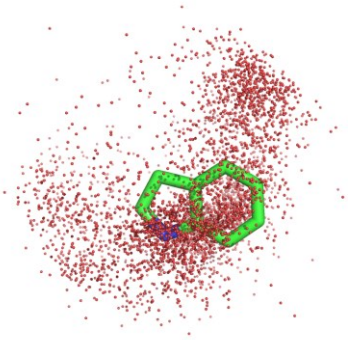
Nejprve jsem zobrazil rozmístění aminokyselin okolo tryptofanu pro každou vybranou sekundární strukturu s tím, že jsem zkoumal dvě různé situace. Aminokyselina musela být v sekvenci vzdálena od tryptofanu 1) do deseti aminokyselin včetně (budu odkazovat jako sekvenčně blízké) nebo 2) od jedenácti aminokyselin a více (sekvenčně vzdálené). Hranici deseti aminokyselin jsem zvolil na základě α-helixu, kde na jeden závit připadá 3,6 aminokyselin. Na této hranici určitě aminokyseliny mohou interagovat, viz například obrázek 6.9 (A). Pro zajištění dostatečné vzdálenosti jsem zvolil tři otáčky α-helixu, což odpovídá právě jedenácti aminokyselinám.

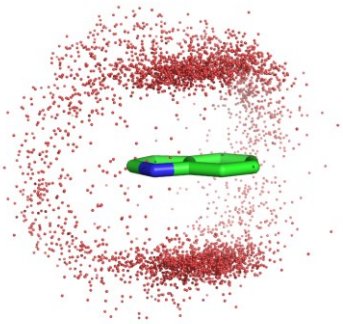
Od každého tryptofanu jsem vybral pouze jeho nejbližší aminokyselinu, která byla ve stejném typu sekundární struktury jako daný tryptofan a výsledek je na obrázku 8.1. Je zobrazeno pouze 4 000 aminokyselin ze 4 000 tryptofanů z důvodu malého množství aminokyselin bez sekundární struktury a chtěl jsem, aby v každém obrázku byl stejný počet bodů. U blízkých v sekvenci (A, C, a E) je u sekundárních struktur patrná (především u α-helixu) tendence ke vzniku oblaků preferenčního výskytu.

A - Bez sekundárních struktur (sekvenčně blízké)

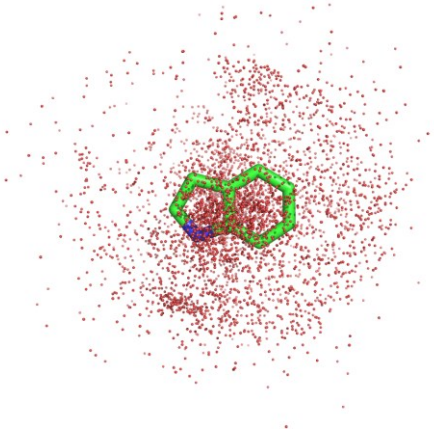
B - Bez sekundárních struktur (sekvenčně vzdálené)

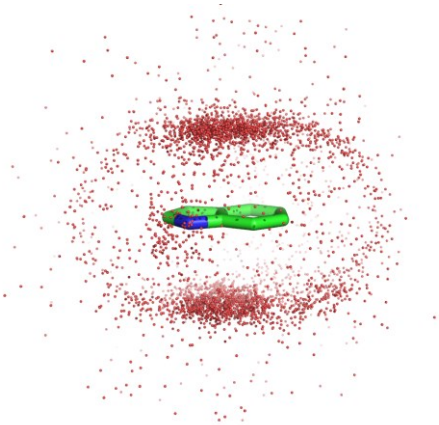
C - α-helixy (sekvenčně blízké)



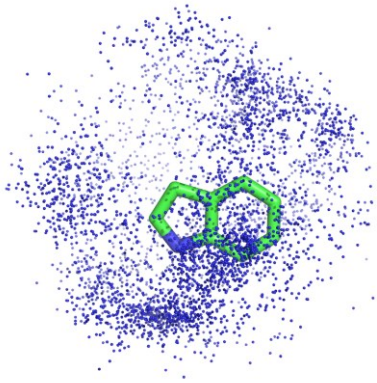


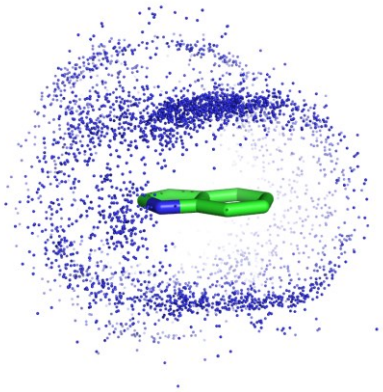
D - α-helixy (sekvenčně vzdálené)



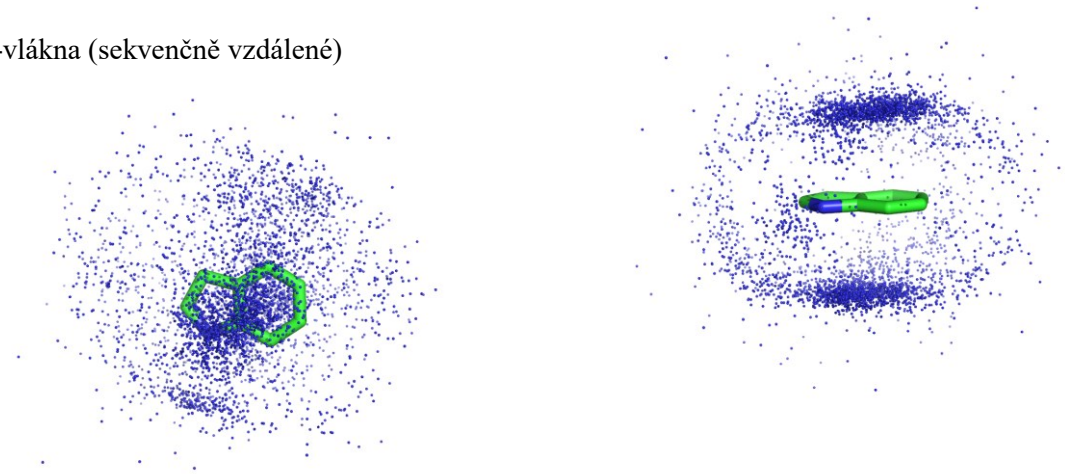


E - β-vlákna (sekvenčně blízké)



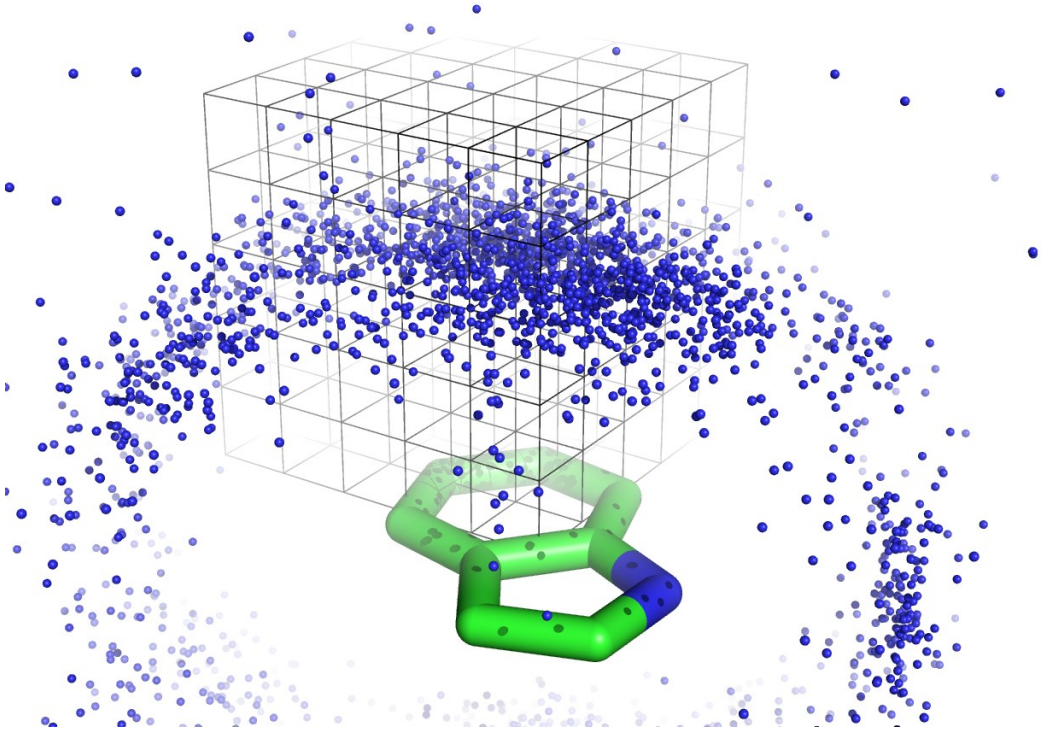


F - β-vlákna (sekvenčně vzdálené)



Obrázek 8.1: Zobrazených 4 000 aminokyselin, pro 4 000 unikátních tryptofanů. Každý tryptofan má zobrazenu jednu prostorově nejbližší aminokyselinu odpovídající stejné sekundární struktuře (A-F) jako daný tryptofan. Zobrazeny jsou vždy dva různé pohledy na okolí tryptofanu.

Pro statistické porovnání výskytů jsem zvolil trochu složitější postup. Prostor okolo tryptofanu z obrázku 8.1 jsem rozdělil do krychlí o délce hrany 1 Å. Pro představu je na obrázku 8.2 ukázán výřez.



Obrázek 8.2: Výřez prostoru okolo tryptofanu rozděleného na krychle o délce hrany 1 Å.

Po tomto rozdělení jsem porovnával výskyt v jednotlivých krychlích pro jednotlivé dvojice z obrázku 8.1 (A-B, C-D, E-F, A-C, A-E, B-D, B-F). Jde o to, aby byl zvýrazněn rozdíl mezi těmito dvojicemi. Postup vysvětlím na dvojici obrázků C-D. Nejprve sečtu výskyt aminokyselin vždy mezi jednotlivými krychlemi z obrázku C a jejími ekvivalenty z obrázku D. Takže pro každou krychli je dáno číslo, které vznikne součtem těchto krychlí z obrázků C a D. Následně stanovím hranici, pod kterou nebudu danou krychli analyzovat. To znamená, že pokud součet bodů v obou krychlích je menší, než jsou součty 25 % nejobsazenějších krychlí, tak daná krychle nebude analyzována. Pro obrázky C-D byla tato hranice 6, to znamená, že výskyt v daných dvou krychlích musí být po sečtení minimálně 6 aminokyselin (v drtivé většině případů je výskyt 1 nebo 2). Četnosti v odpovídajících krychlích v porovnávaných situacích následně vzájemně vydělím způsobem zobrazeným na obrázku 8.3. Například pokud jsou v krychli z obrázku C 2 body a v D deset bodů, tak jejich součet je více než 6, proto není krychle zahozena a proběhne následující výpočet. Protože je v C méně, tak vypočet bude 10/2=5, a jelikož došlo k prohození čitatele a jmenovatele, je podíl vynásoben -1. Následně se přičte 1 a konečný výsledek je hodnota -4.

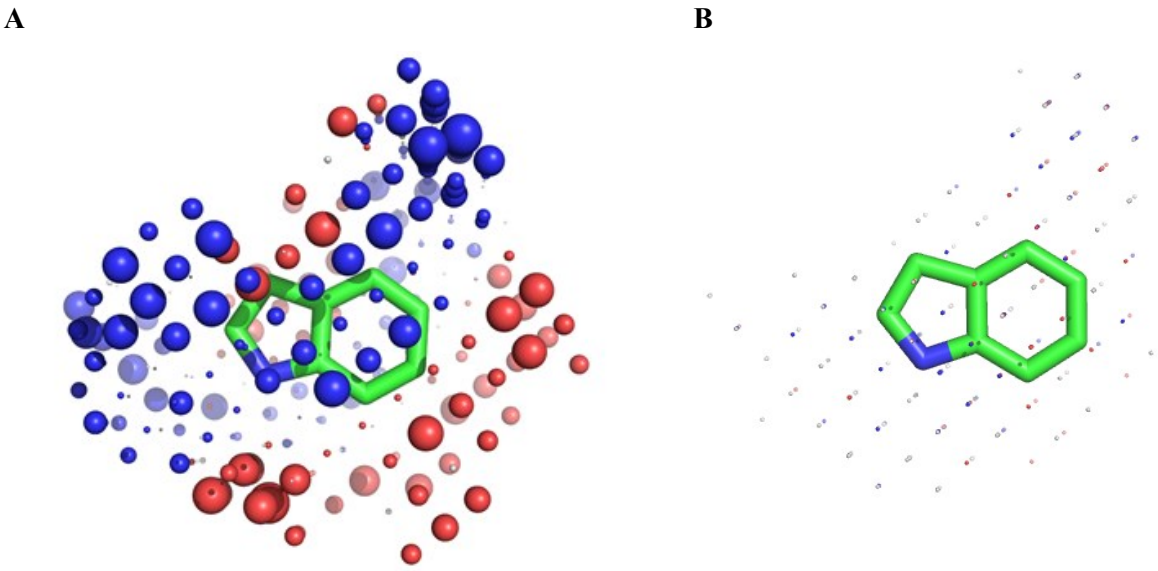
C > D: $\alpha = \frac{C}{D} - 1$

C < D: $\alpha = -\left(\frac{D}{C}\right) + 1$

Obrázek 8.3: Porovnání dvou četností (C, D) tak, aby při stejných hodnotách C = D vyšla hodnota α = 0. Pokud například platí C = 2 a D = 10, tak α = -(10/2)+1 = -4. Tato transformace je využívána pro porovnávání četností v různých zobrazeních v této práci.

Takový postup se vykoná pro každou krychli a parametr podílu α se následně zobrazí v podobě sfér. Průměr sféry odpovídá velikosti vypočítané hodnoty rozdílu četností, pokud byl čitatel menší než jmenovatel, je červená, jinak je modrá. Navíc pokud je rozdíl v četnosti menší než odmocnina z vyšší hodnoty, tak bude sféra bílá. Výsledek porovnání situací C/D (z obrázku 8.1) je na obrázku 8.4 (A). Je vidět, jak D má zřetelně nadhodnocenou oblast, zatímco červené D pouze vyplňuje prázdný prostor, který je v C. Tento obrázek je zajímavý spíše tím, že je na něm dobře ukázaný postup, protože rozdělení je celkem zřejmé už v původních obrázcích. Je patrný lokalizovaný pruh modrých sfér okolo Cβ, což pravděpodobně souvisí s α-helixem. Místa červených sfér jsou dána spíše prázdných prostorem v C, než výrazným výskytem v D. Pro negativní kontrolu byl proveden stejný postup pro data z obrázku 8.1 (C), která byla rozdělena na dvě náhodné poloviny. Tyto

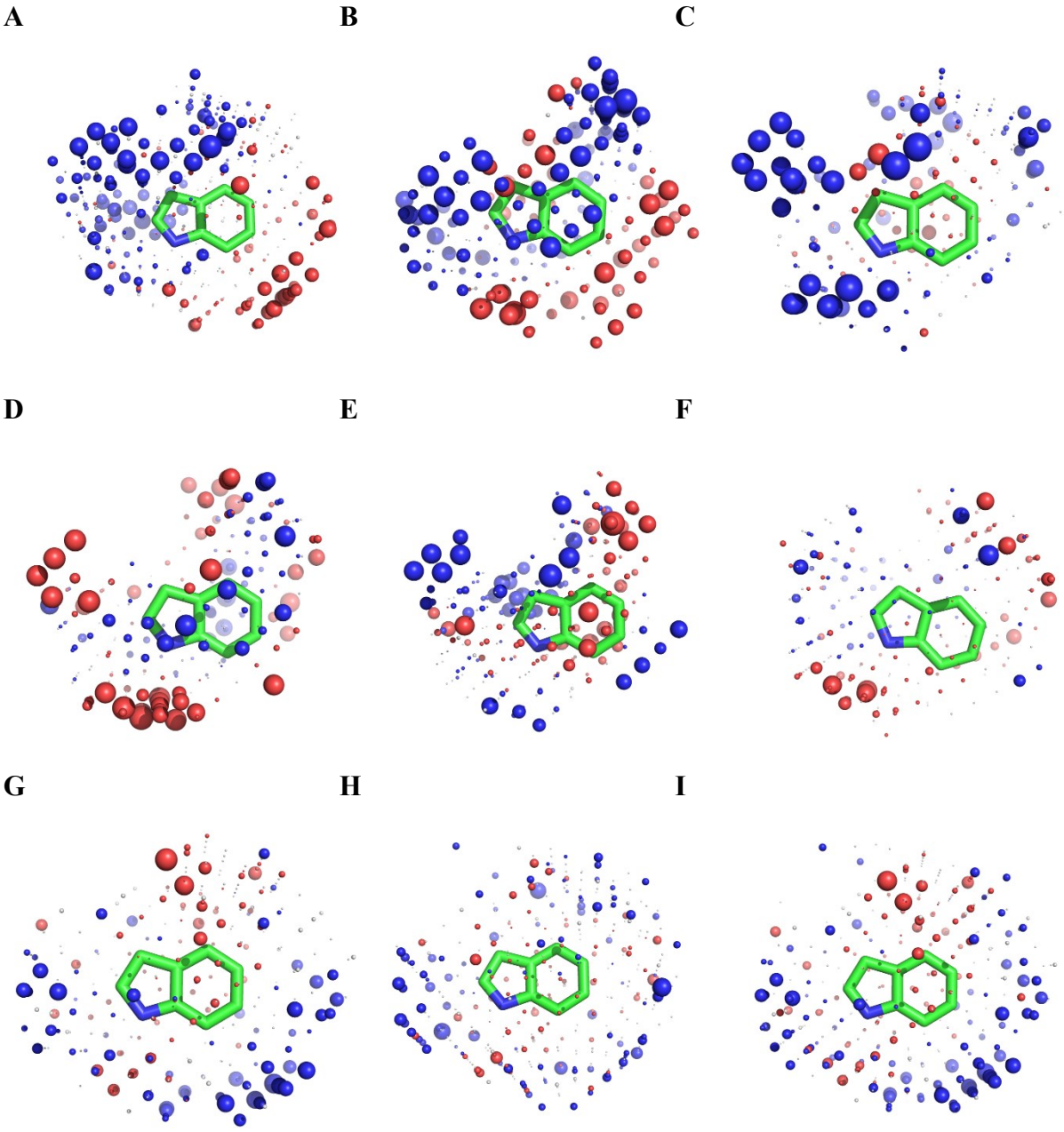
poloviny byly následně mezi sebou porovnány stejným postupem popsaném výše a výsledek je na obrázku 8.4 (B). Je patrné, že žádná preference v prostoru není pozorovatelná.



Obrázek 8.4: (A) Zobrazeno porovnání obrázků C/D z obrázku 8.1. Modré sféry odpovídají nadbytku v C, zatímco červené je pro D. (B) Negativní kontrola. Porovnání dvou náhodných polovin dat z obrázku 8.1 (C). Velikost kuliček vyjadřuje zvýšenou/sníženou četnost výskytu aminokyseliny v daném prostorovém segmentu.

Kompletní výsledky jsou na obrázku 8.5. Z mého pohledu jsou nejdůležitější obrázky E a H, protože porovnávají stejné sekundární struktury, jednou sekvenčně blízké (obrázek 8.5 (E)) a podruhé sekvenčně vzdálené aminokyseliny (obrázek 8.5 (H). Je vidět, že obrázek H nemá oproti E téměř žádnou preferenci (žádné velké lokalizované sféry). Jedině u β-vláken je určitá prostorová preference u sekvenčně vzdálených aminokyselin, a to pravděpodobně kvůli tvořícím se β-listům. Důležitý je i obrázek A, který vypovídá o preferenci u sekvenčně blízkých, i přesto, že nejsou v žádné definované/rozpoznané sekundární struktuře.

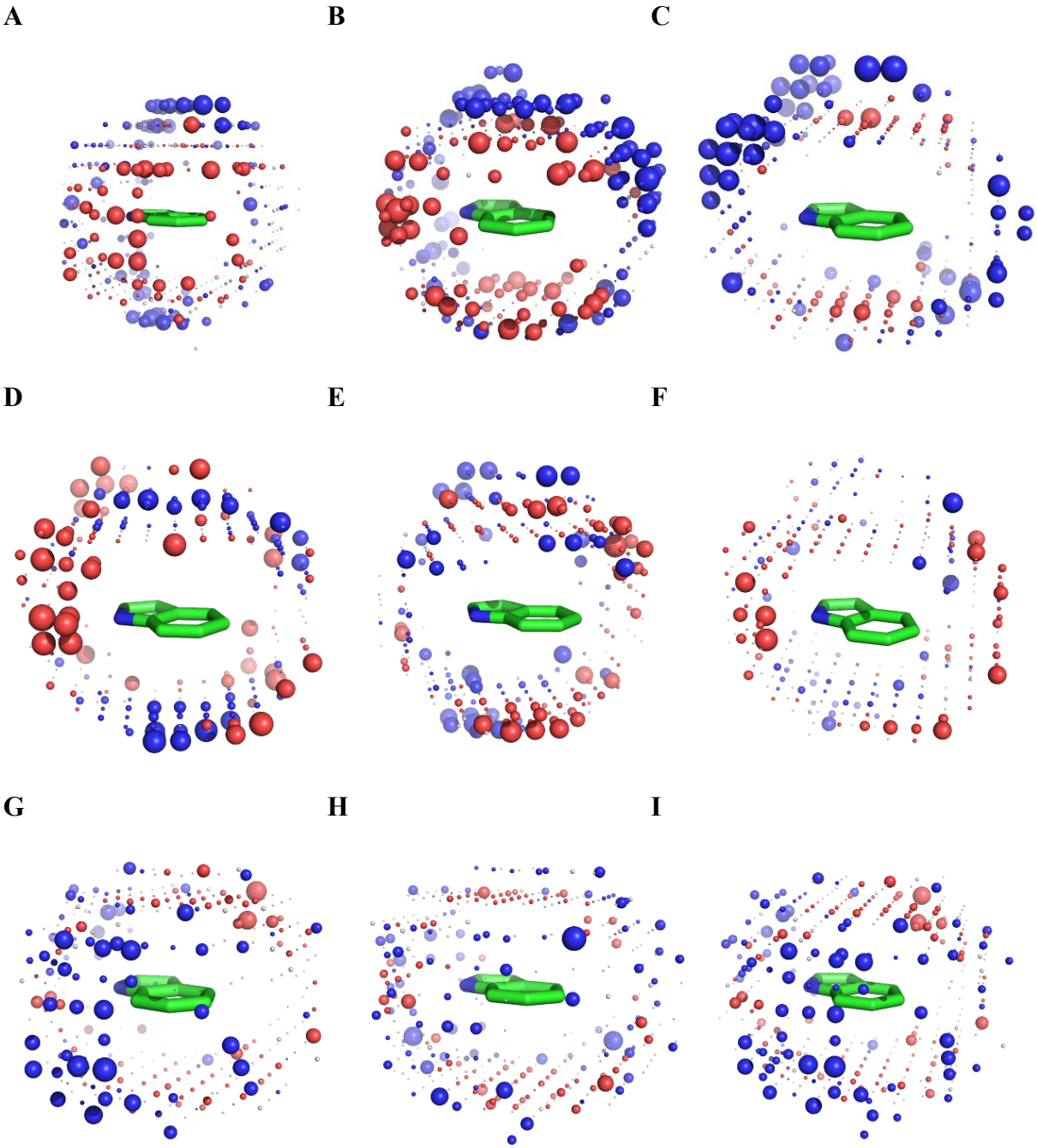
Z výsledků je evidentní, že sekvenčně blízké aminokyseliny, bez ohledu na sekundární struktury, se budou častěji vyskytovat na vynucených pozicích (především pokud jsou v α-helixu). Téměř všichni autoři uvedení v přehledu literatury tento problém pří analýze okolí a různých interakcí ignorují. Jediný, kdo upozorňoval na možné ovlivnění, byl Thomas et al. (2002).



Obrázek 8.5: Zobrazeno porovnání dat z obrázku 8.1. Modrá barva značí preferenci dat čitatele, červená jmenovatele (podrobné vysvětlení viz text):
A) Bez sekundárních struktur (sekv. blízké) / Bez sekundárních struktur (sekv. vzdálené).
B) α-helixy (sekv. blízké) / α-helixy (sekv. vzdálené).
C) β-vlákna (sekv. blízké) / β-vlákna (sekv. vzdálené).

D) α-helixy (sekv. blízké) / β-vlákna (sekv. blízké).
E) Bez sekundární struktury (sekv. blízké) / α-helixy (sekv. blízké).
F) Bez sekundární struktury (sekv. blízké) / β-vlákna (sekv. blízké).

G) α-helixy (sekv. vzdálené) / β-vlákna (sekv. vzdálené).
H) Bez sekundární struktury (sekv. vzdálené) / α-helixy (sekv. vzdálené).
I) Bez sekundární struktury (sekv. vzdálené) / β-vlákna (sekv. vzdálené).



Obrázek 8.6: Zobrazena data z obrázku 8.5, pouze z jiného pohledu. Modrá barva značí preferenci dat čitatele, červená jmenovatele (podrobné vysvětlení viz text):
A) Bez sekundárních struktur (sekv. blízké) / Bez sekundárních struktur (sekv. vzdálené).
B) α-helixy (sekv. blízké) / α-helixy (sekv. vzdálené).
C) β-vlákna (sekv. blízké) / β-vlákna (sekv.vzdálené).

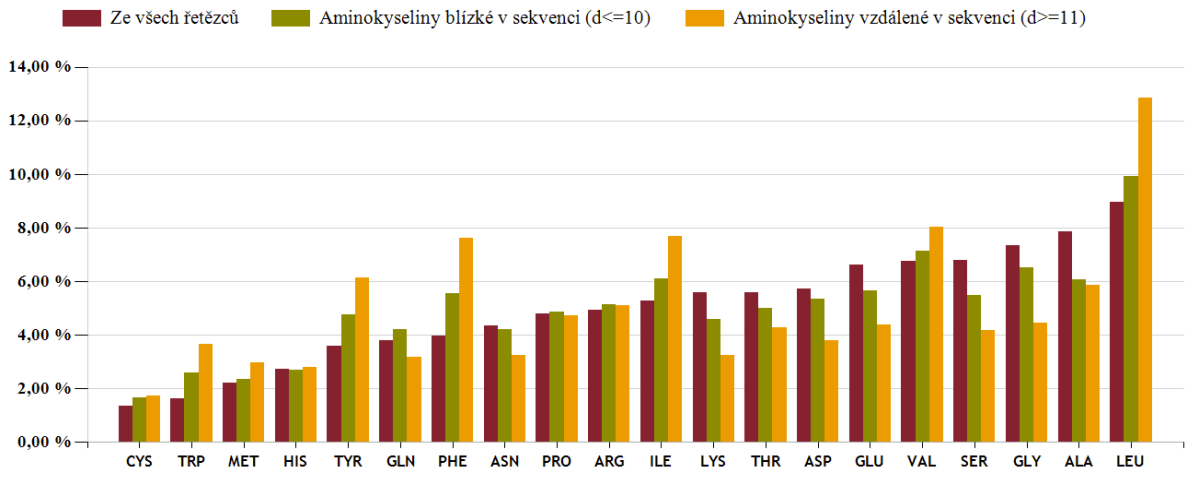
D) α-helixy (sekv. blízké) / β-vlákna (sekv. blízké).
E) Bez sekundární struktury (sekv. blízké) / α-helixy (sekv. blízké).
F) Bez sekundární struktury (sekv. blízké) / β-vlákna (sekv. blízké).

G) α-helixy (sekv. vzdálené) / β-vlákna (sekv. vzdálené).
H) Bez sekundární struktury (sekv. vzdálené) / α-helixy (sekv. vzdálené).
I) Bez sekundární struktury (sekv. vzdálené) / β-vlákna (sekv. vzdálené).

8.1.3 Analýza prostorového okolí tryptofanu

Nyní se budu zabývat samotným okolím tryptofanu. To znamená především tendencemi jednotlivých aminokyselin být v blízkosti indolu. Dále jednotlivými interakcemi popsanými v kapitole 6.4 a dalšími průběžně vymezenými cíli.

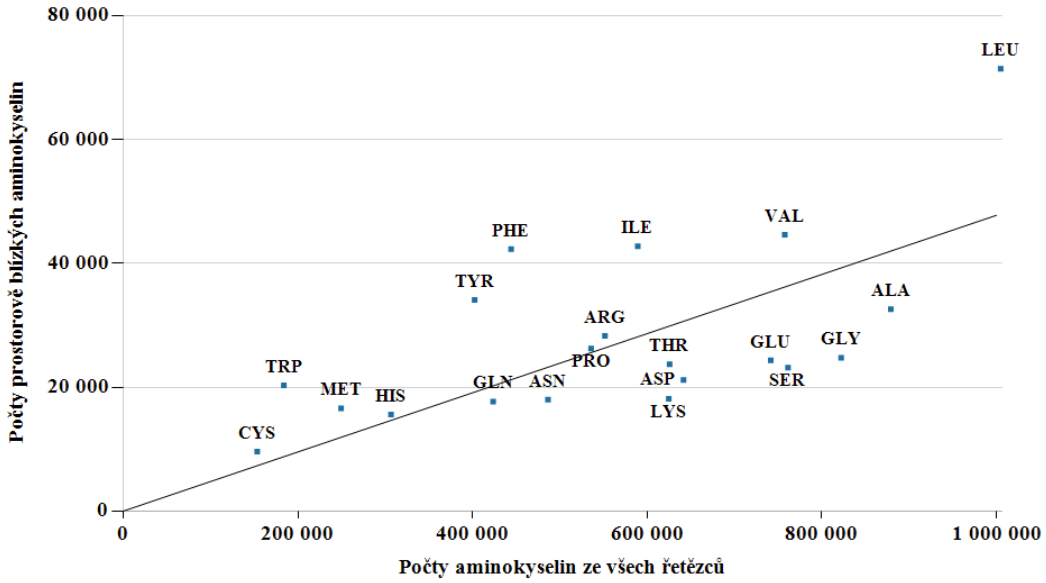
Pro tuto analýzu jsem opět použil prvních deset nejbližších aminokyselin v prostoru okolo tryptofanu, které jsem již využil pro odstraňování homologie řetězců v kapitole 7.5.3. deset aminokyselin je dostačujících, protože většinou obalí celou indolovou skupinu (viz obrázek 7.7). Připomínám, že aminokyselina je zde pracovně definována nejbližším atomem jejího postranního řetězce od těžiště indolu. Analýza v kapitole 8.1.2 ukázala, že výskyt aminokyselin u tryptofanu je ovlivněn sekvenční blízkostí. Pro ujištění jsem ještě vytvořil graf 8.7, který zobrazuje procentuální výskyt aminokyselin ze všech analyzovaných řetězců a procentuální výskyt aminokyselin v deseti prostorově nejbližších aminokyselinách od těžiště indolu. Těchto deset aminokyselin jsem navíc rozdělil na sekvenčně blízké (do vzdálenosti deset od tryptofanu, které jsou pravděpodobně ovlivněny sekundární strukturou nebo sekvencí) a sekvenčně vzdálené (od jedenácti). Na tomto grafu je patrná odstupňovaná preference jednotlivých aminokyselin k tryptofanu, jelikož sekvenčně blízké aminokyseliny (zelený sloupec) jsou téměř vždy svojí četností mezi všemi (hnědý sloupec) a sekvenčně vzdálenými (žlutý sloupec) aminokyselinami.



Obrázek 8.7: Procentuální četnosti aminokyselin v blízkosti indolu. Barevně jsou odlišeny následující skupiny: 1) Četnosti aminokyselin ze všech analyzovaných řetězců bez ohledu na jejich sekvenční a prostorovou vzdálenost (11 183 960 aminokyselin), seřazeno vzestupně. 2) Četnosti z deseti prostorově nejbližších od tryptofanu, které jsou od něj v sekvenci do desáté pozice (367 431 aminokyselin). 3) Procentuální četnosti z deseti prostorově nejbližších aminokyselin od tryptofanu, které jsou od něj v sekvenci na jedenácté pozici a dále (555 539 aminokyselin). Aminokyseliny jsou seřazeny podle vzrůstající četnosti výskytu v analyzovaných řetězcích.

Tento jev by se dal interpretovat tak, že na sekvenčně blízkých aminokyselinách je již vidět preference k tryptofanu, ale pořád jsou nějak ovlivněny (sekvencí, sekundárními strukturami). Teprve sekvenčně vzdálené aminokyseliny (žlutý sloupec) mají volnost k plné interakci s tryptofanem. Na grafu je tedy jasně vidět, že v sekvenci vzdálené aminokyseliny mají jasnější specifitu k tryptofanu než blízké, a proto se v dalších analýzách zaměřím pouze na aminokyseliny sekvenčně vzdálené.

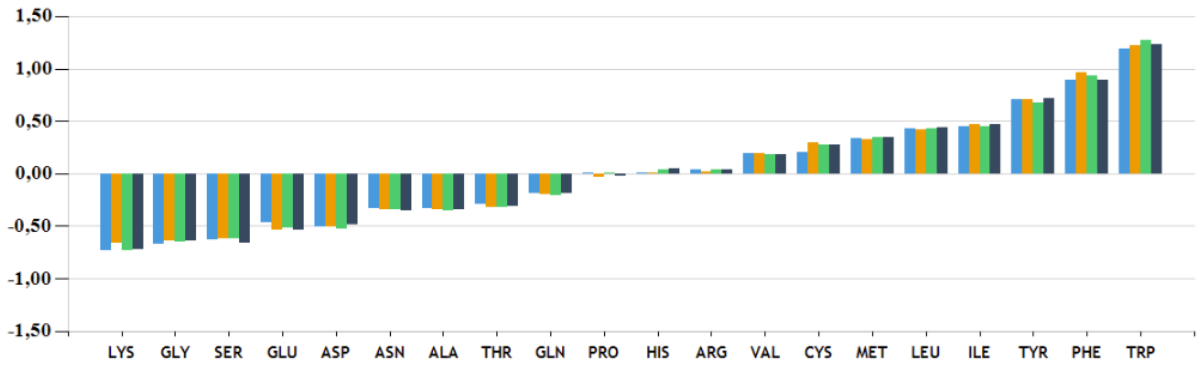
Pro lepší zobrazení preference jsem na grafu 8.8 vynesl četnosti těchto sekvenčně vzdálených aminokyselin oproti sumárním četnostem v analyzovaných řetězcích. Například prolin leží v grafu 8.8 prakticky na přímce, což odpovídá jeho „nulové“ preferenci vyskytovat se v blízkosti indolu. Totéž se dá usoudit ze sloupcového grafu 8.7, kde prolin vykazuje shodné relativní četnosti ve všech zkoumaných vzdálenostních skupinách. Tento graf nejen vypovídá o preferenci, ale i o rozsahu prováděné analýzy. Například i nejméně početného cysteinu bylo analyzováno skoro 10 000. I přesto, že tento graf ukazuje kladnou nebo zápornou tendenci jednotlivých aminokyselin k blízkému výskytu u tryptofanu (výskyt nad/pod regresní přímkou), tak se domnívám, že pořád není zřejmé pořadí jednotlivých aminokyselin při zkoumání této jejich vlastnosti. Je potřeba preferenci nějak kvantifikovat.



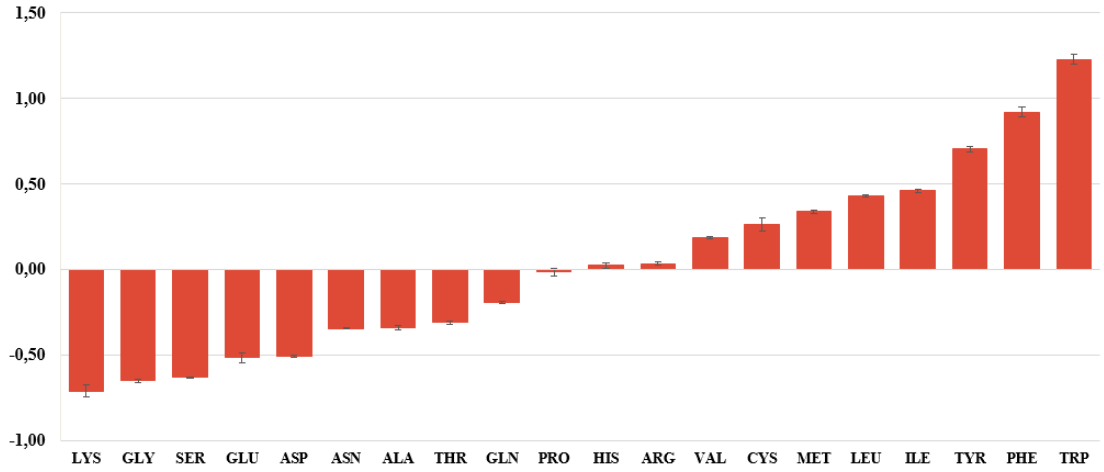
Obrázek 8.8: Vynesení absolutních četností deseti prostorově blízkých a zároveň sekvenčně vzdálených aminokyselin proti všem aminokyselinám z analyzovaných řetězců. Body nad regresní přímkou by měly naznačovat tendenci dané aminokyseliny vyhledávat blízkost indolové skupiny.

Jako nejlepší zobrazení pro kvantifikaci mi přišlo výsledky pro sekvenčně vzdálené aminokyseliny normalizovat na počty aminokyselin ze všech analyzovaných řetězců.

Tato normalizace je provedena na základě grafu 8.7, což znamená, že relativní četnosti (%) sekvenčně vzdálených a zároveň prostorově blízkých aminokyselin jsou vyděleny relativními četnostmi (%) stejné aminokyseliny ve všech analyzovaných řetězcích (hodnoty žlutého sloupce vyděleny hodnotami hnědého sloupce). Výsledkem je tedy bezrozměrná veličina. Při výpočtu je opět využita transformace definována na obrázku 8.3. Abych ukázal statisticky validní data, použil jsem metodu Bootstrap (Efron, 1979; Markus a Groenen, 1998), která je využívána pro statistickou analýzu dat. Klasický přístup metody je založen na rozdělení studovaných dat na stejně velké díly a jejich porovnání, jak je zobrazeno na grafu 8.9, kde jsou data rozdělena na čtyři díly. Zjednodušeně řečeno, pokud vykazuje pozorovaný trend každá část (například čtvrtina) z původních studovaných dat, pak lze usuzovat, že se jedná o skutečný jev a nikoli pouze náhodu. Výsledný graf je na obrázku 8.10 a obsahuje průměrnou hodnotu preference a směrodatnou odchylku, která je vypočítaná z těchto náhodných čtvrtin. Na základě malé odchylky by prakticky všechny pozorované rozdíly v preferenci aminokyselin k blízkosti indolu měly být statistický signifikantní.



Obrázek 8.9: Preference deseti prostorově nejbližších aminokyselin od tryptofanu, které jsou v sekvenci od tryptofanu dále než deset aminokyselin, normalizované na četnosti všech aminokyselin z analyzovaných řetězců (z grafu 8.7). Data jsou rozdělena na náhodné čtvrtiny a transformována podle obrázku 8.3.



Obrázek 8.10: Kvantifikované preference jednotlivých aminokyselin k indolu. Zobrazena celková data z předchozího grafu 8.9, kde rozdíl v náhodných čtvrtinách po Bootstrap analýze je znázorněn směrodatnou odchylkou.

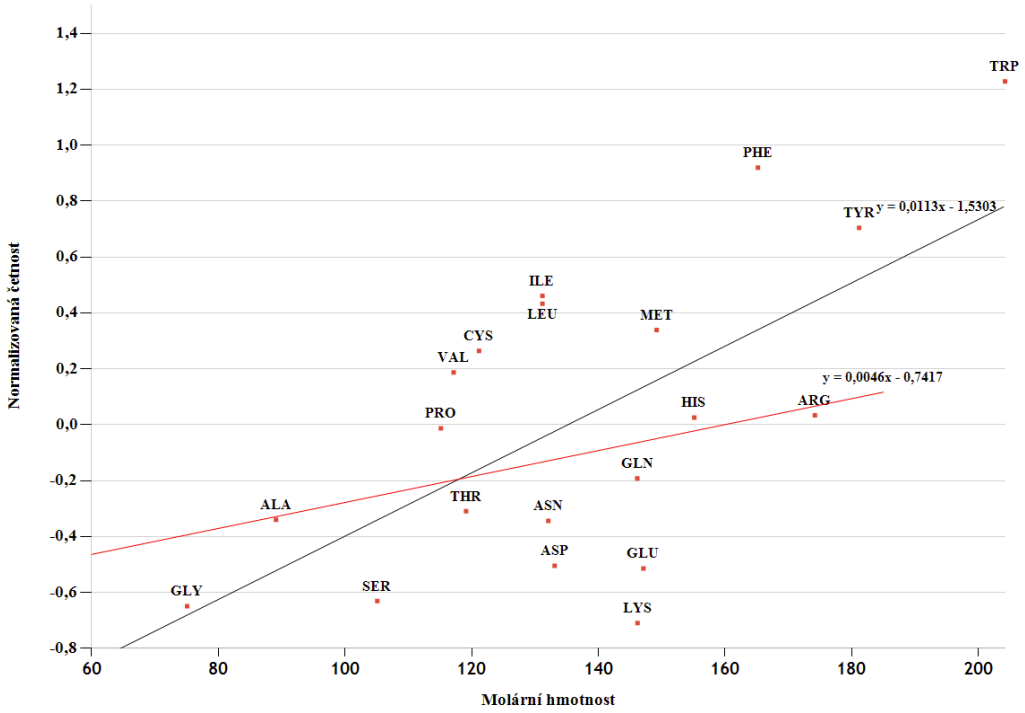
Graf 8.10 nabízí mnohá překvapení. Například články, které se zabývají kationt-π interakcí (viz kapitola 6.4.2), odkazují na preferenci argininu a lysinu k ploše indolu. Proč se tedy v grafu ukazuje, že arginin není u tryptofanů o moc více procentuálně zastoupený než ve zbytku proteinu? A lysin se dokonce u tryptofanu vyskytuje o hodně méně.

Preferenci leucinu a isoleucinu k tryptofanu reportovali ve výsledcích Samanta et al. (2000). Jedním z vysvětlení může být, že svým hydrofobním řetězcem vyplňují prostor mezi aromáty. Následují methionin a cystein, což pravděpodobně souvisí s popsanou sulfur-π interakcí (viz kapitola 6.4.6). Přes popsanou aniont-π interakci (viz kapitola 6.4.3) jsou záporně nabité aminokyseliny (Glu, Asp) v levé části grafu (bez preference, případně s „negativní preferencí“).

Samanta et al. uvádějí, že malé (Gly, Ala), negativně nabité (Asp, Glu) a polární (Ser, Thr) aminokyseliny se vyhýbají indolovému kruhu. V grafu se tyto aminokyseliny opravdu nalézají v levé části (bez preference), ale podle mého názoru se dá polemizovat nad tím, jestli se skutečně vyhýbají, nebo jsou například vytlačeny aminokyselinami se silnější preferencí. Tito autoři dále uvádějí, že aromatické aminokyseliny mají značnou preferenci, pouze pokud jsou posuzovány společně. Na grafu 8.10 ale jednotlivé aromatické aminokyseliny vykazují k tryptofanu jednoznačně největší preferenci. Tato preference by například mohla souviset s aromatickými klastry, tak jak uvádí Lanzarotti et al. (2011).

Tento graf zobrazuje preferenci určitých aminokyselin k tryptofanu, ale bude potřeba detailněji prozkoumat jednotlivé aminokyseliny, aby byl poskytnut komplexnější obrázek o prostorovém okolí tryptofanu. Nad konkrétními čísly lze nyní pouze spekulovat.

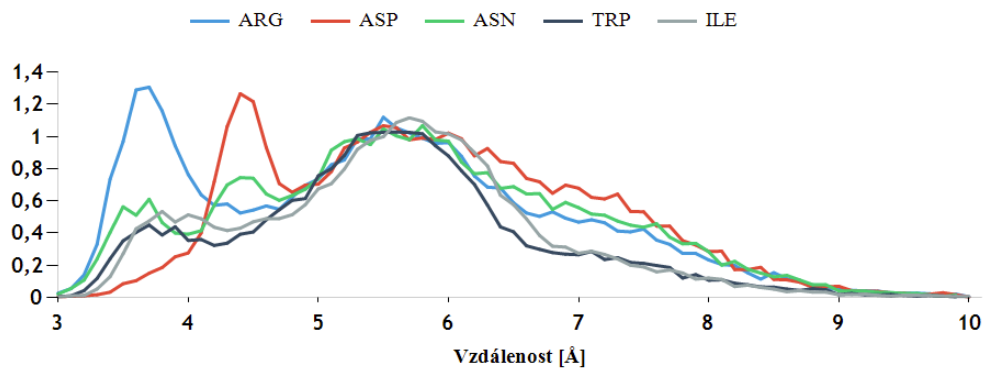
Po vytvoření normalizovaného grafu 8.10 se nabízí otázka, jestli pořadí aminokyselin nemůže nějak souviset s velikostí molekul (postranních řetězců)? Například aromatické aminokyseliny s největší preferencí (pravá část grafu) mají největší molekulovou hmotnost. Dále mnoho malých aminokyselin (glycin, serin, alanin) se zase vyskytuje v levé části grafu. Pro odpověď jsem vytvořil graf 8.11, který zobrazuje výsledky z grafu 8.10 proti molární hmotnosti dané aminokyseliny. Na první pohled se zdá, že se v datech závislost vyskytuje. I lineární regrese ze všech aminokyselin (černá přímka) má stoupající tendenci. Nicméně podle mě je nutné se zamyslet nad aromatickými aminokyseliny, které je nutné brát v tomto případě jako skupinu aromátů, tvořící například aromatické klastry. Tím by mohly být velmi nadhodnoceny. Pokud vypočítám lineární regresi (červená přímka) bez tryptofanu, tyrosinu a fenylalaninu, tak závislost již není tak výrazná. Pro obě lineární regrese jsem u obou hodnot směrnic provedl t-test, jestli opravdu závislost existuje (tedy, jestli se směrnice liší od nuly). Na základě kritické hodnoty (CI 99 %) u lineární regrese všech aminokyselin (černě, směrnice 0,0113, chyba směrnice 0,003206) vyšla dolní mez konfidenčního intervalu směrnice 0,002094. Jelikož je hodnota nad nulou, lze považovat závislost za statisticky významnou. U lineární regrese bez aromátů (červeně, směrnice 0,0046, chyba směrnice 0,003953) vyšla dolní mez konfidenčního intervalu směrnice při stejné kritické hodnotě -0,00701. To znamená, že statisticky významná závislost zde již patrně neexistuje.



Obrázek 8.11: Závislost normalizované četnosti na molární hmotnosti aminokyselin. Zobrazeny výsledky z grafu 8.10, proti molární hmotnosti dané aminokyseliny. Součástí grafu jsou dvě lineární regrese. „Černá“ regrese je přes všechny aminokyseliny. „Červená“ je bez tryptofanu, tyrosinu a fenylalaninu.

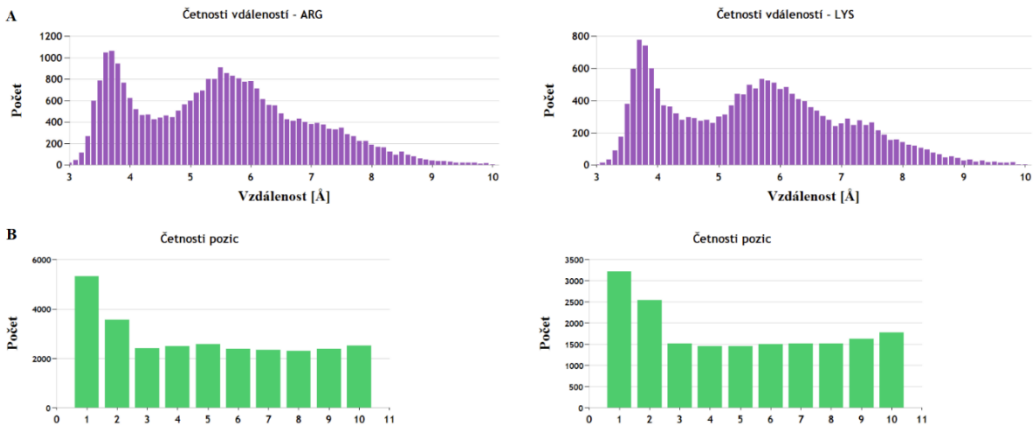
Statistické výsledky na grafu 8.11 vypovídají o závislosti aminokyselin k molekule indolu podle jejich molární hmotnosti. Nicméně při vynechání tří aromatických aminokyselin (tryptofan, tyrosin a fenylalanin) se závislost ztratí. Osobně si myslím, že závislost na molární hmotnosti není a výsledek lineární regrese všech aminokyselin je způsoben něčím jiným, konkrétně aromatickým charakterem nejpreferovanějších aminokyselin.

Nyní bych se rád věnoval analýze prostorového rozmístění jednotlivých aminokyselin od indolové skupiny tryptofanu. Detailně ukázat v grafu všech dvacet aminokyselin by bylo nepřehledné, proto jsem z nich vybral pět reprezentativních pro určité vlastnosti. 1) Arginin reprezentující kladně nabité aminokyseliny. 2) Aspartát zase záporně nabité. 3) Asparagin, který obsahuje v postranním řetězci kyslík i dusík. 4) Tryptofan za aromatické aminokyseliny a 5) isoleucin za hydrofobní. Pro každou aminokyselinu jsem vytvořil histogram vzdálenosti jejího nejbližšího atomu v postranním řetězci od těžiště indolu. Rozlišení histogramu je 0,1 Å. Jednotlivé profily aminokyselin jsem překryl přes sebe, aby byly dobře patrné rozdíly. Překrytí jsem docílil vydělením každé hodnoty průměrem hodnot četností v intervalu 5,2-5,8 Å a výsledek je na obrázku 8.12. Důležitá je levá část grafu do 5,5 Å, protože pravá část je dána omezením zkoumaných aminokyselin na prvních deseti prostorově nejbližších od těžiště indolu. Bez omezení by relativní četnosti v grafu neustále narůstaly z důvodu zvětšujícího se prostoru okolo tryptofanu, jak už je pospáno v kapitole odstraňování homologie 7.5.3. Jednotlivé aminokyseliny mají v levé části grafu specifický profil, který následně podrobím detailnější analýze. U tohoto grafu bych rád vysvětlil, proč se nebudu zabývat normalizací prostoru, tak jak to vytýkám některým autorům v kapitole 6.4. Je to z důvodu, že budu porovnávat u jednotlivých aminokyselin jejich profily a ne absolutní četnosti. Navíc jako ideální graf se ukázal být histogram pozic (pořadí od těžiště indolu), který v kombinaci s histogramem vzdáleností ideálně zobrazuje preferenci aminokyselin k těžišti indolu. Histogram pozic není z definice potřeba normalizovat.



Obrázek 8.12: Profily relativní četnosti výskytu různých typů aminokyselin v různých vzdálenostech od těžiště indolu.

Detailní analýza aminokyselin z grafu 8.12 se bude skládat ze tří zobrazení. Pro vybranou aminokyselinu vždy vytvořím histogram vzdáleností od těžiště indolu odpovídající datům v grafu 8.12, a pod ním histogram pozic, který udává, kolikátá je daná aminokyselina v pořadí od těžiště indolu. Následně vytvořím prostorové zobrazení daných aminokyselin okolo indolu. Každá aminokyselina bude reprezentovaná svým nejbližším atomem z postranního řetězce a tyto atomy budou obarveny podle jejich typu (uhlík, kyslík, dusík). Histogramy vytvořím i pro další podobnou aminokyselinu, aby byly výsledky relevantnější. Jak už jsem uvedl výše, kombinace grafů vzdáleností (fialový) a grafu pozic (zelený) se ukázaly jako nejideálnější zobrazení preferencí jednotlivých aminokyselin k tryptofanu. Pro zajímavé aminokyseliny se vždy pokusím zobrazit příklady konkrétních struktur z PDB databáze. Ukázkové pozice atomů a příklady konkrétních konfigurací aminokyselin vychází z datasetu tryptofanů, který je použit pro analýzy.



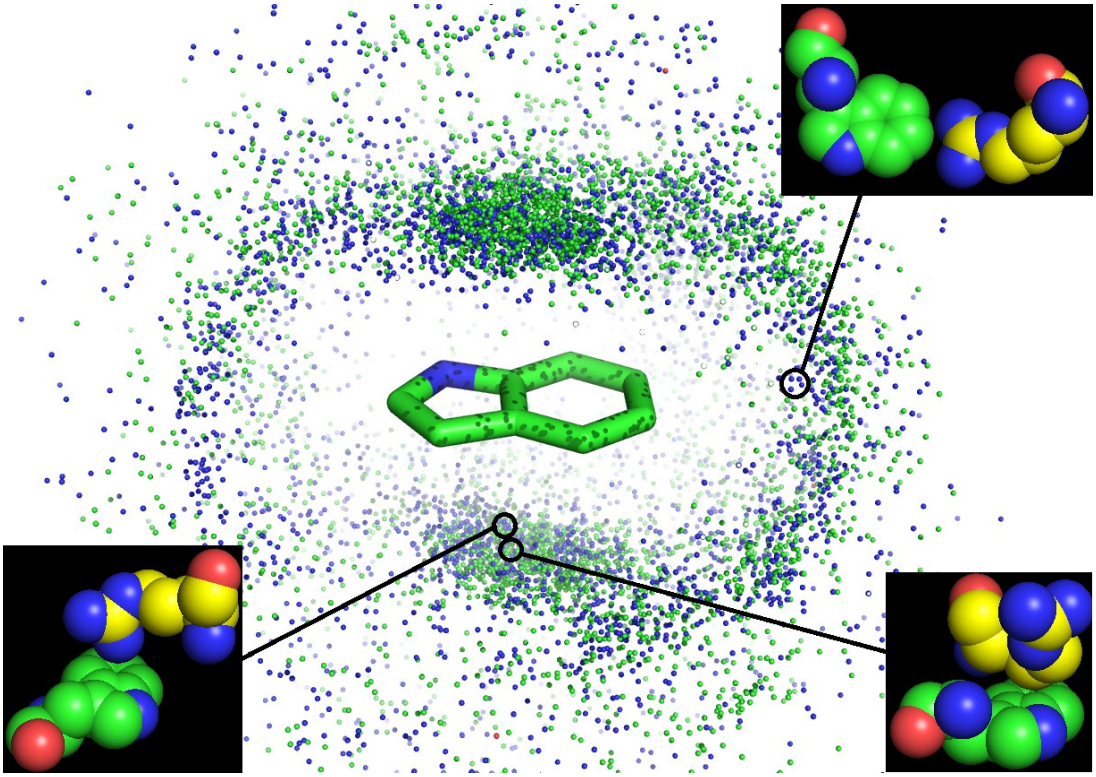
Obrázek 8.13: A) Histogram vzdáleností argininu a lyzinu (první nalezený atom z postranního řetězce) dané aminokyseliny od těžiště tryptofanu. B) Zelený graf ukazuje, kolikátá v pořadí deseti nejbližších aminokyselin v prostoru byla nalezena daná aminokyselina.

Jako první aminokyselinu jsem zvolil arginin, který má oproti jiným aminokyselinám zvýšený výskyt ve vzdálenosti 3,7 Å. K argininu se nejlépe hodí lysin, který má také kladný náboj postranního řetězce. Výsledné grafy jsou na obrázku 8.13 a na první pohled jsou profily grafů v podstatě totožné. Maximum ve vzdálenosti 3,7 Å u obou aminokyselin indikuje výraznou tendenci být u těžiště indolu, což odpovídá kationt-π interakci popsané v kapitole 6.4.2. Graf pozic tuto preferenci potvrzuje. Obě aminokyseliny mají oproti vzdálenějším pozicím (pravděpodobně již náhodným) více než dvojnásobný výskyt u těžiště indolu (obrázek 8.13 (B)). Nicméně proč je arginin v grafu 8.10 bez preference a lysin dokonce jako nejméně preferovaná aminokyselina? Nabízí se hypotéza, že arginin se u tryptofanu

nevyskytuje nijak preferenčně a lysin se dokonce u tryptofanu vyskytuje se zápornou preferencí, ale když se vyskytnou v jeho blízkosti, tak interagují těsně s plochou jeho indolové skupiny. Graf pozic lysinu navíc vykazuje ve vzdálenějších pozicích rostoucí tendenci (nárůst na pozicích devět a deset), což by mohlo odpovídat jeho celkově záporné preferenci k tryptofanu.

Na obrázku 8.14 jsou zobrazené nejbližší atomy postranního řetězce argininu v prostoru okolo indolové skupiny tryptofanu. Z obrázku je patrná určitá preference dusíků aminoskupin vyskytovat se spíše podél hran indolu. Na hraně indolu je ale parciální kladný náboj, takže mi není jasné, proč jsou v daném místě v tak hojném počtu. Většina atomů argininu se vyskytuje u těžiště indolu, ale přes popisovanou kationt-π interakci v kapitole 6.4.2 se uhlík u těžiště vyskytuje 1,5krát více než dusík. To může naznačovat, že arginin raději k indolu přiléhá alifatickým řetězcem, než by byl v kationt-π interakci.

Takovéto zobrazení nejbližších atomů má nevýhodu, že není poznat, jak je daná aminokyselina přesně orientovaná k indolu. Analyzovat orientaci vyžaduje značně komplexnější přístup a v kapitole 8.2 se o něj pokusím alespoň pro dvojice tryptofanů. Nicméně i toto prostorové zobrazení může mnohé napovědět o orientaci postranního řetězce, jak dokládají příklady konkrétních struktur.

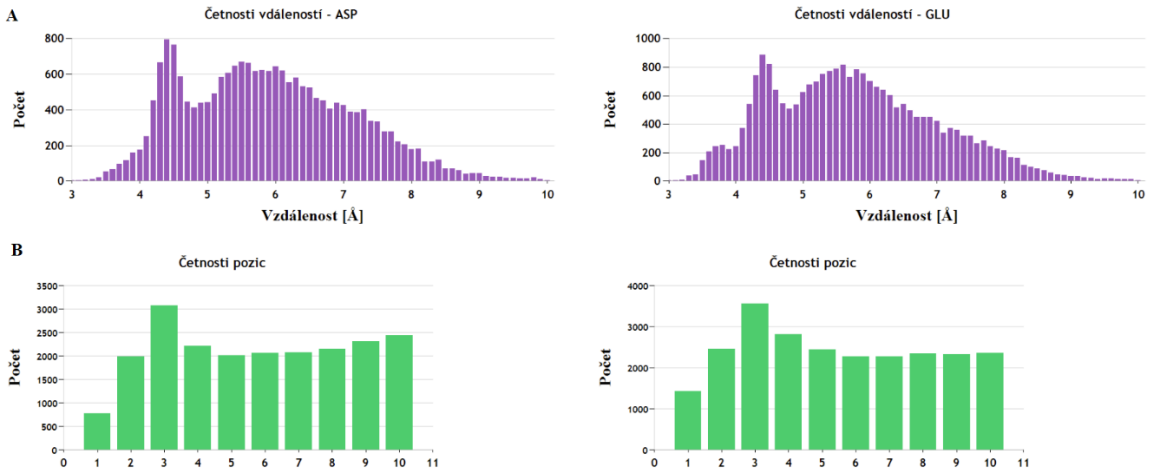


Obrázek 8.14: Prostorové zobrazení nejbližších atomů argininu okolo indolu tryptofanu. Zobrazeno je 10 000 atomů. Modré kuličky značí dusík a zelený je uhlík postranního řetězce argininu. Ukázkové struktury (PDB: 1JSG, 3V5L, 1YIZ) zobrazují tryptofan (zeleně) a arginin (žlutě) v konkrétních konfiguracích.

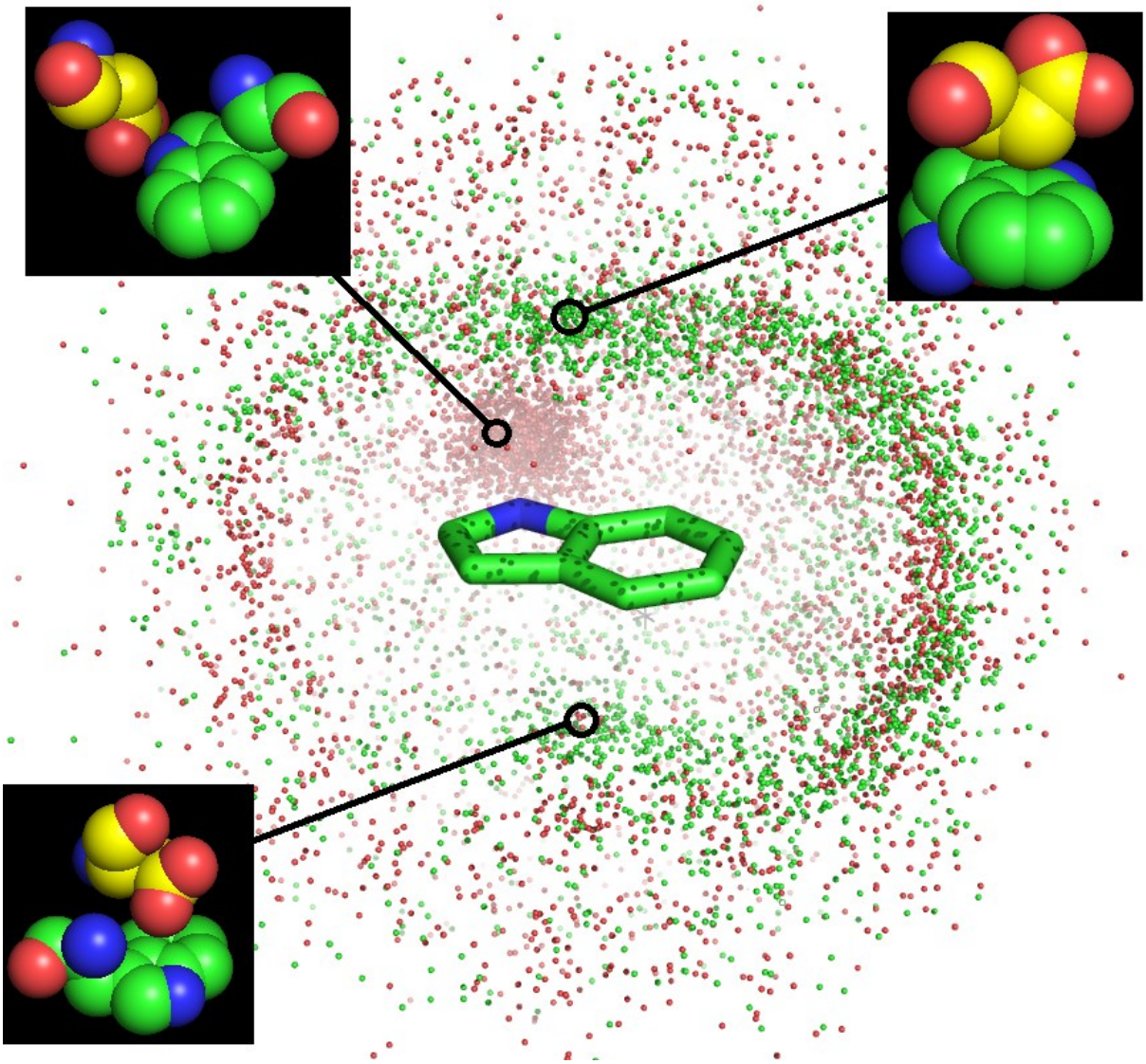
Další aminokyselina v grafu 8.12 je aspartát. Na rozdíl od argininu má maximum výskytu ve vzdálenosti okolo 4,5 Å, což je společné i pro glutamát, jak je vidět na obrázku 8.15. Podle grafu pozic jsou nejčastěji obě aminokyseliny až třetí v pořadí od těžiště indolu (obrázek 8.15 (B)). Pro detailnější popis je nutné zohlednit prostorové zobrazení. Jelikož glutamát má oproti aspartátu mírné navýšení četnosti okolo 3,7 Å, rozhodl jsem se pro jeho zobrazení (obrázek 8.16). Na obrázku je značný výskyt kyslíkových atomů u dusíku indolové skupiny, což odpovídá oné třetí pozici. Jedná se o vodíkový můstek mezi aminovou skupinou indolu a karboxylovou skupinou postranního řetězce glutamátu.

Kapitola 6.4.3 popisuje aniont-π interakce, které byly nalezeny dokonce i u tryptofanu (Lucas et al., 2016). Bohužel v mé práci v žádném zobrazení tato interakce není detekovatelná. Glutamát sice vykazuje mírné navýšení oproti aspartátu ve vzdálenosti 3,7 Å (odpovídá ploše indolu), ale na prostorovém zobrazení (8.16) není žádný velký výskyt karboxylové skupiny pozorován (červené sféry). Naopak se zde vyskytují atomy uhlíku (zelené sféry), které by mohly svědčit o tom, že glutamát na ploše indolu leží alifatickým řetězcem. Větší preference glutamátu k ploše indolu tedy může být z důvodu delšího postranního řetězce, který lépe přilehne k indolu. Z grafu pozic by se dalo usuzovat to samé, jelikož aspartát je na první pozici (nejblíže těžiště indolu) 3krát méně často než na vzdálenějších pozicích (od páté pozice dále). Zatímco glutamát ani ne 2krát méně.

Na závěr musím konstatovat, že aniont-π interakci jsem ve spojení s tryptofanem nenalezl, a proto o této kombinaci pochybuji. V tom mě utvrzuje i graf 8.10, kde aspartát i glutamát vykazují značnou negativní preferenci k tryptofanu.



Obrázek 8.15: A) Histogram vzdáleností aspartátu a glutamátu (první nalezený atom z postranního řetězce) dané aminokyseliny od těžiště tryptofanu. B) Zelený graf ukazuje, kolikátá v pořadí deseti nejbližších aminokyselin v prostoru byla daná aminokyselina.

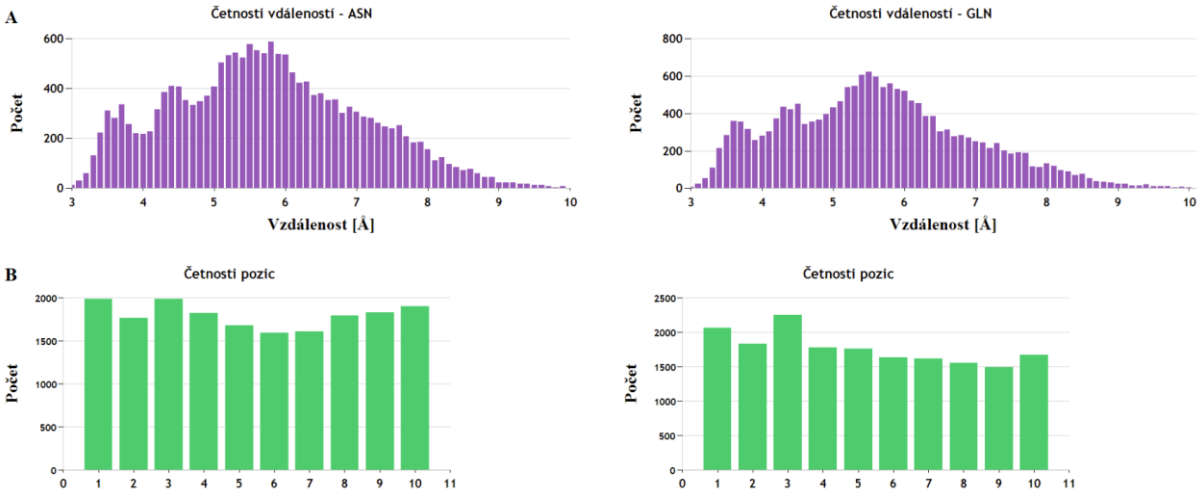


Obrázek 8.16: Prostorové zobrazení nejbližších atomů glutamátu okolo indolu tryptofanu. Zobrazeno je 10 000 glutamátů. Červené kuličky jsou atomy kyslíku a zelený je atom uhlíku. Ukázkové struktury (PDB: 3CL5, 5XWI, 1P49) zobrazují tryptofan (zeleně) a glutamát (žlutě) v konkrétních konfiguracích.

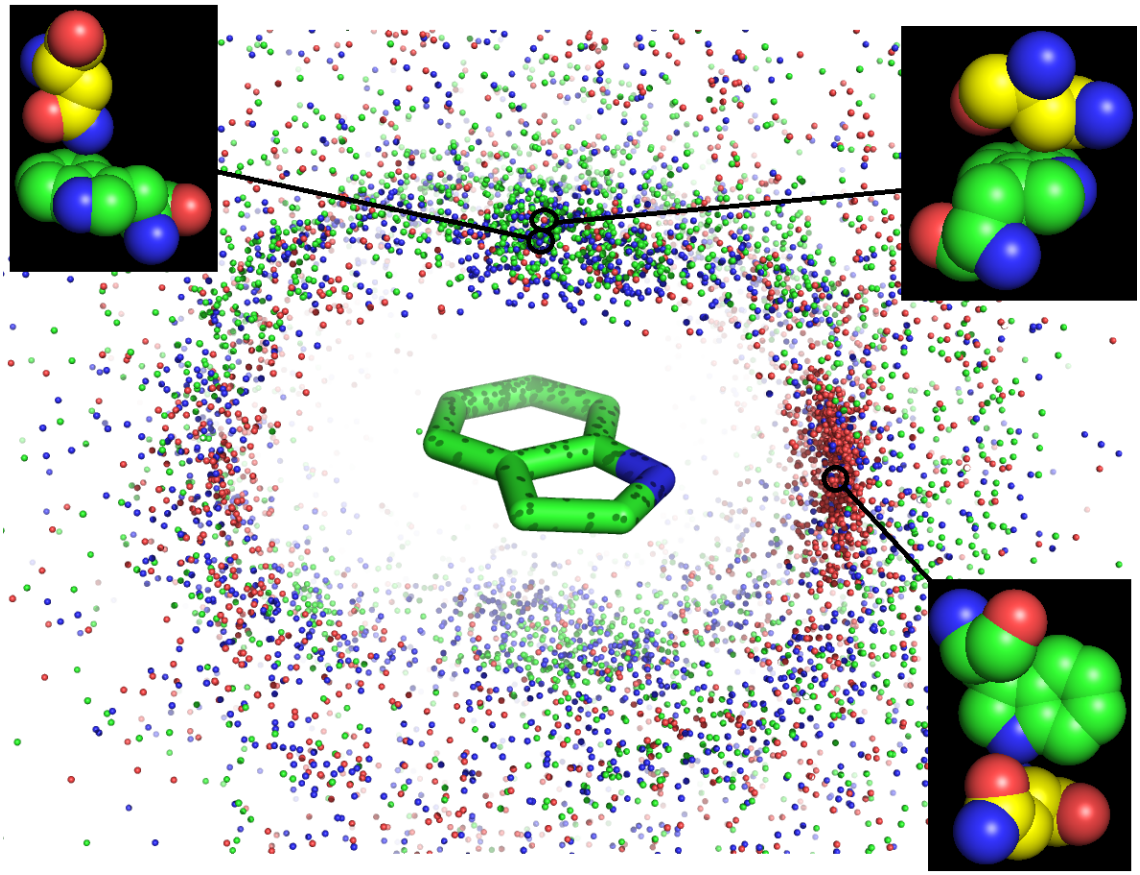
Profily asparaginu a glutaminu na obrázku 8.17 mají zřetelná maxima ve vzdálenostech 3,7 Å a 4,5 Å, která se podobají lysinu (3,7 Å) a aspartátu (4,5 Å). Toto zjištění mě utvrzuje ve správném přístupu k analýze, protože tento profil odráží preference funkčních skupin k jednotlivým částem indolu stejně jako u předchozích aminokyselin. Dusík aminoskupiny v postranních řetězcích asparaginu a glutaminu má tendenci se vyskytovat u těžiště indolu (viz arginin), zatímco kyslík ketonové skupiny zase u dusíku indolové skupiny (viz aspartát). Tyto preference jsou viditelné v prostorovém zobrazení asparaginu na obrázku 8.18.

Je překvapivé, že jsem nenašel žádný článek, který by se těmito dvěma aminokyselinami zabýval v kontextu analýzy PDB databáze. Podle mě jsou zrovna tyto

aminokyseliny velmi zajímavé, jelikož je na nich vidět natáčení postranního řetězce k tryptofanu podle funkčních skupin.



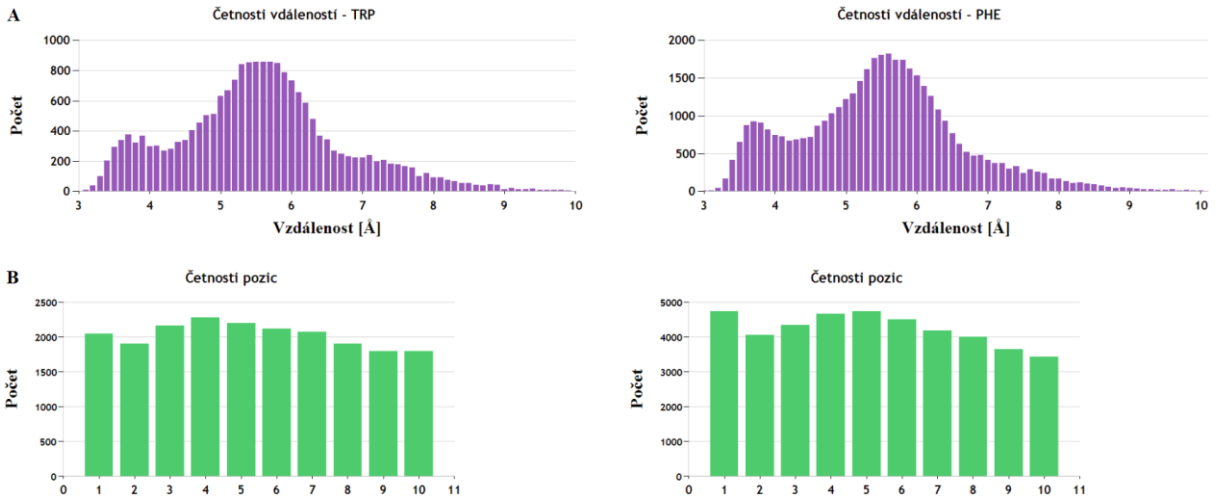
Obrázek 8.17: A) Histogram vzdáleností asparaginu a glutaminu (první nalezený atom z postranního řetězce) dané aminokyseliny od těžiště tryptofanu. B) Zelený graf ukazuje, kolikátá v pořadí deseti nejbližších aminokyselin v prostoru byla daná aminokyselina.



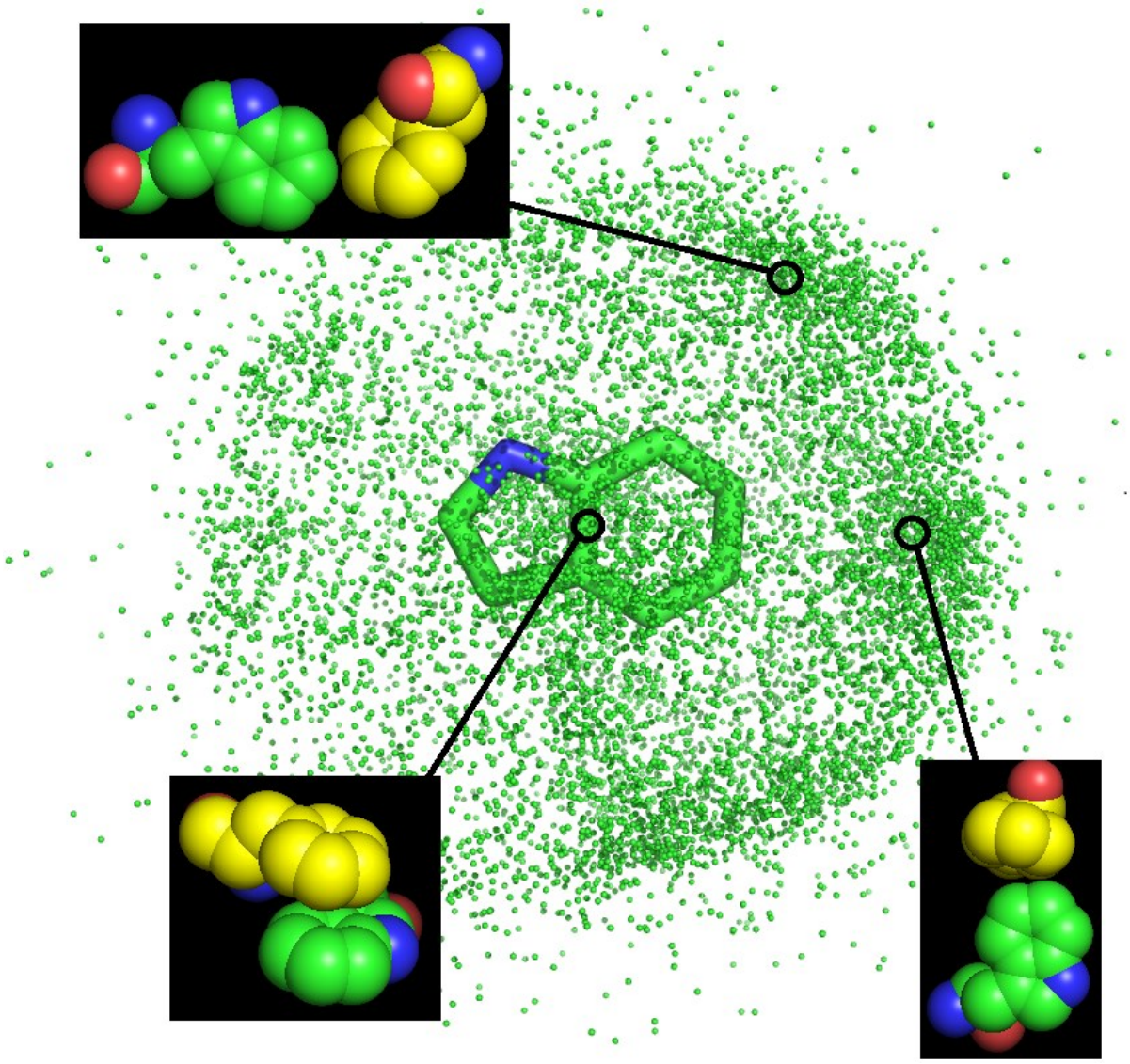
Obrázek 8.18: Prostorové zobrazení nejbližších atomů asparaginu okolo indolu tryptofanu. Zobrazeno je 10 000 asparaginů. Červené kuličky jsou atomy kyslíku, modré jsou atomy dusíku a zelené jsou atomy uhlíku. Ukázkové struktury (PDB: 2CB4, 5Z75, 1UT9) zobrazují tryptofan (zeleně) a asparagin (žlutě) v konkrétních konfiguracích.

Další na řadě jsou aromatické aminokyseliny. K tryptofanu jsem připojil fenylalanin, který by měl mít podobné vlastnosti a na rozdíl od tyrosinu nemá hydroxylovou skupinu. Jejich grafy jsou na obrázku 8.19. Fenylalanin vykazuje navýšení ve vzdálenosti 3,7 Å a na první pozici v grafu pozic. Zajímavější je ale jeho prostorové zobrazení, kde jsou například vidět tendence být spíš u vazeb benzenového cyklu spíše, než u atomů uhlíku. Nabízí se otázka, jestli jsou v těchto místech fenylalaniny na kolmo, a interagují tak s hranou indolu, nebo jsou do těchto míst vytlačeny vodíky, které míří z každého atomu uhlíku indolu do prostoru.

Je evidentní, že tento druh analýzy na aromáty plně nestačí. V kapitole 8.2 se budu věnovat vzájemné pozici a orientaci dvou tryptofanů, kde bude prostor okolo indolu lépe prozkoumán po jednotlivých segmentech.

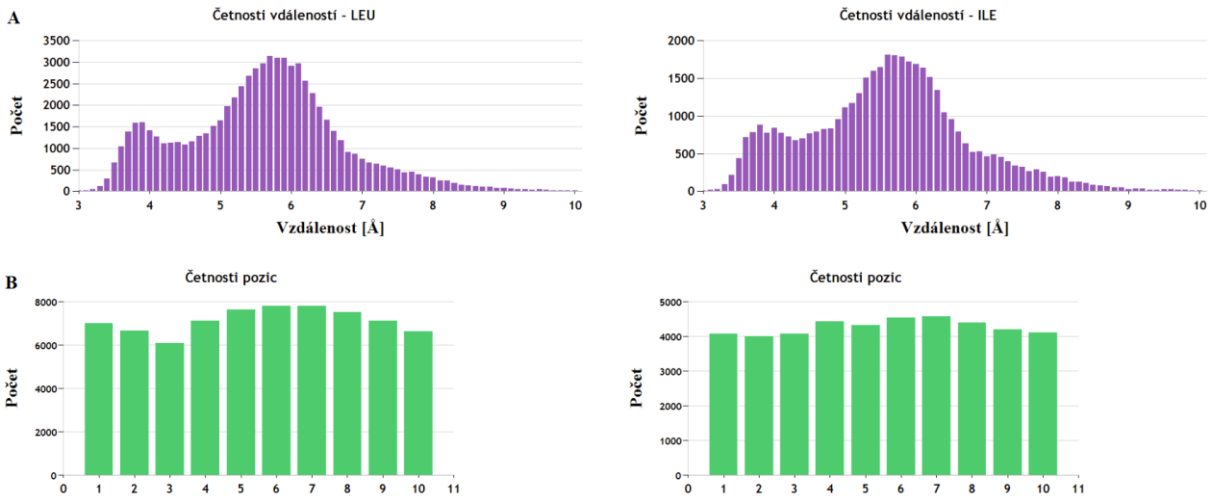


Obrázek 8.19: A) Histogram vzdáleností tryptofanu a fenylalaninu (první nalezený atom z postranního řetězce) dané aminokyseliny od těžiště tryptofanu. B) Zelený graf ukazuje, kolikátá v pořadí deseti nejbližších aminokyselin v prostoru byla daná aminokyselina.

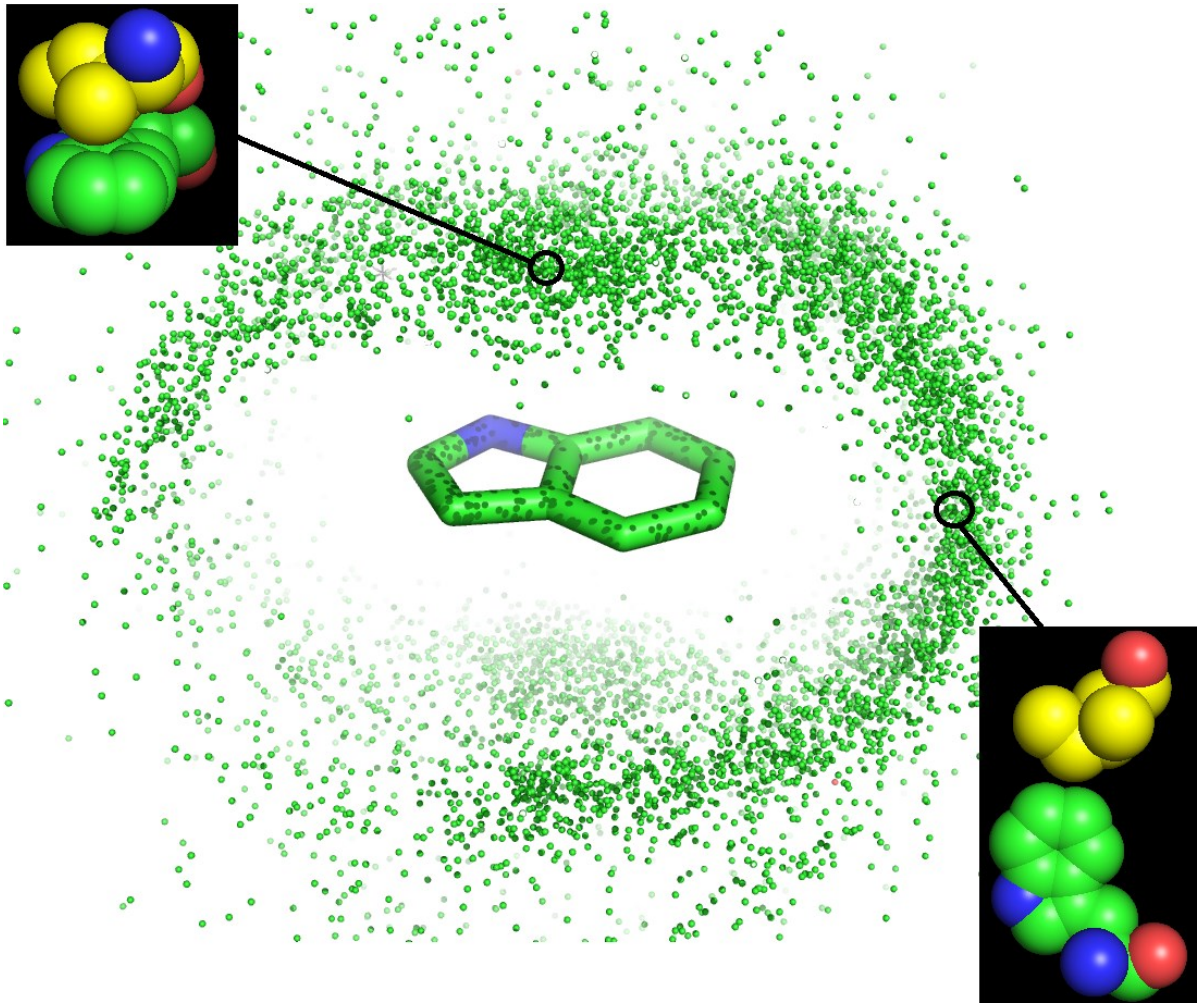


Obrázek 8.20: Prostorové zobrazení nejbližších atomů fenylalaninu okolo indolu tryptofanu. Zobrazeno je 10 000 fenylalaninů. Zelené kuličky jsou atomy uhlíku. Ukázkové struktury (PDB: 1QNM, 5HM4, 5H5Z) zobrazují tryptofan (zeleně) a fenylalanin (žlutě) v konkrétních konfiguracích.

Poslední aminokyseliny, které bych chtěl podrobněji zobrazit, je leucin a isoleucin na obrázku 8.21. Zejména co se týče isoleucinu, který evidentně nemá téměř žádnou poziční preferenci (8.21 (B)). Z tohoto grafu lze usuzovat, že všechny ostatní aminokyseliny, které se liší svým profilem od isoleucinu, vykazují pravděpodobně určitou konkrétní interakci (viz graf 8.12). Vzhledem k tomu, že graf 8.10 ukázal, že leucin i isoleucin jsou obecně u tryptofanu preferovány (což konstatovali i Samanta et al. (2000)), dá se usuzovat, že tyto aminokyseliny interagují s indolem po celém povrchu (například pomocí Van der Waalsovských sil). Určité navýšení v levé části grafu vzdáleností (pod 4,5 Å) je dáno větším rozptylem na třetí a čtvrté pozici (viz obrázek 7.9), což je způsobeno tvarem molekuly indolu.



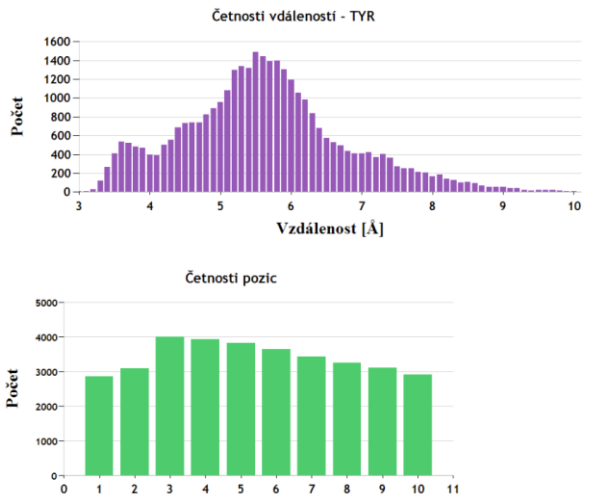
Obrázek 8.21: A) Histogram vzdáleností leucinu a izoleucinu (první nalezený atom z postranního řetězce) dané aminokyseliny od těžiště tryptofanu. B) Zelený graf ukazuje, kolikátá v pořadí deseti nejbližších aminokyselin v prostoru byla daná aminokyselina.



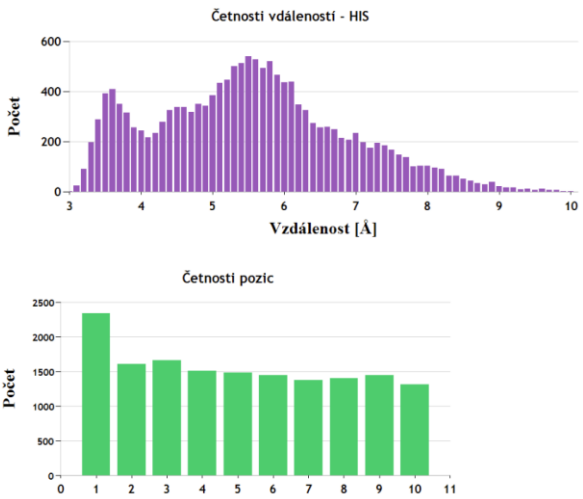
Obrázek 8.22: Prostorové zobrazení nejbližších atomů isoleucinu okolo indolu tryptofanu. Zobrazeno je 10 000 isoleucinů. Zelené kuličky jsou uhlík. Ukázkové struktury (PDB: 3CT5, 4LGJ) zobrazují tryptofan (zeleně) a isoleucin (žlutě) v konkrétních konfiguracích.

Bohužel popisování jednotlivých aminokyselin zabírá značný prostor, proto musím u zbylých aminokyselin zobrazit pouze histogramy (obrázek 8.22).

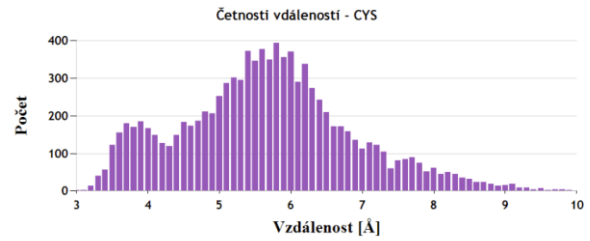
A



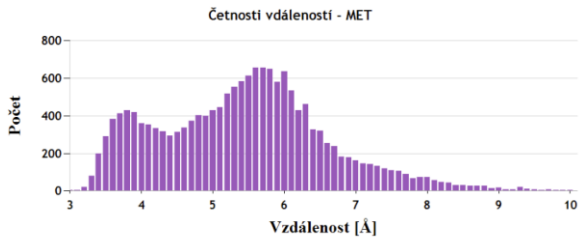
B

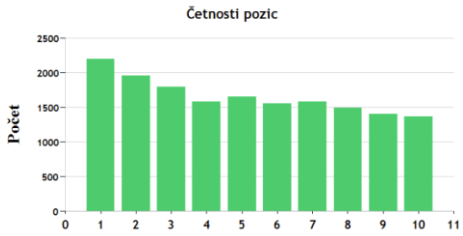


C

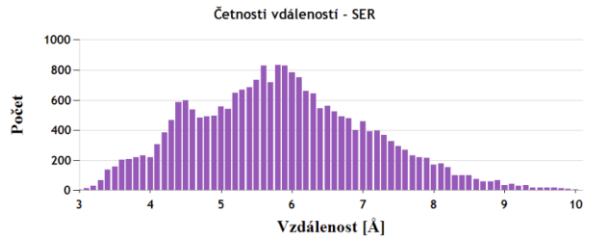


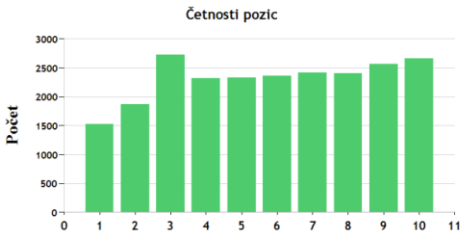
D



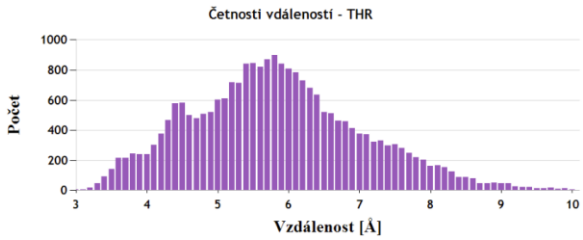


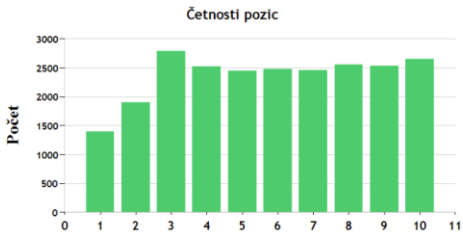
E



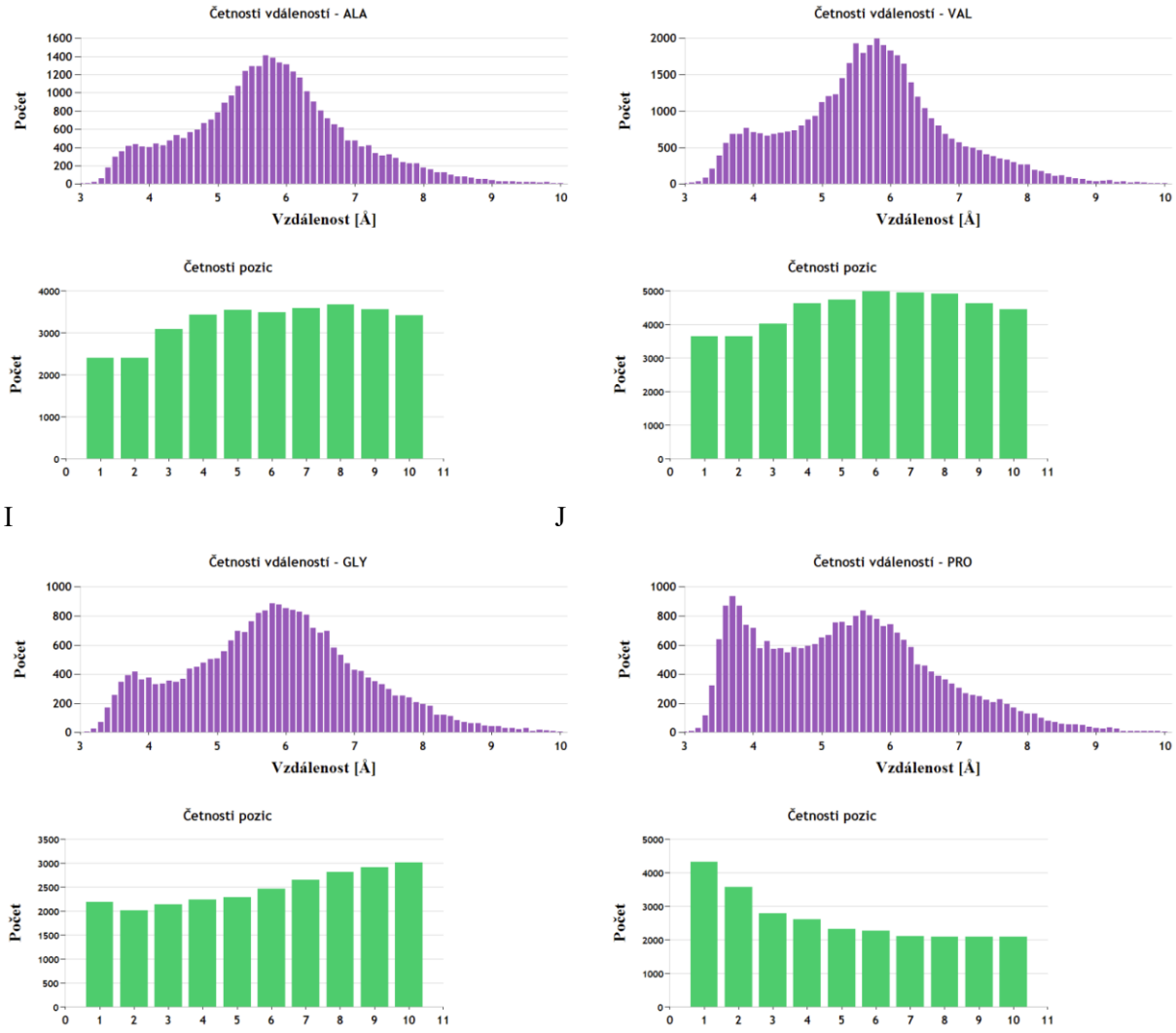


F





G H



I J

Obrázek 8.22: Histogramy vzdáleností (první nalezený atom z postranního řetězce) zbývajících aminokyselin (viz nadpisy na grafy četností) od těžiště tryptofanu. B) Zelený graf ukazuje, kolikátá v pořadí deseti nejbližších aminokyselin v prostoru byla daná aminokyselina.

Z ostatních aminokyselin je patrné, že fyzikálně podobné aminokyseliny mají podobné histogramové profily. Například serin a threonin mají podobné navýšení ve vzdálenosti 4,5 Å (obrázek 8.22 E a F), stejně jako aspartát (obrázek 8.15). Pravděpodobně je to způsobeno jejich hydroxylovou skupinou. Dále je zajímavý prolin (obrázek 8.22 (J)) s preferencí k těžišti tryptofanu, což pravděpodobně odpovídá vlastnostem popsaným v kapitole 6.4.5. Tato preference je zřejmě způsobená CH-π interakcí, kde parciální pozitivně nabitý vodík prolinu interaguje s π oblastí aromatické aminokyseliny (Brandl et al., 2001; Kumar a Balaji, 2014; Umezawa et al., 1999). Dále u cysteinu a methioninu jsou například překvapivě vidět opačné trendy pozic, kde se jedná o sulfur-π interakci (viz kapitola 6.4.6).

V této kapitole jsem chtěl předvést, že pouze exaktní podrobná analýza může ukázat převažující trendy a preference, které se jinak mohou ztratit v nadbytku „náhodných“ konfigurací, které jsou ve skutečnosti vynuceny dalšími aminokyselinami mino prováděnou analýzu. Dále, že by se nemělo v analýze spoléhat pouze na jeden výsledný graf, ale teprve kombinace různých grafů a zobrazení může dát povědomí o studované problematice.

U jednotlivých aminokyselin by se jistě dalo postupovat daleko podrobněji, ale kvůli omezenému prostoru diplomové práce jsem nemohl mnohé zpracovat. Jedná se například o porovnání prostoru u jednotlivých ploch a hrany indolu, kde by bylo možné porovnat četnosti s narůstající vzdáleností od daného indolu. Prozkoumat profily četností okolo indolu, jak se nabízí u prostorového zobrazení fenylalaninu (obrázek 8.20). Zobrazovat více, než jen nejbližší atom postranního řetězce. A možná jako nejdůležitější definovat preferované orientace jednotlivých aminokyselin.

8.2 Dvojice tryptofanů

8.2.1 Úvod

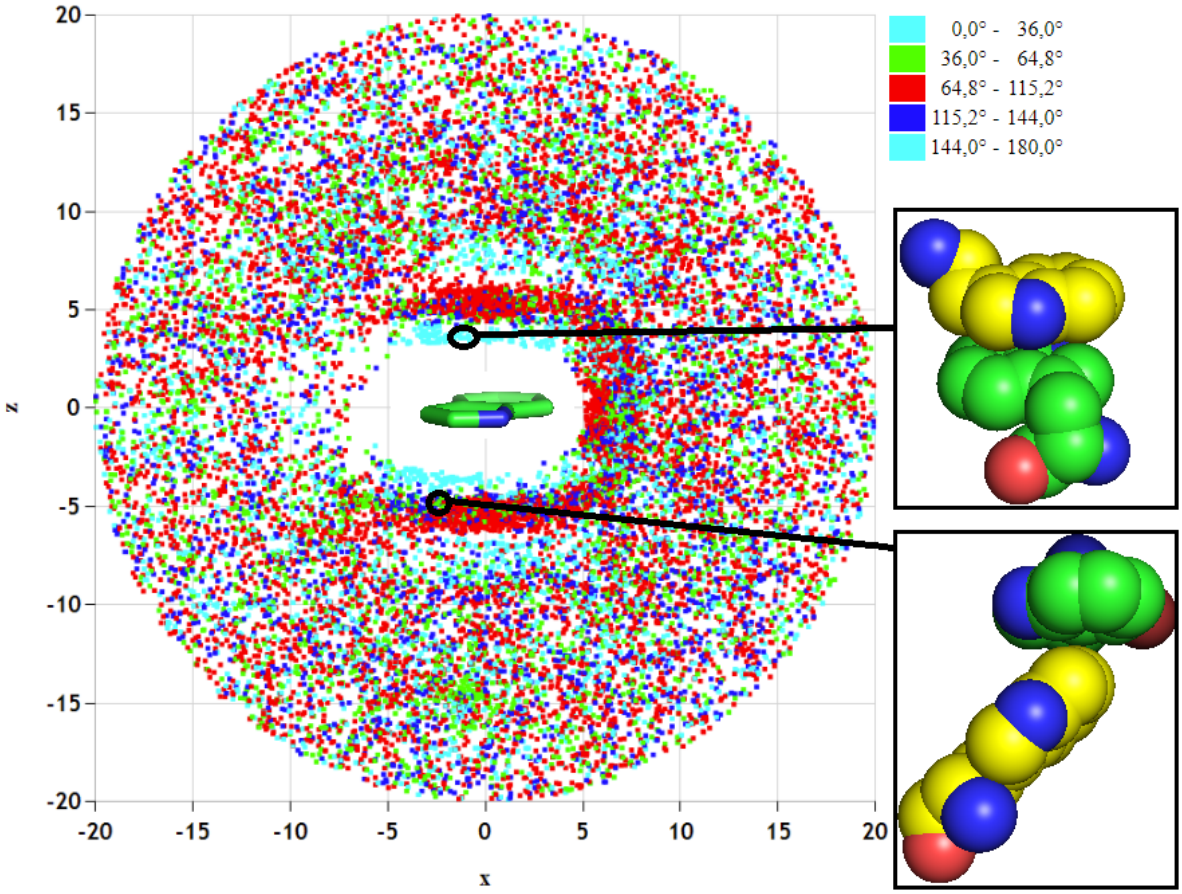
Jak už jsem uvedl v předchozí kapitole, u jednotlivých aminokyselin by se měla vypracovat podrobnější analýza. V této kapitole se proto pokusím zaměřit alespoň na jednu z nich. Vzhledem k tomu, že v grafu 8.10 se ukázala jako nejpreferovanější aminokyselina k tryptofanu jiný tryptofan, soustředím se nyní právě na tuto aminokyselinu. Jednotlivé páry tryptofanů byly získány postupem popsaným v kapitole 7.6.2 a mezi těmito páry byly vypočítány potřebné údaje. Především se jedná o vzájemnou vzdálenost mezi těžišti indolů a výpočet úhlu mezi jejich normálami, tak jak je popsáno v kapitole 7.4.

8.2.2 Analýza

Pro samotnou analýzu jsem disponoval 558 962 páry tryptofanů. Každý pár se v tomto datasetu vyskytuje 2krát, aby se na daný pár v analýzách nahlíželo nezávisle z pohledu každého zúčastněného indolu.

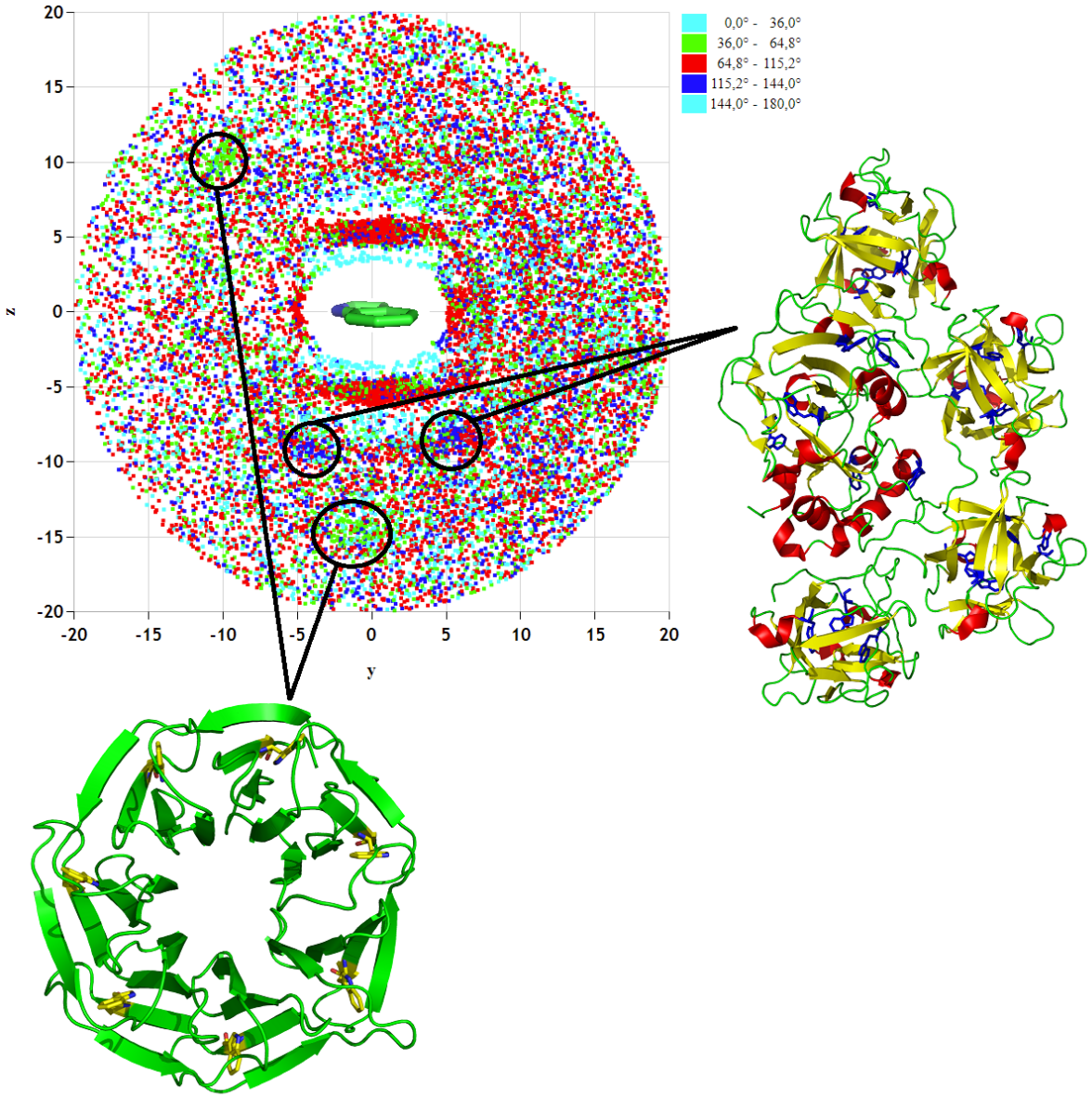
Již v předešlé kapitole se různá prostorová zobrazení ukázala jako velmi názorná, proto nyní budu postupovat stejně. U každého páru jsem indolovou skupinu prvního tryptofanu (dále označovaný jako centrální) umístil do počátku souřadnicového systému tak, jak je znázorněno na obrázku 7.4 v kapitole 7.4. Těžiště druhého indolu jsem přesunul v prostoru s ohledem na první tryptofan. Z důvodu ovlivnění pozic u sekvenčně blízkých aminokyselin (viz kapitola 8.1.2) jsem do analýzy použil pouze páry se sekvenční vzdáleností od jedenácti aminokyselin a jejich vzdálenost těžišť jsem omezil do 20 Å. Vzhledem k tomu, že by zobrazení takto získaných dat vytvořilo neprůhlednou kouli, budu prostor zobrazovat v podobě řezů podél jednotlivých os. Jedná se vždy o vrstvu prostoru v tloušťce 3 Å okolo počátku, tedy s pozicemi těžišť v mezích od -1,5 Å do 1,5 Å. Jednotlivé body (těžiště druhého indolu) jsou obarveny podle úhlů normál. Normály okolo pravého úhlu jsou zobrazeny červeně, paralelní orientace světle modře a úhly mezi zeleně a tmavě modře.

První obrázek 8.23 zobrazuje řez podél osy y. Na první pohled jsou zde patrné vrstvy různých úhlů normál nad a pod plochou indolu. Nejblíže jsou „stacking“ (paralelní) orientace (světle modrá barva), protože jen tak mohou být indoly z definice těžiště nejblíže u sebe. Následně se začínají indoly natáčet (zelená a modrá barva) až přecházejí do pravého úhlu (červená barva). Kolmý úhel je pak v určité vzdálenosti (okolo 5 Å) naprosto dominantní. Na hraně indolu je tomu naopak, nejblíže musí být z definice kolmá orientace (červená barva). Za pozornost stojí i jasný úbytek bodů na místě, kde se vyskytuje Cβ tryptofanu, a tudíž peptidová vazba (vlevo od indolu).



Obrázek 8.23: Prostorové zobrazení tryptofanových párů. Indol prvního tryptofanu je umístěn v počátku souřadnicového systému a z indolu druhého tryptofanu je zobrazeno umístění jeho těžiště. Barva těžiště odpovídá úhlu mezi jejich normálami. Zobrazena je vrstva prostoru podél osy y v tloušťce 3 Å okolo počátku, tedy s pozicemi těžišť v mezích od -1,5 Å do 1,5 Å. Dvě různé ukázkové struktury (PDB: 3WFO, 4GT4) zobrazují první centrální tryptofan (zeleně) a druhý tryptofan (žlutě) v konkrétních konfiguracích.

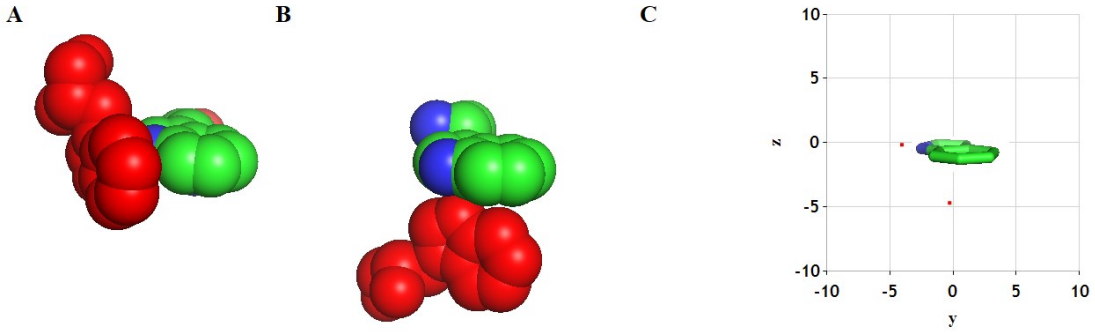
Na obrázku 8.24, na kterém je zobrazený řez podél osy x, je situace obdobná. Navíc jsou zde vidět modré a zelené shluky často se vyskytujících podobných konfigurací, ale vzhledem ke značné vzdálenosti od indolu se pravděpodobně nejedná o reálnou interakci.



Obrázek 8.24: Prostorové zobrazení tryptofanových párů. Indol prvního tryptofanu je umístěn v počátku souřadnicového systému a z indolu druhého tryptofanu je zobrazeno umístění jeho těžiště. Barva těžiště odpovídá úhlu mezi jejich normálami. Zobrazena je vrstva prostoru podél osy x v tloušťce 3 Å okolo počátku, tedy s pozicemi těžišť v mezích od -1,5 Å do 1,5 Å. Označené barevné shluky odpovídají strukturně podobným proteinům (podrobné vysvětlení viz text). Konkrétní příklad struktury z modrého shluku je PDB 2VSE a zeleného PDB 5MZH.

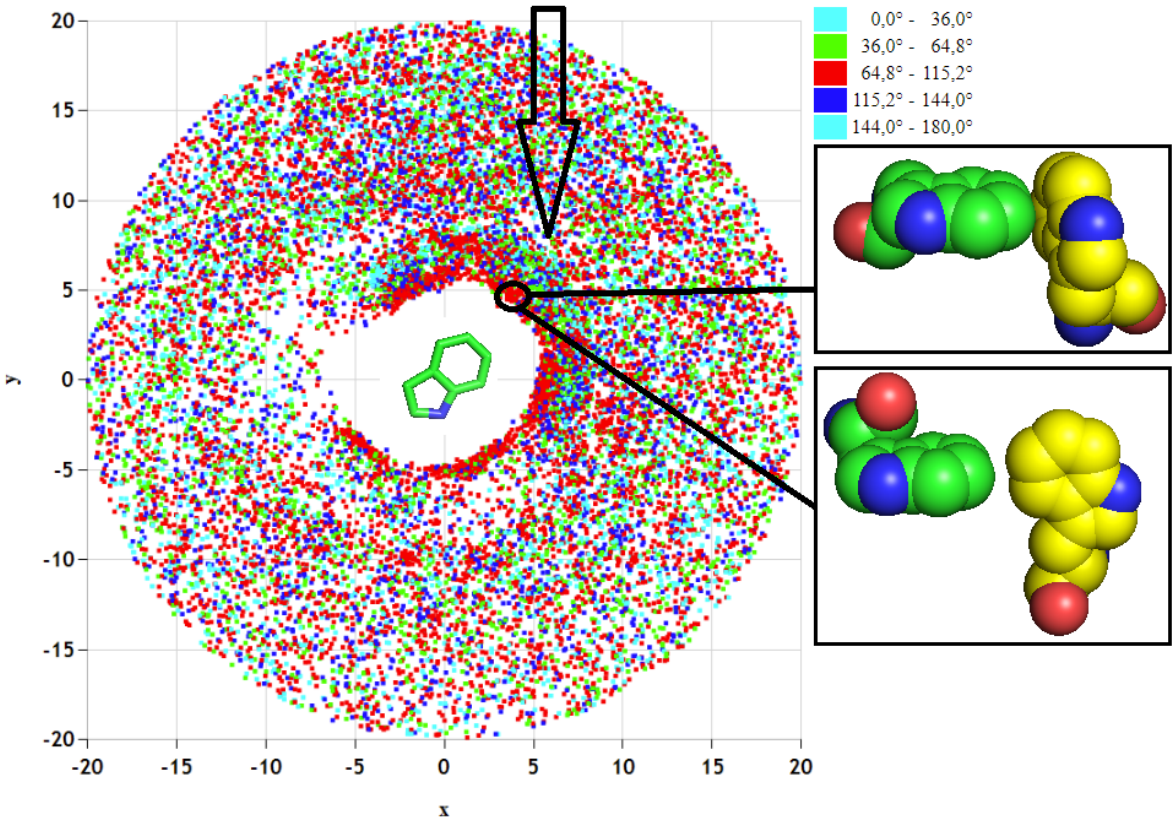
Takový shluk by se mohl vytvořit, pokud by ve studovaném datasetu přetrvávala sekvenční homologie, což by mohlo značně ovlivňovat analýzy. Po bližším přezkoumání se ukázalo,

že se jedná o strukturně podobné proteiny. Zelené shluky se skládají z nehomologních spirálovitých proteinů se symetricky umístěnými tryptofany. U modrých shluků se jedná o nehomologní proteiny složené z menších domén, kdy podobné domény obsahují téměř stejné počty tryptofanů, a tím vytvářejí tato seskupení. Zatím jsem neobjevil způsob, jak podobné struktury z analýzy eliminovat. Každý arteficiální shluk je na obrázku 8.24 duplikován v různých místech. Tato duplikace je z důvodu dvojího zobrazení každého páru. To znamená, že se jedná o stejný shluk, jen pokaždé z pohledu jiného z obou tryptofanů v páru. Zdvojení ukazuje na důležitou vlastnost tohoto prostorového zobrazení, jelikož například relevantní červený oblak na okrajích indolu (kolmá orientace) se z části rovná červenému oblaku nad a pod indolem. Jsou to totiž opět stejné na sebe kolmé páry indolů z různých pohledů. Vysvětlení je uvedeno na obrázku 8.25, pro různá zobrazení stejného páru tryptofanů. Takový úmyslný duplikovaný výskyt páru zvyšuje hustotu výskytu a analýza je proto snazší. Žádná z konfigurací však tímto postupem není uměle obohacována.



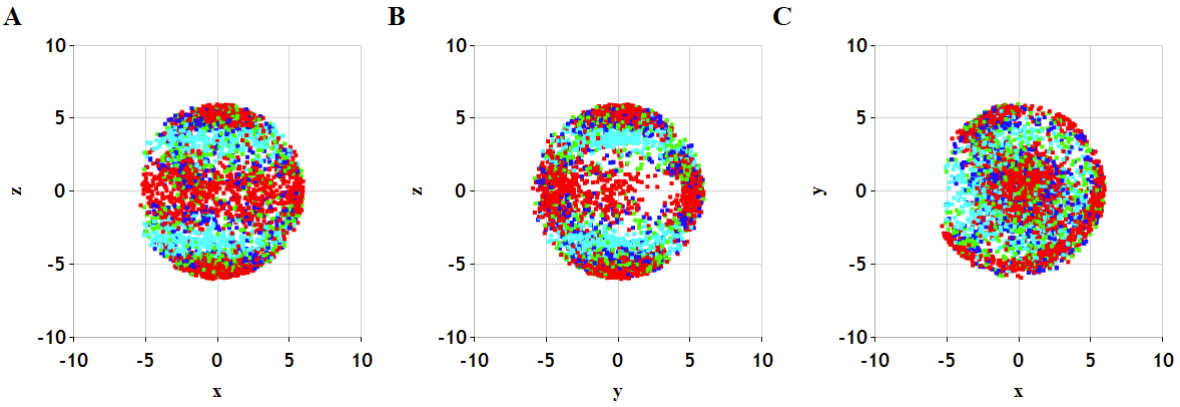
Obrázek 8.25: Dvě různá zobrazení stejného páru tryptofanů (tryptofan 70 a 86 z PDB: 3V10, řetězec A). (A) První tryptofan je z pozice 70 (zelený) a k němu druhý z pozice 86 (červený). (B) Druhý pár má zobrazeny stejné tryptofany, ale obráceně orientovány a přebarveny. (C) Pozice obou párů v prostorovém zobrazení. Zelený tryptofan v počátku souřadnicového systému odpovídá zeleným tryptofanům z obrázků A a B, tudíž „prvním“ tryptofanům.

Na obrázku 8.26 je znázorněn řez podél osy z. Na tomto zobrazení je jasně patrná tendence indolů okolo hrany indolu zaujímat pravý úhel. Dále je opět vidět pochopitelně snížený výskyt bodů na místě peptidové vazby. V tomto řezu je zajímavá vrstva sníženého výskytu ve vzdálenosti 9 Å od těžiště indolu (šipka). Rád bych upozornil, že takováto prostorová zobrazení jsem v žádné práci nenalezl a myslím, že docela dobře vypovídají o vzájemné orientaci tryptofanů. Nicméně informace je stále nekompletní, jelikož například není poznat, jak přesně jsou vůči sobě indoly natočené. Například „kolmost" totiž může být zcela různá, jak je vidět na příkladech konkrétních struktur (8.26 vpravo). Za zmínku stojí, že preferovaný výskyt tryptofanu v okolí jiného tryptofanu se vcelku podobá výskytu fenylalaninu v okolí tryptofanu (viz obrázek 8.20), kde je podobně jako zde patrná značná četnost molekul okolo šestičetného cyklu indolu.

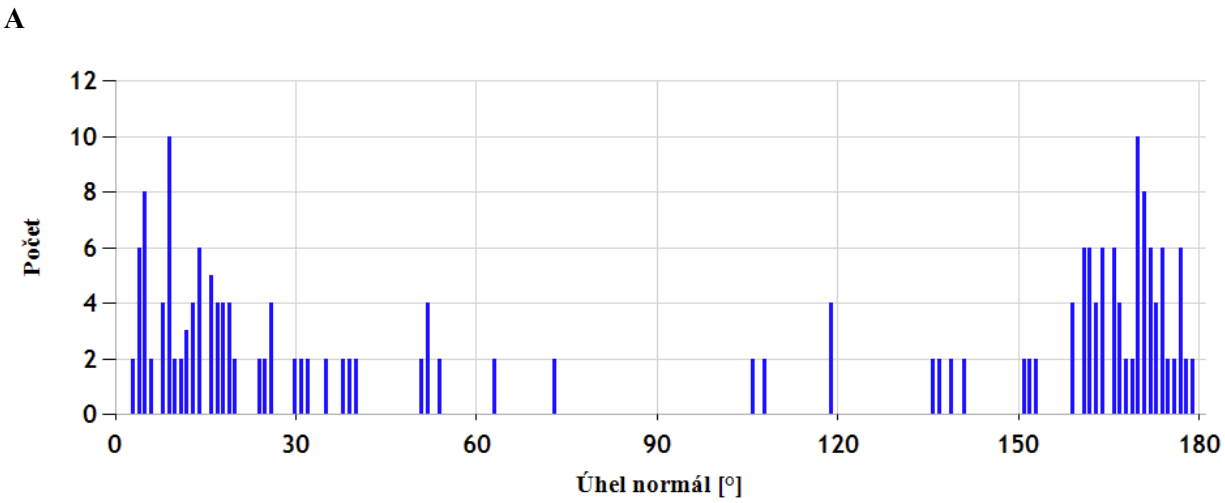


Obrázek 8.26: Prostorové zobrazení tryptofanových párů. Indol prvního tryptofanu je umístěn v počátku souřadnicového systému a z indolu druhého tryptofanu je zobrazeno umístění jeho těžiště. Barva těžiště odpovídá úhlu mezi jejich normálami. Zobrazena je vrstva prostoru podél osy z v tloušťce 3 Å okolo počátku, tedy s pozicemi těžišť v mezích od -1,5 Å do 1,5 Å. Dvě různé ukázkové struktury (PDB: 4P6B, 1RH6) zobrazují první centrální tryptofan (zeleně) a druhý tryptofan (žlutě) v kolmém úhlu normál, ale jinak ve zcela různé orientaci. Šipka poukazuje na určitý nevysvětlený pokles hustoty bodů.

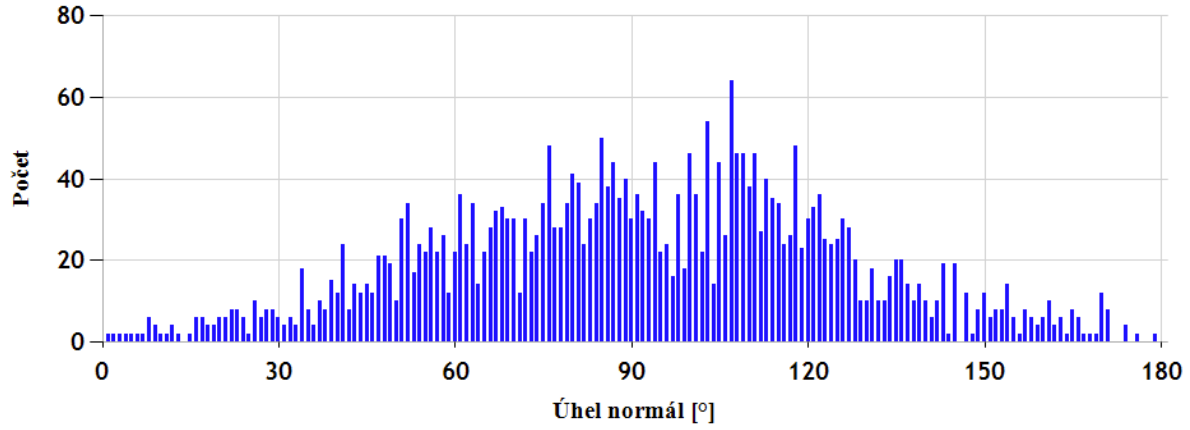
Z těchto prostorových zobrazení by se dalo usuzovat, že orientace do určité malé vzdálenosti jsou ovlivněné fyzikálně (nejbližší indol k ploše jiného indolu musí být paralelní, jinak by nebyl nejbližší). Jedná se zde tedy o zcela jasná předvídatelná pravidla. Orientace jsou dále odstupňovány po určitých vrstvách, jak je vidět na zobrazení prostorové koule výskytu do 6 Å na obrázku 8.27. Nejedná se tedy o řez prostoru, ale o zobrazení celého vymezeného objemu. Pro určité vrstvy vzdáleností od těžiště jsem navíc na obrázku 8.28 zobrazil histogramy úhlů normál mezi indoly. Blízká vrstva 0-4,5 Å obsahuje především paralelní úhly. O něco vzdálenější vrstva 5-6 Å již obsahuje pootočené indoly a vzdálená vrstva 6-8 Å začíná vykazovat určitý zakulacující se profil s tendencí k pravému úhlu.



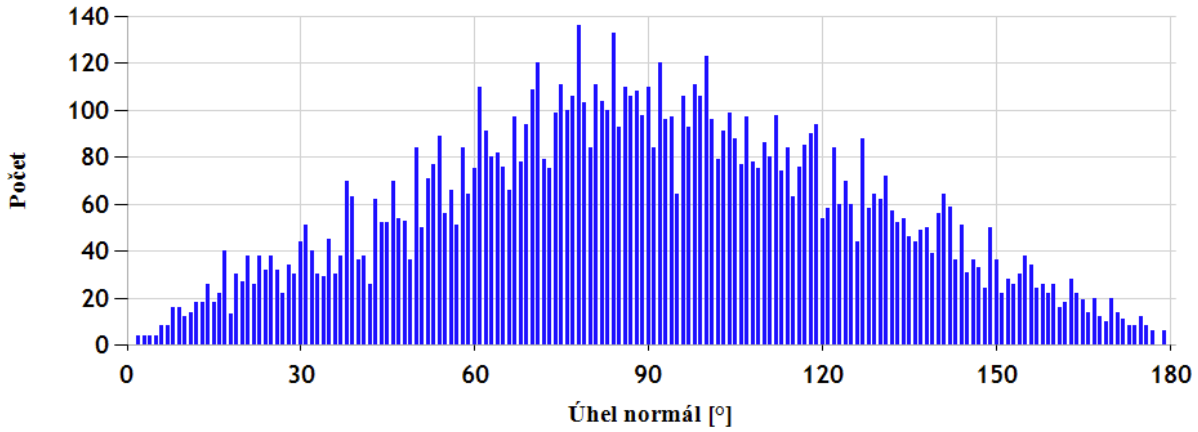
Obrázek 8.27: Rozložení těžišť indolů v prostoru okolo centrálního indolu. Vzdálenost těžišť mezi indoly je omezena do 6 Å. Pohledy jsou z různých směrů. (A) Pohled z osy y, viz obrázek 8.23 (B) Pohled z osy x, viz obrázek 8.24 (C) Pohled z osy z, viz obrázek 8.26. Barva těžiště odpovídá úhlu mezi normálami indolů, viz obrázek 8.26.



B

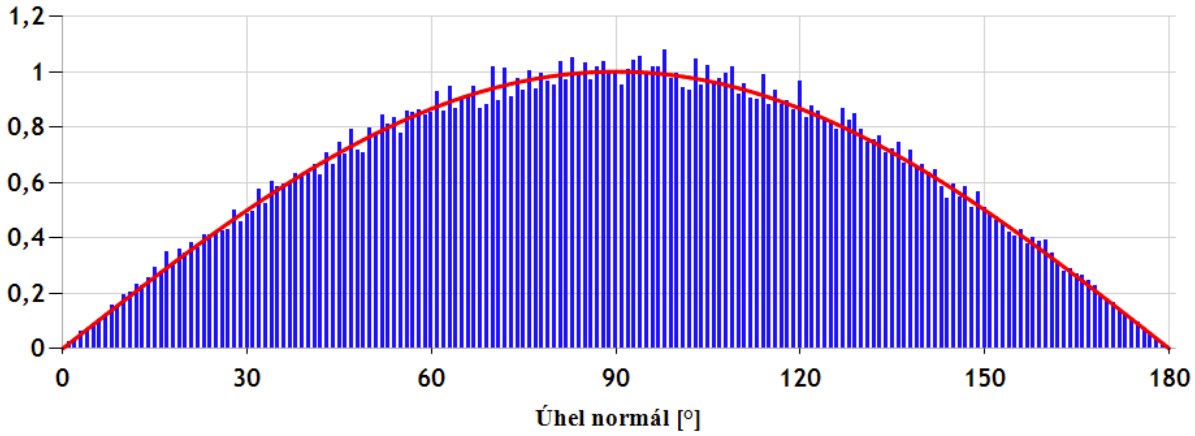


C



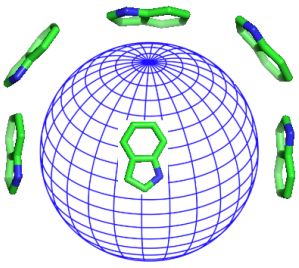
Obrázek 8.28: Histogramy úhlů mezi normálami dvou indolů. Jednotlivé grafy se liší intervalem vzdáleností mezi těžišti. (A) Vzdálenost 0-4,5 Å, 212 párů. (B) Vzdálenost 5-6 Å, 3 199 párů (C) Vzdálenost 6-8 Å, 10 026 párů.

Tendence vzdálenějších vrstev k pravému úhlu může souviset s pravděpodobností, s jakou bychom danou orientaci nalézali v případě velkého množství náhodně rozmístěných aminokyselin, jak je popsáno v kapitole 6.4.1. Odpověď může dát histogram na obrázku 8.29, který zobrazuje úhly normál pro páry tryptofanů, které jsou od sebe od 15 Å do 40 Å. V této vzdálenosti spolu již nejspíš nemají žádnou interakci, a přesto graf má velmi specifický profil. Jedná se totiž o funkci sinus, což dokazuje křivka na stejném obrázku a svědčí o naprosté náhodnosti pozorovaných orientací.



Obrázek 8.29: Normalizovaný histogram úhlů mezi normálami dvou indolů (modře). Zobrazeny jen páry, které mají vzdálenost mezi těžišti indolů 15-40 Å. Tento histogram je možné proložit funkcí sinus (červeně), což značí náhodné rozdělení pozorovaných úhlů. Graf je tvořen 342 349 páry tryptofanů.

Proč náhodné rozdělení dává zrovna funkci sinus? Když si představíme různé úhlové orientace dvou tryptofanů v prostoru a z toho jeden (centrální) by byl normálou svisle a druhý (zkoumaný) libovolně, tak pravděpodobnost s jakou nalezneme určitou orientaci, je úměrná ploše „čtverce“ na glóbu (viz obrázek 8.30). Prakticky platí, že pro úhly blízké 90° (normála zkoumaného tryptofanu je vodorovně) je čtverec „větší“, než čtverec u pólu (při 0° nebo 180°). Velikost oněch zkoumaných čtverců závisí právě na úhlu podle funkce sinus. Zjednodušeně řečeno, pro kolmé uspořádání dvou molekul (měřeno pomocí jakýchkoli jejich vybraných os) je obrovské množství různých možností. Naopak, při paralelním (nebo antiparalelním) uspořádání je vzájemná orientace molekul mnohem přesněji vymezena.



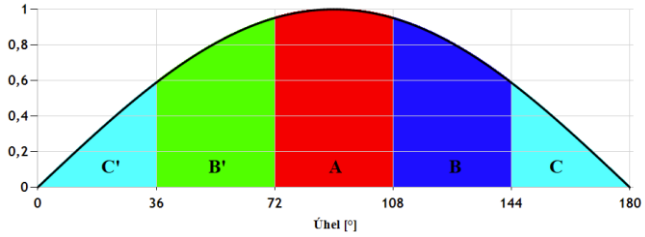
Obrázek 8.30: Zobrazení pravděpodobnosti orientace dvou náhodně orientovaných indolů. Centrální indol je ukryt ve středu glóbu a orientován normálou svisle. Hodnota úhlu normál bude nejčastěji 90°, což odpovídá větší dostupné ploše glóbu podél rovníku. Nejvzácnější bude úhel 0° a 180°, znázorněno indolem na pólu.

Výše zobrazené obrázky a grafy již vcelku vypovídají o rozložení indolů v prostoru. Nyní se zaměřím na konkrétní otázky, které bych chtěl v této kapitole zodpovědět.

Jak už jsem zmínil při popisu obrázku 8.26, ve vzdálenosti 9 Å je vidět zřetelný pokles hustoty bodů. Mohl by to být přechod, kde indoly o sobě přestávají „vědět“, a tudíž od této hranice dále je orientace náhodná. Nyní bych se tedy pokusil odpovědět na otázku, do jaké vzdálenosti se tryptofany ještě ovlivňují? To znamená, do jaké vzdálenosti ještě vykazují „nenáhodnou“ orientaci. Zaměřím se nyní na úhlovou orientaci indolů, protože samotný výskyt tryptofanu v určité vzdálenosti je podmíněn například velikostí proteinů. Jak už jsem ukázal na grafu 8.29, histogram náhodného rozdělení úhlů v trojrozměrném prostoru tvoří funkci sinus. S touto funkcí bych tedy rád porovnal četnosti úhlů v určitých kulovitých vrstvách od těžiště indolu. Jedna z možností je pro každou kulovitou vrstvu vykreslit histogram a zjišťovat, jestli se od funkce sinus liší. To ale není moc praktické, a proto jsem přistoupil k porovnávání určitých intervalových úseků.

Pro každou vzdálenostní vrstvu od centrálního indolu (po 0,2 Å) jsou vypočítány histogramy úhlů normál (stejně jako histogramy na obrázku 8.28). Četnosti každého histogramu jsou následně rozděleny do pěti intervalů úhlů (C’, B’, A, B a C) tak, jak je zobrazeno pro sinus na obrázku 8.31 (A). Výsledná hodnota porovnání je „náhodnost úhlů“ Δ, která je dána výrazem 8.31 (B). Pokud byl pro konkrétní data jmenovatel větší než čitatel, dojde ke stejné transformaci jako je na obrázku 8.3. Výsledkem je tedy konkrétní výsledná transformovaná hodnota náhodnosti Δ‘. Konstanta 1,6178 ve vzorci 8.31 (B) vyrovnává pokles funkce sinus (jedná se o podíl integrálů zkoumaných intervalů funkce sinus), aby při náhodném rozdělení vyšla hodnota 0. Pokud platí Δ‘ > 0, udává hodnota |Δ‘|+1, kolikrát je v dané vrstvě *kolmé* uspořádání častější, než by odpovídalo náhodě. Obdobně při Δ‘<0 udává hodnota |Δ‘|+1, kolikrát je v dané vrstvě *paralelní* uspořádání častější, než by odpovídalo náhodě. Bohužel tento postup nezachytí některé specifické odchylky od funkce sinus (uvnitř intervalů), ale jen odchylky v daných intervalech jako celcích. Například v určitých vzdálenostech jsou dominantní úhly 80° a 100° a určitý pokles u 90° (viz obrázek 8.28 (B)). Takovéto detailní rozdíly popsaný postup nezaznamená, ale určitou představu poskytne.

A

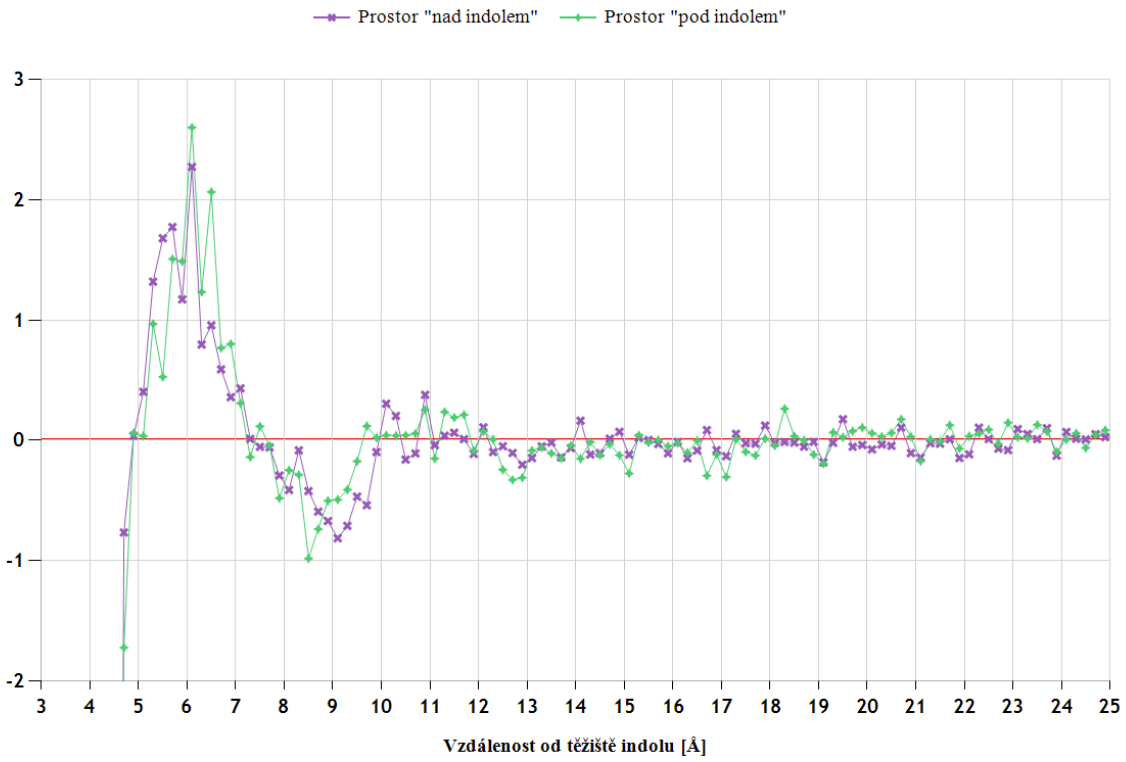


B

$$\Delta = \left(\frac{A}{(C' + C) * 1{,}6178}\right) - 1$$

Obrázek 8.31: (A) Rozdělení funkce na intervaly, přes které jsou porovnávány četnosti úhlů normál pro dané vzdálenostní vrstvy. (B) Vzorec udávající rozdíl oproti funkci sinus.

Výše popsaným postupem jsem vytvořil graf 8.32, který zobrazuje porovnání četností úhlů normál s funkcí sinus pro každou kulovitou vrstvu po 0,2 Å od těžiště centrálního indolu. Zdrojem této analýzy je 253 134 párů tryptofanů, které jsou rozděleny na dvě části podle plochy centrálního indolu (nad a pod). Toto rozdělení je z důvodu verifikace analýzy, aby bylo patrné, že výsledný profil není pouhý šum. První bod na ose x začíná ve vzdálenosti 4,6 Å, jelikož menší vzdálenosti dosahovaly k hodnotě -20, což narušovalo detail zobrazení. Graf tedy ukazuje, že pod 5 Å jsou orientace nenáhodné, konktrétně paralelní nebo antiparalelní. Od 5-7,2 Å orientace indolů tíhne ke kolmému uspořádání a od 7,6-10 Å opět k paralelní či antiparalelní orientaci.

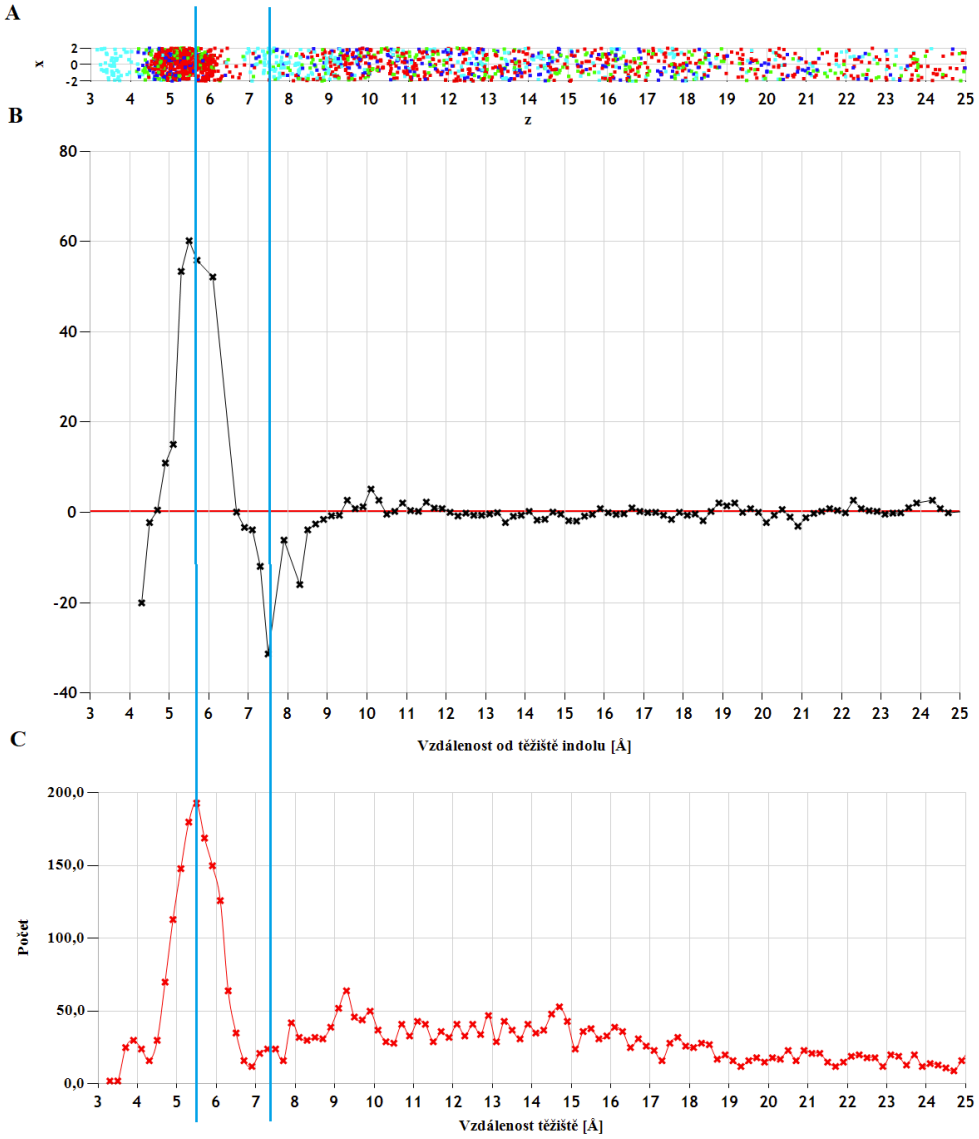


Obrázek 8.32: Graf „náhodnosti“ rozdělení úhlů normál Δ‘ v celém prostoru kolem indolu (po 0,2 Å, do 25 Å). Náhodnost Δ‘ je porovnávána s funkcí sinus (viz obrázek 8.31 a text). Záporné hodnoty odkazují na paralelní orientaci, zatímco kladné hodnoty ukazují na preferenci ke kolmým úhlům. Šum okolo nulové hodnoty značí náhodné rozložení úhlů. 253 134 párů tryptofanů je rozděleno na dvě poloviny podle plochy indolu.

Snaha o současnou analýzu všech vzájemných pozic indolů v určité vzdálenosti může být zavádějící, jelikož z prostorového zobrazení (viz obrázek 8.24) je patrné, že konfigurace indolů v kouli okolo centrálního indolu jsou naprosto nesrovnatelné. Je to dáno tím, že různé směry od středu indolu mají jiné preferované orientace. Autoři, vytvářející prosté histogramy chybují ve slučování prostoru (např. nad a pod plochou indolu a zároveň po jeho hranách), který je ale značně různorodý (McGaughey et al., 1998; Ninkovic et al., 2014; Thomas et al.,

2002). Z tohoto důvodu jsem zkoumané pozice omezil na kvádr 4×4 Å nad a pod plochou centrálního indolu, což dalo 3 947 tryptofanů. Výsledek je na obrázku 8.33.

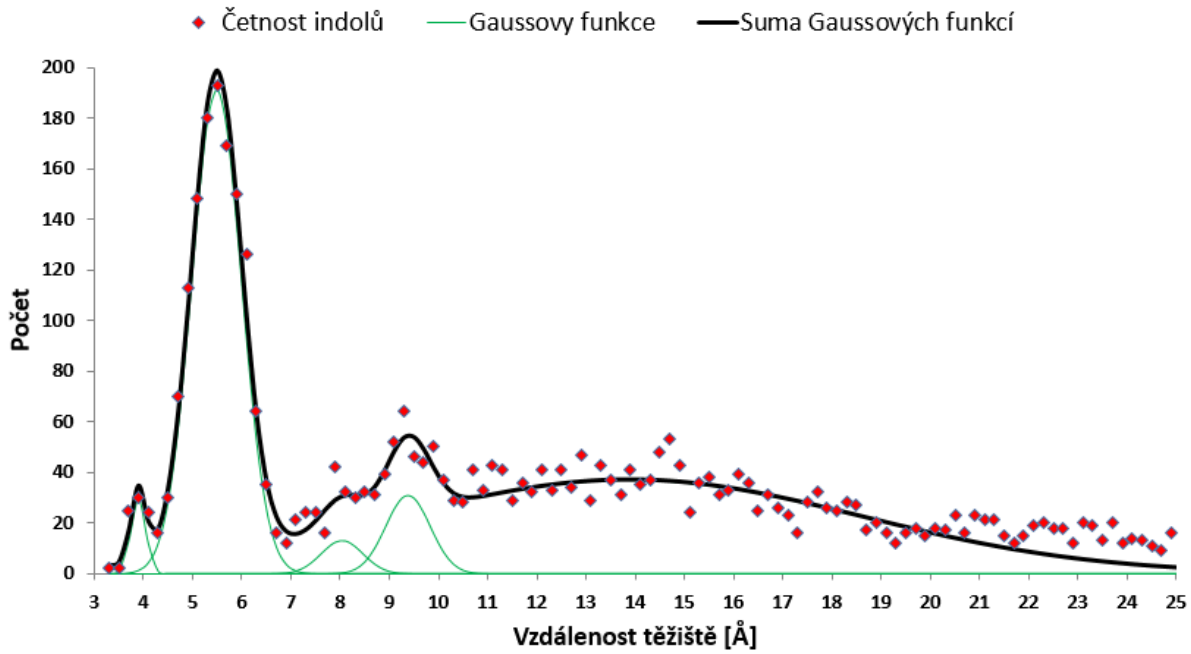
Obrázek 8.33 se skládá ze tří vzájemně porovnatelných částí. Obrázek 8.33 (A) zobrazuje výsek studovaného prostoru z obrázku 8.24 a jedná se o kvádr 4×4 Å. Pro jednoduchost je zobrazen pouze výsek nad plochou indolu. Obrázek 8.33 (B) zobrazuje graf náhodnosti tak, jak je definován na obrázku 8.32. Rozdíl je v tom, že data jsou analyzována dohromady nad i pod plochou indolu, ale pro definovaný kvádrový výsek prostoru. Obrázek 8.33 (C) je prostý histogram četnosti bodů v jednotlivých úsecích po 0,2 Å. Modré čáry přes obrázky upozorňují na vzájemnou porovnatelnost.



Obrázek 8.33: Analýza výřezu prostoru nad a pod plochou indolu. (A) Kvádrový výřez prostorového zobrazení z obrázku 8.24. (B) Graf náhodnosti úhlů normál Δ‘ (viz definice na obrázku 8.32). (C) Histogram četností bodů ve studovaných vrstvách (po 0,2 Å). Detailní popis viz text.

Pro účely kvantifikace jednotlivých vrstev indolů v okolí centrálního indolu jsem proložil histogramem četností z obrázku 8.33 (C) Gaussovy funkce v programu fityk. Konkrétní křivky vybraných proložených funkcí jsou na obrázku 8.34 (A). Parametry proložených funkcí jsou v tabulce na obrázku 8.34 (B). Pro porovnání relativního významu komponent jsem využil hodnotu plochy každé proložené Gaussovy funkce.

A



B

Amplituda (height)	Pozice (center)	Pološířka (hwhm)	Vypočtená plocha	Plocha (%)
29,2539	3,89479	0,178564	5,2236934	1,50
191,033	5,48543	0,594654	113,5985376	32,62
12,7967	8,03501	0,497294	6,36372213	1,83
30,7408	9,37381	0,532162	16,35908561	4,70
37,0876	13,8769	5,57322	206,6973541	59,35

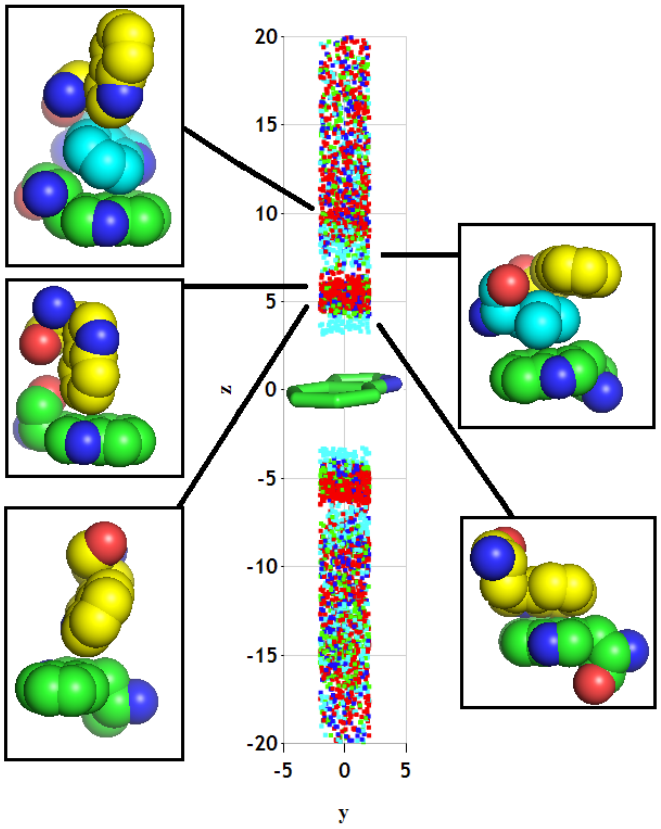
Obrázek 8.34: (A) Kvantifikace histogramu četností bodů ve studovaných vrstvách (viz obrázek 8.33 (C)). Červené body zobrazují četnosti indolů v jednotlivých vrstvách. Jednotlivé početnější oblasti jsou proloženy Gaussovými funkcemi (zeleně) a suma těchto křivek je vykreslena černou křivkou. Gaussova funkce není pro pravou část grafu (od 11 Å) zobrazena, ale do sumy je započtena. (B) Parametry Gaussových funkcí v pořadí od počátku.

Nyní popíši výsledky z obrázků 8.33 a 8.34. První vrstva do 4 Å, (ve vzdálenosti těžišť 3,89 ± 0,18 Å) se skládá z čistě paralelních „stacking“ tryptofanů. Tato vrstva obsahuje 1,5 % z celkového vzorku 3 947 tryptofanů (viz první Gaussova funkce na obrázku 8.34 (B)). Jak jsem již zmínil výše, nejbližší tryptofany musí mít paralelní orientaci, jinak by z definice nemohly mít těžiště nejblíže od centrálního indolu. Nicméně v této vrstvě musí být skutečná preference, jelikož dále od 4 Å do 4,3 Å četnosti klesají. Tento pokles je podle mě dán nepravděpodobností separací dvou indolů o vzdálenost odpovídající velikosti asi půl atomu. Dále se indoly začínají natáčet, až k nejčastější kolmé „T-shaped“ orientaci ve vzdálenosti 5,48 ± 0,59 Å (viz druhá Gaussova funkce na obrázku 8.34 (B)). Toto výrazné maximum se tvoří 33 % zobrazených tryptofanů, a navíc pokud bychom uvažovali pouze indoly do vzdálenosti 7 Å, pak tato skupina kolmých orientací tvoří 96 % vzorku. Co se týká naopak preferovaných orientací ve vzdálenosti 5,5 Å, tak výskyt indolů se skutečně kolmým uspořádáním normál (definováno skupinou A v obrázek 8.31 (A)) je asi 60krát častější, než by odpovídalo náhodnému uspořádání (viz obrázek 8.33 (B)).

S narůstající vzdáleností je detekovatelný opět pokles hustoty četnosti asi do 6,8 Å. Do této vzdálenosti se jedná zjevně o přímou interakci dvou indolů, nejprve paralelně a dále na kolmo uspořádaných. Vzdálenost okolo 7 Å již neumožňuje uskutečnění přímé interakce dvou indolů, snad pouze v případě, že by byly oba paralelně položené ve stejné rovině. Tato konfigurace je kvůli parciálním kladným nábojům méně pravděpodobná, a to by mohlo vysvětlit absolutní pokles četnosti v této konkrétní vzdálenosti indolů (viz obrázek 8.26).

Následuje velmi zajímavá vrstva poměrně vzácných (1,8 %) paralelních orientací. Jedná se o nad sebou umístěné indoly (vzdálenost 8,03 ± 0,49 Å), které mají mezi sebou přesně tolik místa, aby se mezi ně vešel alifatický postranní řetězec jiné aminokyseliny. Paralelní uspořádání je zde až 30krát častější, než by bylo náhodné (obrázek 8.33 (B)). Tato vrstva by však nebyla z grafu četností prakticky patrná, pokud by nereprezentovala obzvláště nečekané paralelní orientace. Podle grafu 8.33 (B) je ještě zajímavé relativně četné (4,7 %) nenáhodné rozdělení okolo 10 Å (9,37 ± 0,53 Å), velmi slabě tíhnoucí ke kolmé orientaci. V této vrstvě jsem nalezl kolmé indoly, které mezi sebou měly například aromatickou aminokyselinu, popřípadě peptidový řetězec. U vzdálenějších indolů se mi již nepodařilo nalézt žádné další charakteristické interakce. Prakticky zcela náhodná orientace tedy zabírá kolem 60 % zkoumaného prostoru.

Pro úplnost zobrazuji vzájemnou orientaci na obou stranách centrálního indolu na obrázku 8.35 (A) s ukázkou konkrétních struktur. Pod rovinou indolu ve vzdálenosti 12-15 Å je patrný zelený oblak bodů, který je již vysvětlen na obrázku 8.24.



Obrázek 8.35: Příklady orientací na obou stranách centrálního indolu (Příklady v pořadí od centrálního tryptofanu PDB: 4A2N, 4P6B, 5G2V, 2H9A, 2EPL).

Analýza popsaná v této kapitole by se dala jistě ještě značně rozšířit, ale bohužel jsem limitován rozsahem diplomové práce. Například by bylo zajímavé přidat i přesné natočení jednotlivých indolů (například podle úhlů přímek daných dipóly La a Lb viz kapitola 7.4), popřípadě i jiné aminokyseliny. Úhly vytvořené mezi normálami a dipóly (La a Lb) by sice měly povoleny libovolné hodnoty, ale například při filtrování na nějaký interval úhlů normál, by zcela zákonitě (a bez souvislosti s indolem) některé úhly (třeba La-Lb) byly zcela matematicky zakázány. Tuto část výsledků mám již sice z velké části zpracovanou, ale do diplomové práce jsem se rozhodl ji nezahrnout.

V této kapitole bylo zřetelně předvedeno, že při analýzách prostoru okolo aminokyselin, je nutné tento prostor vymezovat a zkoumat pouze v konkrétních segmentech. I k těmto segmentům je potřeba přistupovat obezřetně, jelikož například i u velkých vzdáleností se dá najít nenáhodné uspořádání (viz trojice aminokyselin na obrázku 8.35). Použitá prostorová zobrazení se ukázala jako nesmírně cenný zdroj informací, který jsem v jiných pracích nedohledal. Jen bych rád zmínil, že při nedostatečném odstranění homologních proteinů, se v prostorovém zobrazení objevovalo mnoho různých shluků, což zavádělo k nesprávným závěrům.

9 Souhrn

Značná část této práce se skládá z popisu odstraňování homologie proteinů. Ačkoli se tento postup může zdát jako velmi komplikovaný a zdlouhavý, tak tento vlastní přístup mi dovolil pracovat s pravděpodobně největším datasetem, se kterým jsem se v literatuře doposud setkal. Navíc jsem přesně díky tomu věděl, s jakými daty pracuji a na některé nápady do dalších analýz jsem přišel právě při odstraňování homologie proteinů. Díky takto robustnímu datasetu proteinů mi následně bylo umožněno studovat interakce do značných detailů. Při odstraňování homologie se projevila zvláštnost, že pokud se určitá aminokyselina vyskytuje u jedné plochy indolu, tak druhá nejbližší aminokyselina je nejčastěji o 0,5 Å dále od těžiště indolu, pokud se nachází na opačné straně jeho plochy. Pro tento jev zatím nemám uspokojivé vysvětlení. Ani nedokonalé rozlišení strukturních modelů nedokáže samo o sobě efekt vyvolat.

K shrnutí většiny výsledků v této práci potřebuji kontext textu, aby měl čtenář možnost konkrétní závěr plně pochopit. Proto jsem zde zmínil pouze ty nejzajímavější a z mého pohledu nejdůležitější výsledky.

Při analýze prostorového okolí tryptofanů jsem prokázal ovlivnění prostoru sekvenčně blízkými aminokyselinami, a proto by při studiu interakcí mezi aminokyselinami neměly být zahrnuty do analýz. Hlavní výstup v analýze prostorového okolí v této práci je graf 8.10, který zobrazuje preference jednotlivých aminokyselin k tryptofanu. Jako nejpreferovanější aminokyseliny se ukázaly být aromáty, což může souviset s popsanými aromatickými klastry, kde aromáty mají tendenci být u sebe. Nicméně analýza konkrétních výskytů aromátů okolo indolu neukázala nijak výrazné preferované pozice (jen shluky fenylalaninů u vazeb šesti uhlíkového cyklu indolu). Oproti tomu arginin s lysinem vykázaly silnou preferenci k ploše (těžišti) indolu, ale přitom se arginin u tryptofanu nevyskytuje více než v celém zbytku proteinu a lysin dokonce vykazuje nejvíce zápornou preferenci. Záporně nabité aminokyseliny a aminokyseliny s hydroxylovou nebo ketonovou skupinou hojně interagují s atomem dusíku indolové skupiny. K tryptofanu ale celkově nemají žádnou preferenci. Pouze dochází k natočení jejich postranního řetězce.

Většinu interakcí popsaných v kapitole 6.4 jsem pozoroval také. Nicméně například aniont-π interakci spojenou s tryptofanem jsem sice pozoroval, ale výskyt byl naprosto zanedbatelný. Proto se domnívám, že autoři Xavier Lukas a spol. (2016) postupovali ve své analýze naprosto chybně. Je pravda, že tuto konfiguraci nalezli, ale srovnal bych ji s náhodným výskytem jako arginin vyskytující se u hrany indolu, kde má být parciální kladný

náboj (viz obrázek 8.14). Porovnání síly jednotlivých interakcí jsem bohužel z důvodu omezení rozsahu práce nevykonal, ale rád bych se o to v dalším studiu pokusil. Mnoho výsledků v této práci potřebují kontext textu, proto jsem zde zmínil pouze ty nejzajímavější a z mého pohledu nejdůležitější.

Analýza tryptofanových párů se ukázala jako velmi přínosná. Jedná se především o prostorová zobrazení okolo indolové skupiny tryptofanu, která jasně ukázala, jak strukturovaný tento prostor je. Proto všechny analýzy prostorů okolo tryptofanu (pravděpodobně i jiných aminokyselin) je potřeba zkoumat jen v určitých směrech a ne jako kouli o určitém průměru. Práce zabývající se preferovanou orientací aromátů vůbec neberou v úvahu fakt, že prostor okolo indolu je rozdělen na vrstvy a v každé vrstvě se nalézá jiná preferovaná orientace. Například ve směru nad a pod plochou indolu se ve vzdálenosti 3,89 ± 0,18 Å od indolu nalézá vrstva paralelních „stacking" tryptofanů, která obsahuje 1,5 % z celkového vzorku 3 947 tryptofanů. Hned za ní je pokles výskytu a následuje široká vrstva kolmým „T-shaped" orientací ve vzdálenosti 5,48 ± 0,59 Å. Ty zabírají ve studovaném směru dokonce 33 %. A nejvíce překvapivé je, že nad touto vrstvou je opět úsek paralelních orientací, jak je zobrazeno na obrázku 8.35.

Jak je zmíněno v kapitole 6.4.1, autoři často zvolí určitou vzdálenost mezi aminokyselinami a v této kouli provádějí analýzu (například 5,5 Å (Thomas et al., 2002), 7 Å (Burley a Petsko, 1985), 7,5 Å (McGaughey et al., 1998)). V mé práci jsem ale došel k závěrům, že nenáhodné rozdělení dvou tryptofanů je možné dohledat i ve vzdálenosti 10 Å (viz obrázek 8.32 a 8.33 (B)), což pravděpodobně souvisí s interakcí se třetí aminokyselinou. Sám jsem prakticky vzdálenost zkoumaných aminokyselin neomezoval. Díky tomu jsem se tedy mohl dovědět, do jaké vzdálenosti se mění například četnosti jejich výskytu.

Analýza, v které by byla zahrnuta třetí aminokyselina, by jistě poskytla cenné informace. Bohužel by byla značně rozsáhlá a doufám, že se v budoucnu s touto problematikou vypořádám.

V této práci jsem se zaměřil především na nové postupy a jiný pohled na zkoumání prostoru okolo aminokyselin v konkrétní podobě tryptofanu. Ačkoli se práce zabývá především tryptofanem, většina postupů by se dala použít i na jakoukoli jinou aminokyselinu.

10 Seznam použité literatury

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990, BASIC LOCAL ALIGNMENT SEARCH TOOL: Journal of Molecular Biology, v. 215, p. 403-410.

Amicangelo, J. C., and P. B. Armentrout, 2000, Absolute binding energies of alkali-metal cation complexes with benzene determined by threshold collision-induced dissociation experiments and ab initio theory: Journal of Physical Chemistry A, v. 104, p. 11420-11432.

Balaji, P. V., 2011, Contribution of C-H center dot center dot center dot pi Interactions to the Affinity and Specificity of Carbohydrate Binding Sites: Mini-Reviews in Organic Chemistry, v. 8, p. 222-228.

Bhattacharyya, R., and P. Chakrabarti, 2003, Stereospecific interactions of proline residues in protein structures and complexes: Journal of Molecular Biology, v. 331, p. 925-940.

Bhattacharyya, R., U. Samanta, and P. Chakrabarti, 2002, Aromatic-aromatic interactions in and around alpha-helices: Protein Engineering, v. 15, p. 91-100.

Blundell, T., J. Singh, J. Thornton, S. K. Burley, and G. A. Petsko, 1986, AROMATIC INTERACTIONS: Science, v. 234, p. 1005-1005.

Brandl, M., M. S. Weiss, A. Jabs, J. Suhnel, and R. Hilgenfeld, 2001, C-H center dot center dot center dot pi-interactions in proteins: Journal of Molecular Biology, v. 307, p. 357-377.

Budyak, I. L., A. Zhuravleva, and L. M. Gierasch, 2013, The Role of Aromatic-Aromatic Interactions in Strand-Strand Stabilization of beta-Sheets: Journal of Molecular Biology, v. 425, p. 3522-3535.

Burley, S. K., and G. A. Petsko, 1985, AROMATIC-AROMATIC INTERACTION - A MECHANISM OF PROTEIN-STRUCTURE STABILIZATION: Science, v. 229, p. 23-28.

Burley, S. K., and G. A. Petsko, 1986, AMINO-AROMATIC INTERACTIONS IN PROTEINS: Febs Letters, v. 203, p. 139-143.

Butterfield, S. M., P. R. Patel, and M. L. Waters, 2002, Contribution of aromatic interactions to alpha-helix stability: Journal of the American Chemical Society, v. 124, p. 9751-9755.

Chakravarty, S., Z. Z. Sheng, B. Iverson, and B. Moore, 2012, "eta(6)"-Type anion-pi in biomolecular recognition: Febs Letters, v. 586, p. 4180-4185.

Chipot, C., R. Jaffe, B. Maigret, D. A. Pearlman, and P. A. Kollman, 1996, Benzene dimer: A good model for pi-pi interactions in proteins? A comparison between the benzene and the toluene dimers in the cas phase and in an aqueous solution: Journal of the American Chemical Society, v. 118, p. 11217-11224.

Chothia, C., 1976, NATURE OF ACCESSIBLE AND BURIED SURFACES IN PROTEINS: Journal of Molecular Biology, v. 105, p. 1-14.

Chourasia, M., G. M. Sastry, and G. N. Sastry, 2011, Aromatic-Aromatic Interactions Database, A(2)ID: An analysis of aromatic pi-networks in proteins: International Journal of Biological Macromolecules, v. 48, p. 540-552.

Cochran, A. G., N. J. Skelton, and M. A. Starovasnik, 2001, Tryptophan zippers: Stable, monomeric beta-hairpins: Proceedings of the National Academy of Sciences of the United States of America, v. 98, p. 5578-5583.

Colloc'h, N., L. Gabison, G. Monard, M. Altarsha, M. Chiadmi, G. Marassio, J. Santos, M. El Hajji, B. Castro, J. H. Abraini, and T. Prange, 2008, Oxygen pressurized X-ray crystallography: Probing the dioxygen binding site in cofactorless urate oxidase and implications for its catalytic mechanism: Biophysical Journal, v. 95, p. 2415-2422.

Conley, T. G., and D. G. Priest, 1980, NON-CLASSICAL INHIBITION OF URICASE BY CYANIDE: Biochemical Journal, v. 187, p. 733-738.

Cotelle, Y., V. Lebrun, N. Sakai, T. R. Ward, and S. Matile, 2016, Anion-pi Enzymes: Acs Central Science, v. 2, p. 388-393.

Cotte, N., M. N. Balestre, A. Aumelas, E. Mahe, S. Phalipou, D. Morin, M. Hibert, M. Manning, T. Durroux, C. Barberis, and B. Mouillac, 2000, Conserved aromatic residues in the transmembrane region VI of the V-1a vasopressin receptor differentiate agonist vs. antagonist ligand binding: European Journal of Biochemistry, v. 267, p. 4253-4263.

Daeffler, K. N. M., H. A. Lester, and D. A. Dougherty, 2012, Functionally Important Aromatic-Aromatic and Sulfur-pi Interactions in the D2 Dopamine Receptor: Journal of the American Chemical Society, v. 134, p. 14890-14896.

de Freitas, R. F., and M. Schapira, 2017, A systematic analysis of atomic protein-ligand interactions in the PDB: Medchemcomm, v. 8, p. 1970-1981.

Diana, D., C. Di Salvo, V. Celentano, L. De Rosa, A. Romanelli, R. Fattorusso, and L. D. D'Andrea, 2018, Conformational stabilization of a beta-hairpin through a triazole-tryptophan interaction: Organic & Biomolecular Chemistry, v. 16, p. 787-795.

Dougherty, D. A., 2007, Cation-pi interactions involving aromatic amino acids: Journal of Nutrition, v. 137, p. 1504S-1508S.

Dougherty, D. A., 2013, The Cation-pi Interaction: Accounts of Chemical Research, v. 46, p. 885-893.

Efron, B., 1979, 1977 RIETZ LECTURE - BOOTSTRAP METHODS - ANOTHER LOOK AT THE JACKKNIFE: Annals of Statistics, v. 7, p. 1-26.

Ellenbarger, J. F., I. V. Krieger, H. L. Huang, S. Gomez-Coca, T. R. Ioerger, J. C. Sacchettini, S. E. Wheeler, and K. R. Dunbar, 2018, Anion-pi Interactions in Computer-Aided Drug Design: Modeling the Inhibition of Malate Synthase by Phenyl-Diketo Acids: Journal of Chemical Information and Modeling, v. 58, p. 2085-2091.

Estarellas, C., A. Frontera, D. Quinonero, and P. M. Deya, 2011a, Anion-pi Interactions in Flavoproteins: Chemistry-an Asian Journal, v. 6, p. 2316-2318.

Estarellas, C., A. Frontera, D. Quinonero, and P. M. Deya, 2011b, Relevant Anion-pi Interactions in Biological Systems: The Case of Urate Oxidase: Angewandte Chemie-International Edition, v. 50, p. 415-418.

Fesinmeyer, R. M., F. M. Hudson, and N. H. Andersen, 2004, Enhanced hairpin stability through loop design: The case of the protein G B1 domain hairpin: Journal of the American Chemical Society, v. 126, p. 7238-7243.

Fischer, G., 1994, PEPTIDYL-PROLYL CIS/TRANS ISOMERASES AND THEIR EFFECTORS: Angewandte Chemie-International Edition in English, v. 33, p. 1415-1436.

Flocco, M. M., and S. L. Mowbray, 1994, PLANAR STACKING INTERACTIONS OF ARGININE AND AROMATIC SIDE-CHAINS IN PROTEINS: Journal of Molecular Biology, v. 235, p. 709-717.

Forbes, C. R., S. K. Sinha, H. K. Ganguly, S. Bai, G. P. A. Yap, S. Patel, and N. J. Zondlo, 2017, Insights into Thiol-Aromatic Interactions: A Stereoelectronic Basis for S-H/pi Interactions: Journal of the American Chemical Society, v. 139, p. 1842-1855.

Gallivan, J. P., and D. A. Dougherty, 1999, Cation-pi interactions in structural biology: Proceedings of the National Academy of Sciences of the United States of America, v. 96, p. 9459-9464.

Gerhard, R., H. Tatge, I. Just, and F. Hofmann, 2005, Characterisation of the tryptophan-101 mutant of clostridium difficile toxin A: Naunyn-Schmiedebergs Archives of Pharmacology, v. 371, p. R48-R49.

Griep, S., and U. Hobohm, 2010, PDBselect 1992-2009 and PDBfilter-select: Nucleic Acids Research, v. 38, p. D318-D319.

Hobohm, U., and C. Sander, 1994, ENLARGED REPRESENTATIVE SET OF PROTEIN STRUCTURES: Protein Science, v. 3, p. 522-524.

Hunter, C. A., and J. K. M. Sanders, 1990, THE NATURE OF PI-PI INTERACTIONS: Journal of the American Chemical Society, v. 112, p. 5525-5534.

Ishikawa, S., T. Ebata, H. Ishikawa, T. Inoue, and N. Mikami, 1996, Hole-burning and stimulated Raman-UV double resonance spectroscopies of jet-cooled toluene dimer: Journal of Physical Chemistry, v. 100, p. 10531-10535.

Jaffe, R. L., and G. D. Smith, 1996, A quantum chemistry study of benzene dimer: Journal of Chemical Physics, v. 105, p. 2780-2788.

Jenkins, D. D., J. B. Harris, E. E. Howell, R. J. Hinde, and J. Baudry, 2013, STAAR: Statistical analysis of aromatic rings: Journal of Computational Chemistry, v. 34, p. 518-522.

Kabsch, W., and C. Sander, 1983, DICTIONARY OF PROTEIN SECONDARY STRUCTURE - PATTERN-RECOGNITION OF HYDROGEN-BONDED AND GEOMETRICAL FEATURES: Biopolymers, v. 22, p. 2577-2637.

Kadam, R. U., D. Garg, J. Schwartz, R. Visini, M. Sattler, A. Stocker, T. Darbre, and J. L. Reymond, 2013, CH-pi "T-Shape" Interaction with Histidine Explains Binding of Aromatic Galactosides to Pseudomonas aeruginosa Lectin LecA: Acs Chemical Biology, v. 8, p. 1925-1930.

Kannan, N., and S. Vishveshwara, 2000, Aromatic clusters: a determinant of thermal stability of thermophilic proteins: Protein Engineering, v. 13, p. 753-761.

Kumar, K., S. M. Woo, T. Siu, W. A. Cortopassi, F. Duarte, and R. S. Paton, 2018, Cation-pi interactions in protein-ligand binding: theory and data-mining reveal different roles for lysine and arginine: Chemical Science, v. 9, p. 2655-2665.

Kumar, M., and P. V. Balaji, 2014, C-H center dot center dot center dot pi interactions in proteins: prevalence, pattern of occurrence, residue propensities, location, and contribution to protein stability: Journal of Molecular Modeling, v. 20, p. 14.

Kunthic, T., B. Promdonkoy, T. Srikhirin, and P. Boonserm, 2011, Essential role of tryptophan residues in toxicity of binary toxin from Bacillus sphaericus: Bmb Reports, v. 44, p. 674-679.

Lanzarotti, E., R. R. Biekofsky, D. A. Estrin, M. A. Marti, and A. G. Turjanski, 2011, Aromatic-Aromatic Interactions in Proteins: Beyond the Dimer: Journal of Chemical Information and Modeling, v. 51, p. 1623-1633.

Lee, E. C., D. Kim, P. Jurecka, P. Tarakeshwar, P. Hobza, and K. S. Kim, 2007, Understanding of assembly phenomena by aromatic-aromatic interactions: Benzene dimer and the substituted systems: Journal of Physical Chemistry A, v. 111, p. 3446-3457.

Lee, H., F. Dehez, C. Chipot, H. K. Lim, and H. Kim, 2019, Enthalpy-Entropy Interplay in pi-Stacking Interaction of Benzene Dimer in Water: Journal of Chemical Theory and Computation, v. 15, p. 1538-1545.

Liu, Y. M., Y. C. Liu, A. A. Gallo, K. D. Knierim, E. R. Taylor, and N. F. Tzeng, 2015, Performances of DFT methods implemented in G09 for simulations of the dispersion-dominated CH-pi in ligand-protein complex: A case study with glycerol-GDH: Journal of Molecular Structure, v. 1084, p. 223-228.

Lucas, X., A. Bauza, A. Frontera, and D. Quinonero, 2016, A thorough anion-pi interaction study in biomolecules: on the importance of cooperativity effects: Chemical Science, v. 7, p. 1038-1050.

Ma, B. Y., T. Elkayam, H. Wolfson, and R. Nussinov, 2003, Protein-protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces: Proceedings of the National Academy of Sciences of the United States of America, v. 100, p. 5772-5777.

Ma, J. C., and D. A. Dougherty, 1997, The cation-pi interaction: Chemical Reviews, v. 97, p. 1303-1324.

Macarthur, M. W., and J. M. Thornton, 1991, INFLUENCE OF PROLINE RESIDUES ON PROTEIN CONFORMATION: Journal of Molecular Biology, v. 218, p. 397-412.

Mahadevi, A. S., and G. N. Sastry, 2013, Cation-pi Interaction: Its Role and Relevance in Chemistry, Biology, and Material Science: Chemical Reviews, v. 113, p. 2100-2138.

Marion, D., 2013, An Introduction to Biological NMR Spectroscopy: Molecular & Cellular Proteomics, v. 12, p. 3006-3025.

Markus, M. T., and P. J. F. Groenen, 1998, An introduction to the bootstrap: Psychometrika, v. 63, p. 97-101.

Martinez, C. R., and B. L. Iverson, 2012, Rethinking the term "pi-stacking": Chemical Science, v. 3, p. 2191-2201.

McCaslin, T. G., C. V. Pagba, S. H. Chi, H. J. Hwang, J. C. Gumbart, J. W. Perry, C. Olivieri, F. Porcelli, G. Veglia, Z. J. Guo, M. McDaniel, and B. A. Barry, 2019, Structure and Function of Tryptophan-Tyrosine Dyads in Biomimetic beta Hairpins: Journal of Physical Chemistry B, v. 123, p. 2780-2791.

McGaughey, G. B., M. Gagne, and A. K. Rappe, 1998, pi-stacking interactions - Alive and well in proteins: Journal of Biological Chemistry, v. 273, p. 15458-15463.

Mitchell, J. B. O., C. L. Nandi, I. K. McDonald, J. M. Thornton, and S. L. Price, 1994, AMINO/AROMATIC INTERACTIONS IN PROTEINS - IS THE EVIDENCE STACKED AGAINST HYDROGEN-BONDING: Journal of Molecular Biology, v. 239, p. 315-331.

Musacchio, A., M. Saraste, and M. Wilmanns, 1994, HIGH-RESOLUTION CRYSTAL-STRUCTURES OF TYROSINE KINASE SH3 DOMAINS COMPLEXED WITH PROLINE-RICH PEPTIDES: Nature Structural Biology, v. 1, p. 546-551.

Navarro, G., 2001, A guided tour to approximate string matching: Acm Computing Surveys, v. 33, p. 31-88.

Neidigh, J. W., R. M. Fesinmeyer, and N. H. Andersen, 2002, Designing a 20-residue protein: Nature Structural Biology, v. 9, p. 425-430.

Ninkovic, D. B., J. M. Andric, S. N. Malkov, and S. D. Zaric, 2014, What are the preferred horizontal displacements of aromatic-aromatic interactions in proteins? Comparison with the calculated benzene-benzene potential energy surface: Physical Chemistry Chemical Physics, v. 16, p. 11173-11177.

Nishio, M., 2012, The CH/pi hydrogen bond: Implication in chemistry: Journal of Molecular Structure, v. 1018, p. 2-7.

Pace, C. N., G. R. Grimsley, and J. M. Scholtz, 2009, Protein Ionizable Groups: pK Values and Their Contribution to Protein Stability and Solubility: Journal of Biological Chemistry, v. 284, p. 13285-13289.

Padilla, C., L. Pardo-Lopez, G. de la Riva, I. Gomez, J. Sanchez, G. Hernandez, M. E. Nunez, M. P. Carey, D. H. Dean, O. Alzate, M. Soberon, and A. Bravo, 2006, Role of tryptophan residues in toxicity of Cry1Ab toxin from Bacillus thuringiensis: Applied and Environmental Microbiology, v. 72, p. 901-907.

Peter, B., A. A. Polyansky, S. Fanucchi, and H. W. Dirr, 2014, A Lys-Trp Cation-pi Interaction Mediates the Dimerization and Function of the Chloride Intracellular Channel Protein 1 Transmembrane Domain: Biochemistry, v. 53, p. 57-67.

Pinheiro, S., I. Soteras, J. L. Gelpi, F. Dehez, C. Chipot, F. J. Luque, and C. Curutchet, 2017, Structural and energetic study of cation-pi-cation interactions in proteins: Physical Chemistry Chemical Physics, v. 19, p. 9849-9861.

Pucci, F., and M. Rooman, 2016, Improved insights into protein thermal stability: from the molecular to the structurome scale: Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences, v. 374, p. 12.

Quinonero, D., C. Garau, A. Frontera, P. Ballester, A. Costa, and P. M. Deya, 2002a, Counterintuitive interaction of anions with benzene derivatives: Chemical Physics Letters, v. 359, p. 486-492.

Quinonero, D., C. Garau, C. Rotger, A. Frontera, P. Ballester, A. Costa, and P. M. Deya, 2002b, Anion-pi interactions: Do they exist?: Angewandte Chemie-International Edition, v. 41, p. 3389-3392.

Raimondi, S., N. Barbarini, P. Mangione, G. Esposito, S. Ricagno, M. Bolognesi, I. Zorzoli, L. Marchese, C. Soria, R. Bellazzi, M. Monti, M. Stoppini, M. Stefanelli, P. Magni, and V. Bellotti, 2011, The two tryptophans of beta 2-microglobulin have distinct roles in function and folding and might represent two independent responses to evolutionary pressure: Bmc Evolutionary Biology, v. 11, p. 12.

Reid, K. S. C., P. F. Lindley, and J. M. Thornton, 1985, SULFUR-AROMATIC INTERACTIONS IN PROTEINS: Febs Letters, v. 190, p. 209-213.

Robertazzi, A., F. Krull, E. W. Knapp, and P. Gamez, 2011, Recent advances in anion-pi interactions: Crystengcomm, v. 13, p. 3293-3300.

Rose, P. W., A. Prlic, A. Altunkaya, C. X. Bi, A. R. Bradley, C. H. Christie, L. Di Costanzo, J. M. Duarte, S. Dutta, Z. K. Feng, R. K. Green, D. S. Goodsell, B. Hudson, T. Kalro, R. Lowe, E. Peisach, C. Randle, A. S. Rose, C. H. Shao, Y. P. Tao, Y. Valasatava, M. Voigt, J. D. Westbrook, J. Woo, H. W. Yang, J. Y. Young, C. Zardecki, H. M. Berman, and S. K. Burley, 2017, The RCSB protein data bank: integrative view of protein, gene and 3D structural information: Nucleic Acids Research, v. 45, p. D271-D281.

Samanta, U., D. Pal, and P. Chakrabarti, 2000, Environment of tryptophan side chains in proteins: Proteins-Structure Function and Genetics, v. 38, p. 288-300.

Santiveri, C. M., and M. A. Jimenez, 2010, Tryptophan Residues: Scarce in Proteins but Strong Stabilizers of beta-Hairpin Peptides: Biopolymers, v. 94, p. 779-790.

Sencanski, M., L. Dosen-Micovic, V. Sukalovic, and S. Kostic-Rajacic, 2015, Theoretical insight into sulfur aromatic interactions with extension to D-2 receptor activation mechanism: Structural Chemistry, v. 26, p. 1139-1149.

Serrano, L., M. Bycroft, and A. R. Fersht, 1991, AROMATIC AROMATIC INTERACTIONS AND PROTEIN STABILITY - INVESTIGATION BY DOUBLE-MUTANT CYCLES: Journal of Molecular Biology, v. 218, p. 465-475.

Singh, J., and J. M. Thornton, 1985, THE INTERACTION BETWEEN PHENYLALANINE RINGS IN PROTEINS: Febs Letters, v. 191, p. 1-6.

Situ, A. J., S. M. Kang, B. B. Frey, W. An, C. Kim, and T. S. Ulmer, 2018, Membrane Anchoring of alpha-Helical Proteins: Role of Tryptophan: Journal of Physical Chemistry B, v. 122, p. 1185-1194.

Smith, M. S., E. E. K. Lawrence, W. M. Billings, K. S. Larsen, N. A. Becar, and J. L. Price, 2017, An Anion-pi Interaction Strongly Stabilizes the beta-Sheet Protein WW: Acs Chemical Biology, v. 12, p. 2535-2537.

Smyth, M. S., and J. H. J. Martin, 2000, x Ray crystallography: Journal of Clinical Pathology-Molecular Pathology, v. 53, p. 8-14.

Snyder, D. A., Y. Chen, N. G. Denissova, T. Acton, J. M. Aramini, M. Ciano, R. Karlin, J. F. Liu, P. Manor, P. A. Rajan, P. Rossi, G. V. T. Swapna, R. Xiao, B. Rost, J. Hunt, and G. T. Montelione, 2005, Comparisons of NMR spectral quality and success in crystallization demonstrate that NMR and X-ray crystallography are complementary methods for small protein structure determination: Journal of the American Chemical Society, v. 127, p. 16505-16511.

Spiwok, V., P. Lipovova, T. Skalova, E. Buchtelova, J. Hasek, and B. Kralova, 2004, Role of CH/pi interactions in substrate binding by Escherichia coli beta-galactosidase: Carbohydrate Research, v. 339, p. 2275-2280.

Stewart, D. E., A. Sarkar, and J. E. Wampler, 1990, OCCURRENCE AND ROLE OF CIS PEPTIDE-BONDS IN PROTEIN STRUCTURES: Journal of Molecular Biology, v. 214, p. 253-260.

Stubbe, J., and W. A. van der Donk, 1998, Protein radicals in enzyme catalysis: Chemical Reviews, v. 98, p. 705-762.

Sun, S., and E. R. Bernstein, 1996, Aromatic van der Waals clusters: Structure and nonrigidity: Journal of Physical Chemistry, v. 100, p. 13348-13366.

Sunner, J., K. Nishizawa, and P. Kebarle, 1981, ION-SOLVENT MOLECULE INTERACTIONS IN THE GAS-PHASE - THE POTASSIUM-ION AND BENZENE: Journal of Physical Chemistry, v. 85, p. 1814-1820.

Thomas, A., O. Bouffioux, D. Geeurickx, and R. Brasseur, 2001, Pex, analytical tools for PDB files. I. GF-Pex: Basic file to describe a protein: Proteins-Structure Function and Bioinformatics, v. 43, p. 28-36.

Thomas, A., R. Meurisse, B. Charloteaux, and R. Brasseur, 2002, Aromatic side-chain interactions in proteins. I. Main structural features: Proteins-Structure Function and Bioinformatics, v. 48, p. 628-634.

Thornton, J. M., J. Singh, S. Campbell, and T. L. Blundell, 1988, PROTEIN PROTEIN RECOGNITION VIA SIDE-CHAIN INTERACTIONS: Biochemical Society Transactions, v. 16, p. 927-930.

Tsuzuki, S., K. Honda, T. Uchimaru, M. Mikami, and K. Tanabe, 2000, The magnitude of the CH/pi interaction between benzene and some model hydrocarbons: Journal of the American Chemical Society, v. 122, p. 3746-3753.

Tu, T., Y. P. Li, Y. Wang, B. Yao, and H. Y. Luo, 2017, Probing the role of cation-pi interaction in the thermotolerance and catalytic performance of endo-polygalacturonases: Abstracts of Papers of the American Chemical Society, v. 254, p. 2.

Umezawa, Y., and M. Nishio, 2005, CH/pi hydrogen bonds as evidenced in the substrate specificity of acetylcholine esterase: Biopolymers, v. 79, p. 248-258.

Umezawa, Y., S. Tsuboyama, H. Takahashi, J. Uzawa, and M. Nishio, 1999, CH/pi interaction in the conformation of peptides. A database study: Bioorganic & Medicinal Chemistry, v. 7, p. 2021-2026.

Valley, C. C., A. Cembran, J. D. Perlmutter, A. K. Lewis, N. P. Labello, J. Gao, and J. N. Sachs, 2012, The Methionine-aromatic Motif Plays a Unique Role in Stabilizing Protein Structure: Journal of Biological Chemistry, v. 287, p. 34979-34991.

Warren, J. J., M. E. Ener, A. Vlcek, J. R. Winkler, and H. B. Gray, 2012, Electron hopping through proteins: Coordination Chemistry Reviews, v. 256, p. 2478-2487.

Waters, M. L., 2004, Aromatic interactions in peptides: Impact on structure and function: Biopolymers, v. 76, p. 435-445.

Weber, D. S., and J. J. Warren, 2018, A survey of methionine-aromatic interaction geometries in the oxidoreductase class of enzymes: What could Met-aromatic interactions be doing near metal sites?: Journal of Inorganic Biochemistry, v. 186, p. 34-41.

Wu, C. H., R. Apweiler, A. Bairoch, D. A. Natale, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Z. Huang, R. Lopez, M. Magrane, M. J. Martin, R. Mazumder, C. O'Donovan, N. Redaschi, and B. Suzek, 2006, The Universal Protein Resource (UniProt): an expanding universe of protein information: Nucleic Acids Research, v. 34, p. D187-D191.

Wu, W. J., and D. P. Raleigh, 1998, Local control of peptide conformation: Stabilization of cis proline peptide bonds by aromatic proline interactions: Biopolymers, v. 45, p. 381-394.

Yamamoto, Y., and J. Tanaka, 1972, POLARIZED ABSORPTION-SPECTRA OF CRYSTALS OF INDOLE AND ITS RELATED COMPOUNDS: Bulletin of the Chemical Society of Japan, v. 45, p. 1362-+.

Zhao, Y. J., Y. Cotelle, L. Liu, J. Lopez-Andarias, A. B. Bornhof, M. Akamatsu, N. Sakai, and S. Matile, 2018, The Emergence of Anion-pi Catalysis: Accounts of Chemical Research, v. 51, p. 2255-2263.

Zhuang, W. R., Y. Wang, P. F. Cui, L. Xing, J. Lee, D. Kim, H. L. Jiang, and Y. K. Oh, 2019, Applications of pi-pi stacking interactions in the design of drug-delivery systems: Journal of Controlled Release, v. 294, p. 311-326.

Zondlo, N. J., 2013, Aromatic-Proline Interactions: Electronically Tunable CH/pi Interactions: Accounts of Chemical Research, v. 46, p. 1039-1049.