VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ANALÝZA TRANSPORTU A DOKOVÁNÍ MALÝCH MOLEKUL UVNITŘ PROTEINOVÝCH TUNELŮ

LARGE-SCALE ANALYSIS OF THE LIGAND TRANSPORT AND DOCKING INSIDE OF THE PROTEIN TUNELS

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE **ANDREJ JEŽÍK**
AUTHOR

VEDÚCI PRÁCE **Ing. MILOŠ MUSIL**
SUPERVISOR

BRNO 2021

Vysoké učení technické v Brně
Fakulta informačních technologií

Ústav informačních systémů (UIFS) Akademický rok 2020/2021

Zadání bakalářské práce

Student: **Ježík Andrej**
Program: Informační technologie
Název: **Analýza transportu a dokování malých molekul uvnitř proteinových tunelů**
Large-Scale Analysis of the Ligand Transport and Docking inside of the Protein Tunels
Kategorie: Bioinformatika

Zadání:

1. Nastudujte problematiku tunelů a transportu malých molekul.
2. Nastudujte současnou strukturu nástroje Caver Web.
3. Získejte a předzpracujte datovou sadu léčiv schválených FDA pro použití programem Caver Web.
4. Upravte kód výpočetního jádra tak, aby bylo možné efektivně spouštět jednotlivé výpočty na superpočítači a zpracovávat jejich výsledky.
5. Navrhněte a implementujte uživatelské rozhraní pro komparativní analýzu výsledků.
6. Proveďte evaluaci získaných výsledků.

Literatura:

- Chovancova E, et al. CAVER3.0: a tool for the analysis of transport pathways in dynamic protein structures. PLOS computational biology. 2012, e1002708.
- Jurcik A, et al. CAVER Analyst 2.0: analysis and visualization of channels and tunnels in protein structures and molecular dynamics trajectories. Bioinformatics. 2018, 34, 3586-3588.
- Stourac J, et al. Caver Web 1.0: identification of tunnels and channels in proteins and analysis of ligand transport. Nucleic Acids Research, 2019, 47, 414-422.
- Vavra O, et al. CaverDock: a molecular docking-based tool to analyse ligand transport through protein tunnels and channels.
- Pinto G, et al. Fast screening of inhibitor binding/unbinding using novel software tool CaverDock. Frontiers in Chemistry. 2019, 7, 709.

Pro udělení zápočtu za první semestr je požadováno:

- První tři body zadání.

Podrobné závazné pokyny pro vypracování práce viz https://www.fit.vut.cz/study/theses/

Vedoucí práce: **Musil Miloš, Ing.**
Vedoucí ústavu: Kolář Dušan, doc. Dr. Ing.
Datum zadání: 1. listopadu 2020
Datum odevzdání: 12. května 2021
Datum schválení: 27. října 2020

Abstrakt

Táto práca sa zaoberá hromadnou analýzou transportu a dokovania malých molekúl, nazývaných ligandy, vnútri proteínových tunelov. Interakcie medzi proteínom a ligandom zahŕňajú procesy ako medzibunková signalizácia, transport, metabolizmus, regulácia, génová expresia a enzýmová aktivita. Porozumenie vzájomného pôsobenia medzi týmito molekulami je dôležitou súčasťou pri hľadaní nových liečiv. Procedúra proteín-ligand dokovania zahŕňa nasledujúce kroky: (i) nájdenie štruktúry proteínu (receptoru) a ligandu, (ii) identifikácia väzobných miest ligandu, (iii) zváženie flexibility ligandu a proteínu, a (iv) vypočet energie interakcie. V snahe zefektívniť danú procedúru pre veľké sety ligandov, bola v nástroji CaverWeb pridaná funkcionalita, ktorá umožní hromadnú analýzu dokovania ligandov do vybraného proteínu s kompletnou sadou liečiv, ktoré budú predspracované, čo umožní efektívnejší a plynulejší pracovný tok.

Abstract

This thesis discusses large-scale analysis of the ligand transport and docking inside of the protein tunnels. Protein-ligand interactions are involved in processes such as cell signalling, transport, metabolism, regulation, gene expression, and enzyme activity. To understand the interaction between these molecules is vitally important for the research for new pharmaceuticals. The procedure of protein–ligand docking involves the following steps: (i) finding the structures of proteins (receptors) and ligands, (ii) identifying ligand binding sites, (iii) considering receptor/ligand flexibility, and (iv) computing interaction energy between the receptor and the ligand. Additional functionality will be implemented to allow CaverWeb to test a complete set of pre-processed drug ligands on a protein, in an effort to enhance the efficiency of the procedure for large sets of ligands, which will allow a much smoother workflow.

Kľúčové slová

ligand, proteín, dokovanie, skórovacie funkcie, CaverWeb

Keywords

ligand, protein, docking, scoring functions, CaverWeb

Citácia

JEŽÍK, Andrej. *Analýza transportu a dokování malých molekul uvnitř proteinových tunelů.* Brno, 2021. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedúci práce Ing. Miloš Musil

Analýza transportu a dokování malých molekul uvnitř proteinových tunelů

Prehlásenie

Čestne prehlasujem, že som túto bakalársku prácu vypracoval samostatne pod vedením pána Ing. Miloša Musila. Uviedol som všetky literárne pramene, publikácie a ďalšie zdroje, z ktorých som čerpal.

.....
Andrej Ježík
10. mája 2021

Poďakovanie

Ďakujem pánovi Ing. Milošovi Musilovi za odborné vedenie práce, konzultácie a uvedenie do problematiky. Taktiež výskumnej skupine Loschmid Laboratories za asistenciu pri integrácii implementácie projektu do nástroja CaverWeb.

Obsah

1 Úvod **2**

2 Proteíny **4**
2.1 Aminokyseliny 4
2.2 Preklad DNA sekvencie do proteínu 6
2.3 Základné molekulárne procesy 6
2.4 Rozdelenie proteínov na základe tvaru molekuly 8
2.5 Rozdelenie proteínov na základe funkcie 9
2.6 Rozdelenie proteínov na základe molekulárnej štruktúry 10
2.7 Proteíny ako cieľ pre liečivá 11

3 Aktívne miesta, tunely a ligandy **13**
3.1 Aktívne miesta 13
3.2 Tunely a kanály 15
3.3 Ligandy 15

4 Problematika molekulárneho dokovania **19**
4.1 Príprava proteínu a ligandu 19
4.2 Flexibilita proteínu 22
4.3 Dokovanie ligandu do proteínu 24
4.4 Skórovanie 26

5 Nástroj Caver Web **29**
5.1 Nástroje 29
5.2 Pracovný postup 32
5.3 Detekcia tunelov 34
5.4 Dokovanie proteínu 35

6 Implementácia a výsledky **36**
6.1 Implementácia 36
6.2 Výsledky 41

7 Záver **44**

Literatúra **45**

A Obsah SD karty **49**

Kapitola 1

Úvod

Proteíny alebo bielkoviny sa vyskytujú v každej bunke ľudského tela, zabezpečujú mnoho vitálne dôležitých funkcií. Ako sú metabolizmus, pohyb, obrana, bunková komunikácia a molekulárne rozpoznávanie. Čo vysvetľuje, prečo je odvetvie vedy zaoberajúce sa proteínmi v centre biologického výskumu a je aplikované v disciplínach ako medicína, poľnohospodárstvo a biotechnológie.

Enzýmy patria do skupiny proteínov, ktoré zohrávajú kľúčovú rolu v metabolizme všetkých živých organizmov. Okrem iného, funkciou týchto bielkovín je štiepenie zložitejších molekúl, ktoré sú prijímané z potravy. To následne spôsobuje lepšiu vstrebateľnosť pri trávení. Niektoré enzýmy vyžadujú na uskutočnenie reakcie ďalšie molekuly ako sú napríklad vitamíny alebo minerály. Enzýmy napomáhajú tráveniu, produkcii energie, zrážaniu krvi, svalovým kontrakciám ako aj kopírovaniu genetickej informácie.

Pri návrhu liečiv je dôležitou súčasťou zisťovanie predpokladaného vzájomného pôsobenia liečiva a proteínu. Anatómia a vlastnosti proteínu sú kľúčové pri zisťovaní daných interakcií. Tunely sú nezanedbateľnou súčasťou proteínu. Sú charakterizované počiatkom a koncom, pričom spájajú dve bunkové prostredia a sprostredkovávajú transport rôznych malých molekúl zvaných ligandy do alebo z proteínu. Ligand na splnenie svojej funkcie musí byť schopný dostať sa k aktívnemu miestu v proteíne, kde sa naviaže a slúži ako inhibítor alebo katalyzátor.

Cieľom tejto práce je oboznámiť čitateľa s problematikou dokovania proteínov, priblížiť rôzne prístupy využívané na riešenie problémov spojených s molekulárnym dokovaním a implementovať modul, ktorý umožní používateľom webovej aplikácie Caver Web [36] hromadnú analýzu proteínu s predspracovanou množinou overených liečiv z databázy Zinc [35].

V úvode teoretickej časti 2 sú popísané aminokyseliny ako stavebné bloky proteínov, nasledované popisom proteínov a ich funkciou v živých organizmoch.

Aktívnym miestam je venovaná kapitola 3, ktorá sa zaoberá ich významom, ako aj pozíciu v proteíne ako aj látky, ktoré sa naväzujú na tieto miesta.

Problematika molekulárneho dokovania je rozobraná v kapitole 4. Je rozdelená do troch fáz, ktoré sú neodlučiteľnou časťou tejto problematiky. V prvej časti sú opísané problémy, ktoré sú riešené počas prípravy receptoru a ligandu na proces dokovania. Druhá časť sa venuje už priamo procesu dokovania ligandu do receptoru, kde metódy delíme medzi systematické a stochastické. V tretej časti sú rozoberané rôzne prístupy skórovania výsledkov poskytnutých dokovacími metódami.

Ako jedno z riešení molekulárneho dokovania je popísaný nástroj Caver Web v kapitole 5, do ktorého bude integrovaná funkcionalita hromadného molekulárneho dokovania. Nástroj je postavený na nástrojoch Caver [9], ktorý slúži pre analýzu tunelov v štruktúre

receptoru a nástroji CaverDock [37], ktorý umožňuje simuláciu dokovania ligandu do aktívneho miesta.

V prvej polovici kapitoly venovanej praktickej časti 6 sú opísané použité nástroje na spracovanie molekúl, databáza a dataset, ktorý bol vybraný pre implementáciu modulu. Druhá časť kapitoly sa zaoberá evaluáciou získaných výsledkov. Záver 7 je venovaný sumarizácii práce, so zameraním na vystihnutie najdôležitejších častí.

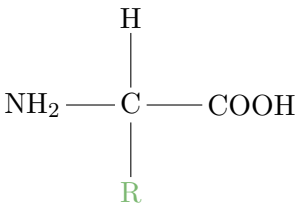
Kapitola 2

Proteíny

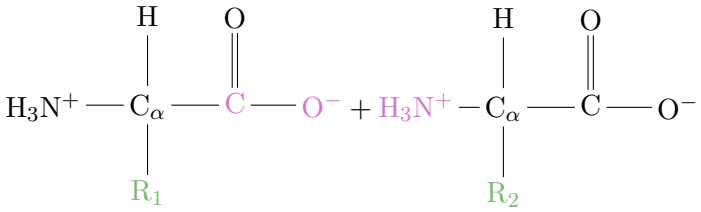
Proteíny spolu s nukleovými kyselinami a polysacharidmi patria medzi tri hlavné biologické makromolekuly. Slúžia ako základná stavebná jednotka všetkých organizmov. Sú to lineárne polyméry obsahujúce sekvenciu niekoľko tisíc aminokyselín prepojených kovalentnými peptidovými väzbami. Zabezpečujú viacero funkcií ako sú: transport iónov a molekúl z jedného orgánu do druhého, regulácia bunkových a fyziologických aktivít, a taktiež spĺňajú funkciu protilátok. Proteíny sú tvorené takzvanými proteínovými sekvenciami, ktoré pozostávajú z 21 rôznych druhov chemických zlúčenín, nazývaných aminokyseliny [29].

2.1 Aminokyseliny

Sú stavebné jednotky proteínov. Formujú štruktúru polypeptidu ako je demonštrované vo figúre 2.2. Proteín môže pozostávať z jedného alebo viacerých polypeptidov. Aminokyseliny majú charakteristickú štruktúru a ich všeobecný vzorec je možné vidieť na obrázku 2.1. Pozostávajú z centrálneho atómu α-uhlíka, na ktorý sú viazané štyri rozdielne chemické skupiny: amino (-NH2) funkčná skupina, karboxylová (-COOH) funkčná skupina, atóm vodíka (H) a bočný reťazec (R), špecifický pre každú aminokyselinu. Aminokyseliny sa klasifikujú na základe viacerých parametrov: veľkosť, tvar, náboj, hydrofóbnosť a chemická reaktivita bočných reťazcov [14, 29].

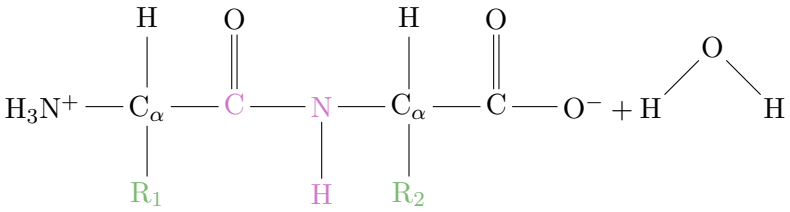


Obr. 2.1: Všeobecný vzorec aminokyseliny.



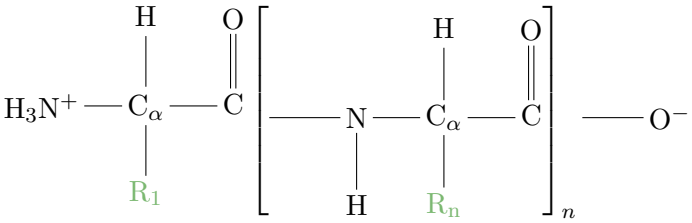
(2.1)

(a) Všeobecný vzorec pre dve aminokyseliny, konkrétny typ je definovaný reťazcom R. Každá aminokyselina obsahuje centrálny atóm uhlíka nazývaného $\alpha-carbon$, na ktorý je naviazaná karboxylová a amino skupina.



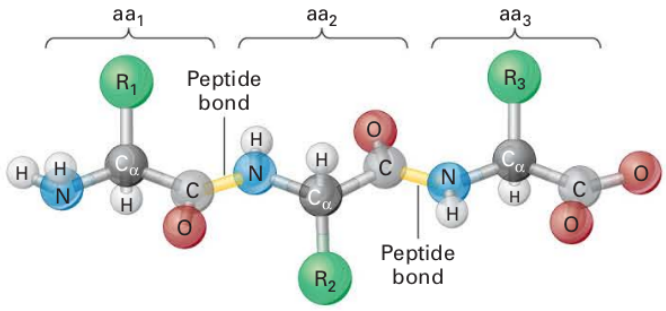
(2.2)

(b) Medzi aminokyselinami bola vytvorená polypeptidová väzba. Ako vedľajší produkt tejto reakcie vznikla molekula vody



(2.3)

(c) Všeobecný vzorec polypeptidu obsahujúceho $n+1$ aminokyselín.



(d) Každý úsek s označením aa_n reprezentuje jednu aminokyselinu v polypeptidovom reťazci

Obr. 2.2: Štruktúra polypeptidu. Prevzaté z [29] a upravené.

2.1.1 Rozdelenie aminokyselín na základe hydrofobicity

Na základe ich tendencie reagovať na prítomnosť vody, aminokyseliny je možné rozdeliť do dvoch skupín: hydrofobické a hydrofilické.

- **Hydrofobické**
  Sú zle rozpustné vo vode. Hydrofóbnosť aminokyseliny je nepriamo úmerná polarite bočného reťazca [29].
  - Nepolárne bočné reťazce: Alanín (Ala), Valín (Val), Leucín (Leu), Izoleucín (Ile).
- **Hydrofilické**
  Aminokyseliny, ktoré obsahujú polárny bočný reťazec, sú dobre rozpustné vo vode. Tieto reťazce sa začleňujú do troch skupín [29].
  - Zásadité aminokyseliny: Arginín (Arg), Lyzín (Lys).
  - Kyslé aminokyseliny: kyselina asparágová (Asp), Asparagín (Asn), kyselina glutámová (Glu), Glutamín (Gln).
  - Aminokyseliny obsahujúce aromatické jadro: Histidín (His).

2.2 Preklad DNA sekvencie do proteínu

Nukleové kyseliny sú lineárne polyméry. Tvoria ich štyri typy nukleotidov. Nukleotidy obsahujú informáciu slúžiacu na rozpoznávanie aminokyselín, štruktúry a funkcie všetkých proteínov bunky. *DNA*[1] je molekula obsahujúca, v jej sekvencii nukleotidov, informáciu potrebnú na stavbu všetkých proteínov, buniek a tkanív organizmu. DNA a RNA (Ribonukleová kyselina) sú zložené z troch zhodných báz: adenín (A), guanín (G), cytozín (C) a jednej rozdielnej. Ako štvrtú bázu má DNA tymín (T) a RNA uracil (U). Časť DNA obsahujúca informáciu, ktorá špecifikuje syntézu polypeptidového reťazca alebo funkčnej RNA sa nazýva *gén*[2] [29].

2.3 Základné molekulárne procesy

V tejto sekcii sú stručne opísané molekulárne procesy, ktoré sú vyobrazené vo figúre 2.3.

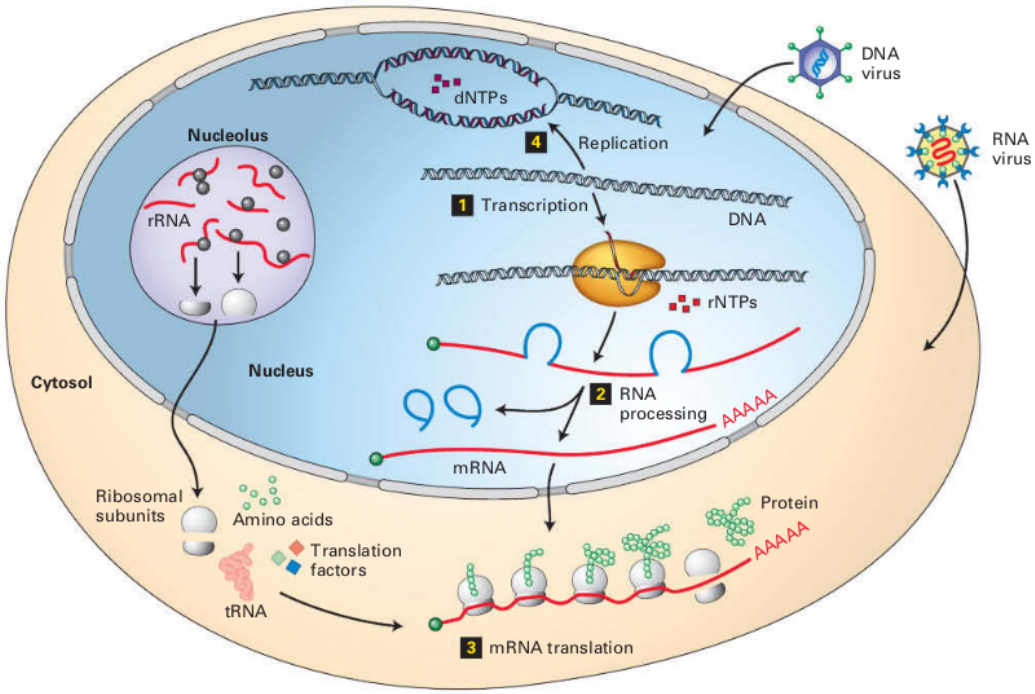
Transkripcia

Transkripcia je proces, počas ktorého RNA polymeráza s pomocou iniciačných faktorov rozpozná a naviaže sa na špecifickú sekvenciu DNA nazývanú promotor. Po naviazaní, RNA polymeráza a iniciačné faktory oddelia DNA vlákna. Tým je vytvorený podklad pre párovanie s bázami ribonukleotidov (rNTP). Proces polymerizácie rNTPs sa začína v bode, keď RNA polymerázy vytvoria transkripčnú bublinu. Po prepísaní určitej oblasti DNA, vlákna ktoré sú od seba oddelené sa spoja a vytvoria dvojzávitnicu. Následne sa RNA polymeráza posúva po DNA v smere 5´- 3´ a syntetizuje RNA polynukleotid. Ukončenie procesu nastáva v momente, keď sa polymeráza dostane na konkrétnu sekvenciu, značiacu koniec kódu [29].

[1] *DNA* alebo Deoxyribonukleovej kyseline je uložená genetická informácia bunky, riadi rast, delenie a regeneráciu bunky.
[2] *Gén* je základná jednotka genetickej informácie.

Translácia

Translácia je prevod genetickej informácie na proteínové sekvencie pomocou ribozómov. Translácia formuje doplnkovú mediátorovú RNA (mRNA) z DNA cez katalýzu RNA polymerázy. RNA polymeráza kopíruje DNA pričom nahrádza tymín za uracil v mRNA. V procese translácie je nukleotidová sekvencia mRNA prečítaná pomocou tRNA (transferová RNA), rRNA (ribozomálna RNA) a asociovaných proteínov. Pokiaľ sú aminokyseliny uložené správne do sekvencie pomocou tRNA, následne budú prepojené peptidovými väzbami a vytvoria proteín. Počas procesu translácie je štvor-bázový kód mRNA dekódovaný do jazyka 21 aminokyselín proteínov. Ribozómy, ktoré prekladajú mRNA kód, sú zložené z dvoch podjednotiek: ribozomálnych proteínov a rRNA. Po transporte do cytoplazmy sú ribozomálne podjednotky spojené s mRNA a vykonávajú proteínovú syntézu za pomoci tRNA a proteínov translačného faktora [29].



Obr. 2.3: Základné molekulárne procesy [29].

Replikácia

Replikácia DNA, je proces ktorý prebieha iba v bunkách pripravujúcich sa na rozdelenie, monoméry deoxyribonukleozid trifosfátu (dNTPs) sú polymerizované, aby vytvorili dve identické kópie každej chromozomálnej DNA molekuly. Dcérska bunka získa jednu z identických kópií. Každé vlákno v materskej duplexnej DNA sa správa ako šablóna pre syntézu, tvoriaca dcérsky duplex. Pravidelné párovanie bázií v dvojzávitnicovej štruktúre DNA popísané Watsonom a Crickom uvádza, že nové vlákna DNA sú syntetizované pomocou existujúceho materského vlákna ako šablóny pri formácii nového dcérskeho vlákna, ktoré je doplnkom pre materské vlákna. Nové vlákna sa tvoria v smere 5´-3´. Replikácia začína na

presne určenom mieste v sekvencii, nazývanom replikačný začiatok. V každej eukaryotickej chromozomálnej DNA sa nachádza niekoľko replikačných začiatkov [29].

2.4 Rozdelenie proteínov na základe tvaru molekuly

Proteíny je možné klasifikovať na základe tvaru do niekoľkých skupín. Fibrilárne proteíny majú polypeptidové reťazce usporiadané do dlhých vlákien. Globulárne proteíny sú tvorené polypeptidovými reťazcami usporiadanými do sférického alebo globulárneho tvaru.

Fibrilárne proteíny

Nie sú rozpustné vo vode, ich tvar pripomína vlnité vlákno. Zohrávajú dôležitú rolu pri anatómii a fyziológii stavovcov. Zabezpečujú vonkajšiu ochranu, podporu, tvar a formu. Zvyčajne sú stavané na základe jednej opakujúcej sa štruktúry pripomínajúcej nite. Príkladom fibrilárnych proteínov sú fibroin a α-keratin, vyskytujúci sa vo vlasoch a nechtoch. Ďalším významným fibrilárnym proteínom je kolagén, ktorý je hlavný proteínový komponent kože, kostí, zubov, šliach a väzív [14].

Globulárne proteíny

Sú rozpustné vo vode, ich tvar pripomína klbko vlákien. Môžu byť klasifikované do štyroch skupín [14]:

- **I.** all-α, v ktorých dominujú α-helixy, $\alpha > 40\%$ a $(\beta) < 5\%$,
- **II.** all-β, v ktorých dominujú β-štruktúry, $\beta > 40\%$ a $(\alpha) < 5\%$,
- **III.** α+β, kde sa vyskytujú α- helixy $> 15\%$ a β-štruktúry $> 10\%$, nemiešajú sa, ale oddeľujú sa od polypeptidového reťazca
- **IV.** α/β, kde sa vyskytujú α- helixy $> 15\%$ a β-štruktúry $> 10\%$, rovnomerne sa miešajú.

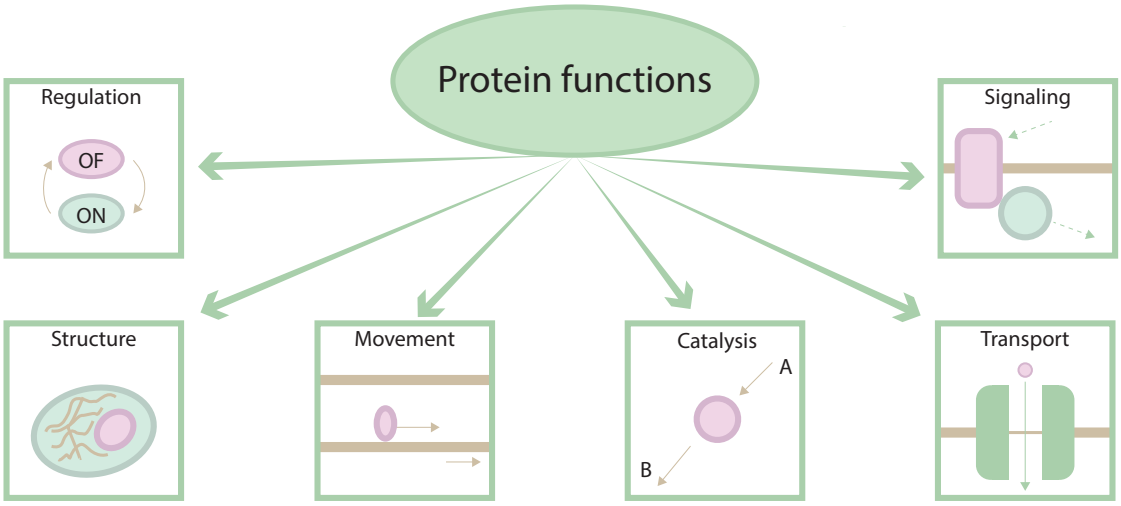
Membránové proteíny

Vyžadujú uchytenie na lipidové dvojvrstvy. Tie musia obsahovať sekvencie aminokyselín, ktoré sa dokážu zložiť spolu s hydrofobickým povrchom pri kontakte s alkánovými reťazcami lipidov. Druhou možnosťou uchytenia je polárny povrch v kontakte s vodnými fázami na oboch stranách membrány. Membránové proteíny zabezpečujú medzibunkové signály, enzymatickú činnosť, prenos iónov a roztokov cez membránu [14].

Vnútorne neusporiadané proteíny

V ich prirodzenom funkčnom stave nemajú fixne usporiadanú štruktúru, čo robí polypeptidové reťazce veľmi flexibilné. Flexibilita neusporiadaných proteínov je kľúčovým faktorom pri viacerých aktivitách. Medzi tieto aktivity patria: vzájomné interakcie s ďalšími proteínmi alebo spojenie sa do preddefinovanej *konformácie*[3] po naviazaní ďalších partnerov. Tieto proteíny slúžia predovšetkým ako signalizačné molekuly, regulátory aktivít molekúl alebo podporné štruktúry pre skupiny proteínov, malé molekuly a ióny.

[3] *Konformácia* opisuje vnútorné usporiadanie atómov.



Obr. 2.4: Rozdelenie proteínov na základe ich funkcie v organizme. Použité z [14] a upravené.

2.5 Rozdelenie proteínov na základe funkcie

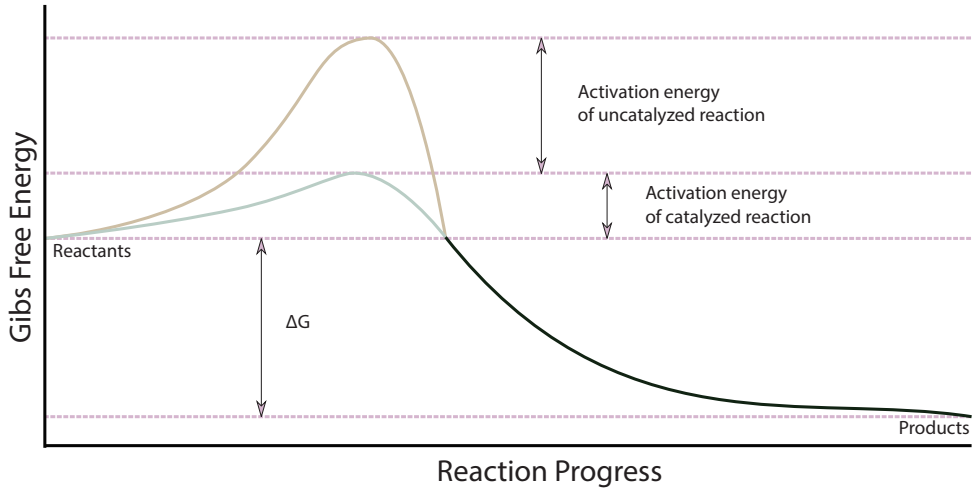
Ako bolo uvedené, proteíny vykonávajú veľa funkcí, ktoré je možné rozdeliť do niekoľkých skupín. Tie sú ukázané v obrázku 2.4.

- **Regulácia (hormóny)**
  Vykonávajú kontrolu proteínových aktivít. Usporiadania v štruktúre proteínu zohrávajú dôležitú rolu pri regulácii. Aktivity všetkých proteínov a ďalších biomolekúl sú regulované tak, aby ich spojenie funkcií bolo optimálnym pre prežitie. Katalytická aktivita enzýmov je regulovaná tak, aby veľkosť reakčných produktov zodpovedala aktuálnym potrebám bunky [29].

- **Štruktúra (tropokolagén, keratín)**
  Vykonávajú organizáciu genómu, organel, cytoplazmy, proteínových komplexov a membrán. V trojdimenzionálnom priestore vytvárajú bunky a tkanivá organizmov [29].

- **Pohyb (aktín, myozín, tubulín, dyneín)**
  Generujú energiu potrebnú pre pohyb prostredníctvom motorických proteínov [29].

- **Katalýza (trypsín, DNA polymerázy a ligázy)**
  Enzýmy katalyzujú chemické reakcie, čiže tvoria a rozbíjajú kovalentné väzby. Znižujú bariéru aktivačnej energie a konvertujú molekuly substrátu, v biochémii nazývaného ligand, na produkty. Skoro každá chemická reakcia prebiehajúca v bunke je katalyzovaná pomocou špecifického katalyzátoru. Grafické zobrazenie vplyvu katalizátora vidíme v grafe 2.5.
  Určité druhy enzýmov sa vyskytujú vo väčšine buniek. Dôvodom je, že katalyzujú syntézu bežných bunečných produktov. Druhou možnosťou častého výskytu je ich úloha zhromažďovania energie zo živín (premena glukózy a kyslíka na oxid uhličitý a vodu). Väčšina enzýmov sa vyskytuje v bunkách, avšak niektoré enzýmy sú vylučované a ich funkcia prebieha v mimobunkových miestach (krv a tráviaci trakt), alebo úplne mimo organizmu [29].



Obr. 2.5: Tento hypotetický graf opisuje zmeny vo voľnej energii (G) počas chemickej reakcie. Reakcia sa vyskytuje spontánne iba pokiaľ G_{sum} produktov je nižšie ako celkové G reaktantov (negatíve ΔG). Všetky chemické reakcie však prechádzajú jedným alebo viacerými vysokoenergetickými prechodovými stavmi. Rýchlosť reakcie je nepriamo úmerná aktivačnej energii, čo znamená rozdiel voľnej energie medzi reaktantmi a prechodovým stavom (najvyšším bodom pozdĺž trvania reakcie). Enzýmy a ostatné katalyzátory zvyšujú výskyt reakcií znížením aktivačnej energie pre prechodové stavy.

- **Transport (hemoglobín, myoglobín, sérový albumín)**
  Prepravujú malé molekuly a ióny cez plazmatické membrány buniek [29].
- **Signalizácia (inzulín, hormón stimulujúci štítnu žľazu)**
  Monitorujú prostredie a informácie, ktoré sú prenášané [29].

2.6 Rozdelenie proteínov na základe molekulárnej štruktúry

Bielkoviny môžu svoje funkcie vykonávať, len pokiaľ sú zbalené do patričného tvaru, tzv. priestorovej konformácie. Správna konformácia bielkovín zabezpečuje špecifickú interakciu s inou molekulou (napr. enzým – substrát), sprostredkovanú slabými nekovalentnými väzbami (vodíkové mostíky, Van der Waalsove sily, iónové alebo hydrofóbne interakcie) medzi aminokyselinovými postrannými reťazcami a danou molekulou (rôznej chemickej povahy). Konformácia bielkoviny je priamo daná jej zložením, t.j. počtom a poradím jednotlivých aminokyselín tvoriacich polypeptidový reťazec. Tak, ako v prípade nukleových kyselín, aj v prípade bielkovín je možné sa stretnúť s rôznym stupňom priestorovej organizácie jej stavebných molekúl (aminokyselín)[14]. Je to možné vidieť vo figúre 2.6.

Primárna štruktúra (reťazec)

Lineárne kovalentné usporiadanie alebo sekvencia aminokyselín v proteíne bez štruktúry. Formácia peptidovej väzby medzi aminovou skupinou jednej aminokyseliny a karboxylovej

skupiny inej aminokyseliny spôsobí vytvorenie molekuly vody, pričom sa voda odštepuje – dehydratačná reakcia [14].

Sekundárna štruktúra (lokálne skladanie)

Stabilné, regulárne priestorové usporiadanie segmentov polypeptidových reťazcov prepojených vodíkovými mostíkmi. Možnosť segmentu polypeptidového reťazca vytvoriť sekundárnu štruktúru závisí na sekvencii aminokyselín. Medzi najdôležitejšie sekundárne konformácie sa radia skrutkovica (α-hélix), skladaný list (β-štruktúra) a krátky do U tvarovaný β-ohyb. V priemere 60% polypeptidových reťazcov tvoria α-helixy a β-štruktúry, zvyšok molekuly tvoria nepravidelné štruktúry, cievky a ohyby. *A*-helixy a β-štruktúry konformácie sú hlavnými vnútornými podpornými elementami v proteínoch [14].

Terciárna štruktúra (celková konformácia)

Odkazuje na priestorový vzťah všetkých aminokyselín v polypeptide. Je to celková konformácia polypeptidového reťazca, úplné trojdimenzionálne usporiadanie všetkých aminokyselín. Celková podoba závisí od vlastností aminokyselín a ich usporiadania v reťazci. Terciárna štruktúra je primárne stabilizovaná hydrofobickými interakciami medzi nepolárnymi stranami reťazca spolu s vodíkovými väzbami obsahujúcimi polárne strany reťazcov. Tvorí sa počas, alebo hneď po polymerizácii, keď sa lineárny reťazec aminokyselín skladá do komplexnejších tvarov, dosahujúcich charakteristickú trojdimenzionálnu štruktúru. Proteíny, ktoré sú polymérmi aminokyselín môžu mať rôzne veľkosti a tvary. Lineárny, nerozvetvetvený polymér aminokyselín obsahujúci ľubovolný proteín sa zloží do jedného alebo zopár veľmi podobných trojdimenzionálnych tvarov, nazývaných konformácie. Funkciu proteínu určuje jeho konformácia spolu s charakteristickými vlastnosťami bočného reťazca aminokyselín. V niektorých prípadoch sa dokáže zmeniť konformácia a takisto aj funkcia proteínu. Tento jav nastane, pokiaľ sa proteín kovalentne alebo nekovalentne spojí s ďalšími molekulami [29, 14].

Kvaternárna štruktúra (multimérna štruktúra)

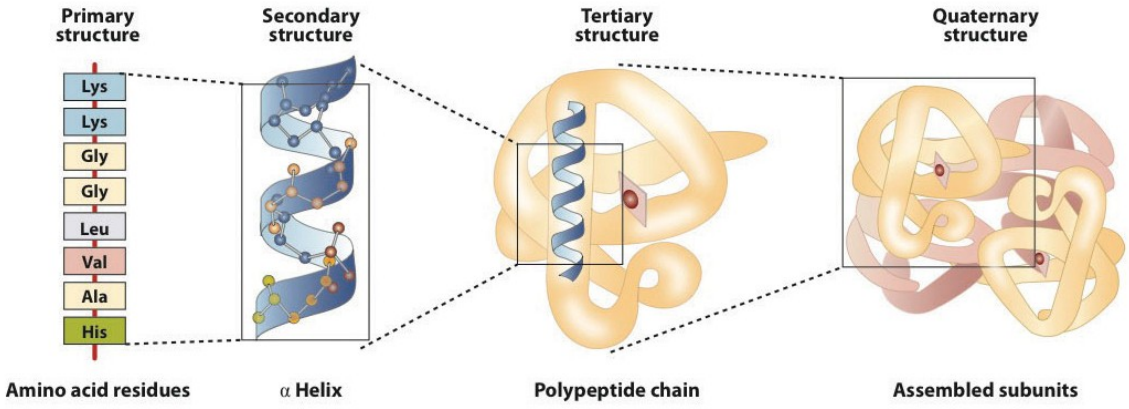
Táto štruktúra odkazuje na priestorový vzťah polypeptidov alebo častí v rámci proteínu. Je to združenie dvoch alebo viacerých polypeptidových reťazcov do oligoméreneho proteínu, alebo do niekoľkých podčastí, ktoré môžu byť identické alebo rozdielne. Kvaternárna štruktúra taktiež obsahuje *kofaktory*[4] a kovy, ktoré tvoria katalytické jednotky a biologicky funkčné proteíny [14].

2.7 Proteíny ako cieľ pre liečivá

Proteíny zaisťujú mnoho funkcií v ľudskom organizme. Celkové zdravie organizmu závisí od správnej funkcie proteínov, akýkoľvek úbytok týchto funkcií môže viesť k vývoju patologického procesu. Všetky zmeny v aktivite proteínu spôsobené dedičnými faktormi, vystavením sa toxínom alebo radiácii sú uložené v bázach mnohých patogénov, ako je napríklad rakovina alebo metabolická choroba.

Vo farmácii sa v súčastnosti venuje výskumu proteínov 80% zdrojov. Väčšina farmaceutických liečiv sa napája na enzým, receptor, iónový mostík, transportný proteín alebo

[4] *Kofaktor* je nízkomolekulová neproteínová zložka enzýmov ovplyvňujúca ich katalytickú funkciu.



Obr. 2.6: Primárna štruktúra je charakterizovaná ako sekvencia aminokyselín. Sekundárna štruktúra z lokálne poskladaných polypeptidových reťazcov formujúcich skrutkovice alebo skladané listy. Terciárna štruktúra predstavuje definitívne priestorové usporiadanie skrutkovíc a skladaných listov v priestore, poprepájaných rôzne zvlneným polypeptidovým vláknom. Kvaternárna štruktúra reprezentuje definitívne usporiadanie terciárnych štruktúr do komplexnej makromolekuly bielkoviny.

nukleárny proteín a reverzibilne menia ich aktivitu. Príkladom sú liečivá SSRI (*selective serotonin reuptake inhibitors*), ktoré sa primárne používajú na liečenie depresií a úzkosti. SSRI sú považované za liečivá s efektívnym účinkom a relatívne miernymi vedľajšími účinkami. Obe tieto výhody sú prisudzované vďaka ich schopnosti selektívne potlačovať transportný proteín smerujúci do mozgu. Ten je zodpovedný za opätovné odchytenie neurotransmiteru serotonínu (proteín, ktorý vracia serotonín z nervovej synapsie naspäť do vylučovacieho neurónu). Opätovné zachytávanie neurotransmitterov zabraňuje nadmernej stimulácii ich cieľových receptorov. Pokiaľ je človek zdravý, táto funkcia ostáva stále aktívna. Pokiaľ človek trpí depresiami, synaptické hladiny serotonínu sú nízke, takže opätovné zachytávanie môže pomôcť proti depresiám. Funkčne väčšina proteínov so zameraním na lieky je klasifikovaných ako receptory spojené s G proteínmi (GPCRs) alebo enzýmami. Lieky so zameraním na enzýmy, ktoré sa správajú ako inhibítory predstavujú 25% obchodu s liečivami [21].

Kapitola 3

Aktívne miesta, tunely a ligandy

Enzýmy obsahujú aktívne miesta, ktoré sú nevyhnutné pre funkciu proteínu. Naväzujú sa na ne ligandy, ktoré aktivujú alebo deaktivujú proteín. Tieto miesta sa môžu vyskytovať skryté v proteíne, kde k ním vedie prístupová cesta nazývaná tunel.

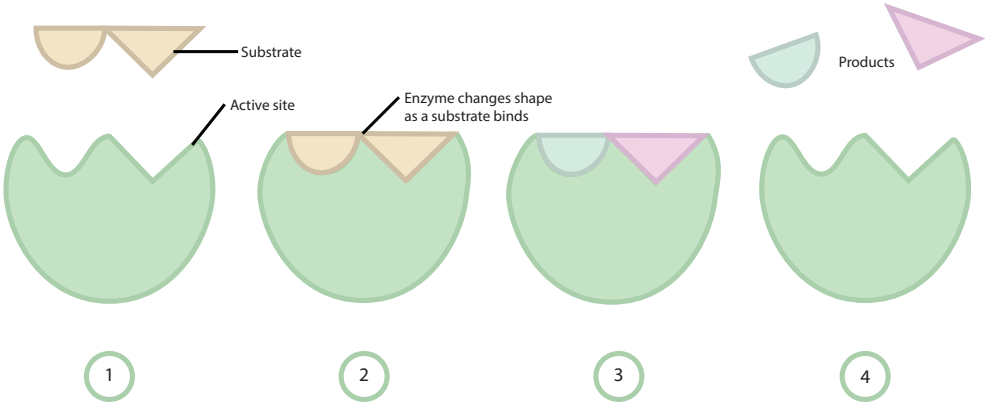
3.1 Aktívne miesta

Sú časti enzýmov, kde sa naväzujú substráty nazývané ligandy. Toto je kľúčové pre katalytickú aktivitu enzýmov. Deje sa to, keď enzýmy vzájomne pôsobia so substrátmi, chemickými reaktantmi takým spôsobom, že reakcia má väčšiu pravdepodobnosť prebehnúť. Táto interakcia prebieha v aktívnom mieste, kde enzým vytvára väzbu so substrátom, za účelom zvýšenia ich šance na reakciu.

3.1.1 Vlastnosti ovplyvňujúce viazanie substrátov

Aktívne miesta enzýmov majú rozličné sekvencie, štruktúry a fyzikálne vlastnosti. Tieto črty zohrávajú dôležitú rolu, vo vytváraní väzby medzi enzýmom a rozličnými substrátmi. Atribúty aktívneho miesta sú [27]:

- **Veľkosť a tvar aktívneho miesta** sú prispôsobené tvaru substrátu, ktorý sa naväzuje na dané aktívne miesto.
- **Polarita** molekuly spôsobuje spájanie polárnych molekúl s ďalšími polárnymi molekulami ako aj spájanie nepolárnych s nepolárnymi. Vďaka tomu môžu určité aminokyseliny, nachádzajúce sa v aktívnom mieste, priťahovať alebo odpudzovať rozličné časti substrátu a tak vytvoriť vhodnejšiu konformáciu.
- **Náboj** navzájom opačné náboje sa priťahujú. To znamená, že pozitívny náboj bude priťahovaný k negatívnemu a negatívny náboj bude priťahovaný k pozitívnemu. Ak sa stretnú zhodné náboje, budú sa navzájom aktívne odpudzovať. Pomocou odpudzovania rovnakých nábojov vzniká ďalšia možnosť ako správne napasovať enzýmy k substrátom alebo častiam substrátom.
- **Hydrofobicita a hydrofilicita** v tomto prípade sa opačné charakteristiky nepriťahujú, priťahujú sa rovnaké. To znamená, že hydrofobické aminokyseliny sú priťahované s ďalšími hydrofobickými molekulami a hydrofilické aminokyseliny sú priťahované s hydrofilickými substrátmi.



Obr. 3.1: Na obrázku je znázornený princíp funkcie aktívneho miesta. (1) Receptor s aktívnym miestom a ligand pred vytvorením väzby. (2) Ligand je naviazaný na aktívne miesto, čím sa jemne pozmenil tvar receptoru. (3) V aktívnom mieste prebehla reakcia, do ktorej vstupoval ligand ako produkt, výsledkom sú dva produkty. (4) Produkty reakcie sú uvoľnené z aktívneho miesta a receptor sa vracia do počiatočného stavu.

- **Špeciálne vlastnosti kofaktorov** napomáhajú enzýmom naviazať sa na substráty. Najčastejšie ide o vitamíny a minerály, ako v prípade B vitamínov, využívaných produkovaní energie. Z tohoto dôvodu mnoho energetických doplnkov obsahuje vitamín B.

3.1.2 Model indukovaného prispôsobenia

Uvádza, že aktívne miesto a substrát nemusia mať plne komplementárny tvar aby sa naviazali. Keď ide o priraďovanie substrátu k aktívnemu miestu, ktoré nemajú rovnaké tvary, neodpudia sa, ale prispôsobia sa. Jeden, druhý alebo obaja zmenia svoj tvar tak, aby do seba zapadli. V tomto modeli ide o neustálu interakciu aktívneho miesta a substrátu, ktorý dáva substrát do jeho novej formy. Po skončení reakcie je vytvorený produkt, ktorý je následne oddelený od proteínu, nakoľko už nie sú spolu kompatibilné.

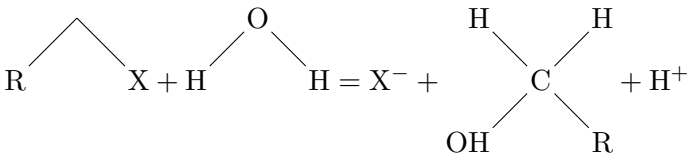
3.1.3 Enviromentálny vplyv

Môže ovplyvniť aktívne miesta receptoru, ako aj mieru výskytu chemických reakcií. Zvyšovanie teploty okolia vo všeobecnosti zvyšuje pravdepodobnosť chemických reakcií. Zrýchľuje pohyb molekúl, čo robí ich stret pravdepodobnejší.

Avšak zvyšovanie alebo znižovanie teploty okolia z optimálneho rozmedzia môže ovplyvniť chemické väzby formujúce proteín. To vedie k zmene jeho tvaru. Po zmene tvaru receptoru sa aktívne miesta nemusia naväzovať na správny ligand a výskyt reakcií môže byť znížený. Dramatické zmeny teploty a pH eventuálne vedú k denaturácii receptoru [6].

3.1.4 Príklad katalyzovanej reakcie

Haloalkán dehalogenáza (ďalej už len HLD) je globulárny proteín. Obsahuje konzervovanú hlavnú doménu a flexibilnú doménu s flexibilným obmedzením prístupu k aktívnemu miestu. Aktívne miesto sa nachádza sa vnútri proteínu a je spojené s povrchom pomocou prístupového tunelu. HLD je enzým, ktorý katalyzuje chemickú reakciu 3.1. V tejto reakcii katalyzuje hydrolytické štiepenie väzieb uhlík-halogén v halogénovaných alifatických zlúčeninách, čo vedie k tvorbe zodpovedajúcich primárnych alkoholov, halogenidových iónov a protónov. Vyjadruje aktivitu halogenázy proti 1-chlóralkánom s dĺžkou reťazca C3 až C10 a tiež vykazuje veľmi slabú aktivitu s 1,2-dichloretánom [38].



(3.1)

3.2 Tunely a kanály

Kanál sa od tunelu v proteínoch odlišuje tým, že prechádza celým proteínom. Má dva otvory, vstup a výstup. Tunel obsahuje iba jeden otvor. Kanály sú charakterizované dvoma otvormi spájajúce rozdielne bunkové prostredia, zohrávajú kľúčovú rolu v transporte iónov a malých molekúl cez biomembrány. Vyskytujú najmä v globulárnych proteínoch s katalytickou funkciou (enzýmy) a slúžia ako vstupy pre substráty, produkty, molekuly vody a inhibitory do skrytých aktívnych miest. Môžu napojiť na dva rozdielne aktívne miesta vo vnútri jedného proteínu. Bolo experimentálne dokázané, že tunely a ich vlastnosti dokážu definovať mnoho dôležitých charakteristík proteínov ako je špecifickosť substrátu, enantioselektivita, stabilita a aktivita. Tunely môžeme definovať na základe nasledujúcich kritérií [36].

- **Priemer najužšieho miesta** opisuje najužšie miesto v tuneli, obmedzuje aká najväčšia molekula je schopná prechodu cez tunel.
- **Dĺžka** definuje vzdialenosť od počiatočného bodu tunela po povrch proteínu.
- **Zakrivenie** opisuje tvar tunelu ako pomer medzi dĺžkou tunelu a najkratšou možnou vzdialenosťou medzi začiatočným a konečným bodom tunelu.
- **Priepustnosť** vyjadruje pravdepodobnosť, že sa cesta používa za účelom transportu ligandov.

3.3 Ligandy

Sú molekuly, ktoré sa naväzujú na proteíny. V určitých prípadoch väzba ligandu spôsobuje zmenu tvaru proteínu. Takéto zmeny v konformácii sú nevyhnutné pre mechanizmus proteínu a dôležité pre regulovanie jeho aktivity. V chémii pod pojmom ligand rozumieme atóm alebo molekulu, pripojenú na centrálny atóm, najčastejšie atómu kovu. Atómy a molekuly vystupujúce ako ligandy sa často správajú ako darcovia elektrónového páru v elektrónovej väzbe vytvorenou s atómom kovu. Príklady bežných ligandov sú neutrálne molekuly vody (H_2O), amoniaku (NH_3), oxidu uhoľnatého (CO), aniónov kyanidu (CN^-), ... ale taktiež môžu byť aj katióny (NO^+, $N_2H_5^+$) a príjemcovia elektrónového páru. Ligandy môžu byť prírodné, čiže ako organické alebo anorganické molekuly, alebo syntetické, čiže vyrobené v laboratóriach. Naviazanie sa ligandu záleží na dvoch charakteristikách proteínu [29].

- **Špecifickosť**
  Schopnosť proteínu naviazať jednu molekulu alebo malú skupinku molekúl uprednostnených pred všetkými ďalšími molekulami.

- **Afinita / Príťažlivosť**
  Afinita určuje čas, počas ktorého je ligand napojený na jeho receptor alebo na špecifický proteín. Poukazuje na silu väzby, je vyjadrená disociačnou konštantou Kd.

Obe charakteristiky proteínu závisia na štruktúre väzbového miesta ligandu. Jednou z najviac skúmaných vlastností ligandov je ich schopnosť napájania protilátok na antigény. Zameriava sa na skupinu proteínov a enzýmov, na ktoré sa naväzujú ligandy vedúce ku katalýze chemických reakcií, nevyhnutných na prežitie a funkciu buniek. Proteíny, ktoré katalyzujú chemické reakcie (vytváranie a rozbíjanie kovalentných väzieb) sa nazývajú enzýmy. Protilátky sa naväzujú na antigény, enzýmy sa naväzujú na reaktanty nazývané substráty. Tie sa následne menia chemickými reakciami na produkty. Ligandy sú schopné naväzovať sa na proteíny, vďaka ich schopnosti naviazať sa na väzobné miesto proteínu [27, 29].

3.3.1 Funkcie ligandov

Funkcie ligandov je možné rozdeliť do nasledujúcich skupín [21].

- **Katalýza**
  Enzýmy reagujú so substrátmi a menia ich na produkty. Rovnako ako substráty, tak aj produkty môžu byť ligandmi. Sú to malé molekuly, peptidy alebo makromolekuly. Vybrané substráty slúžia ako kofaktory. Napríklad NADH a $FADH_2$ sú kofaktory pre viaceré *redoxné reakcie*[1], ktoré sú katalyzované enzýmami.

- **Regulácia**
  Časť organických molekúl je používaných na reguláciu aktivity metabolických enzýmov, proteínov signálnej transdukcie alebo ďalších kľúčových proteínov. Regulácia môže byť jednoduchá, ako napríklad v prípade inhibície produktu používaných pri metabolických cestách alebo sofistikovaná ako v prípade bunkových procesov aktivovaných hormónmi. Sofistikovaná regulácia zväčša zahŕňa komplex signálnych ciest, ktoré aktivujú alebo potláčajú niekoľko cieľov. Príkladom regulácie prostredníctvom inhibície produktu je použitie ATP v úlohe inhibitora spolu s niekoľkými metabolickými enzýmami.

- **Komunikácia**
  Ligandy sa môžu vyskytovať na rôznych miestach pozdĺž bunkových komunikačných dráh: prvý *messenger*[2] (hormón, neurotransmitter, lokálny mediátor), sekundárny messenger(cAMP, IP 3) a downstream regulátor. Činnosť týchto ligandov môže viesť k rozličným výsledkom ako je napríklad bunkový rast, delenie, biosyntéza metabolitov alebo aktivácia obrannej funkcie.

- **Preprava**
  Určité ligandy slúžia ako prostriedky, ktorými sú organely alebo ďalšie makromolekuly rozpoznávané proteínmi. Ako príklad naväzovania na cytoplazmický enzým proteínu kanázy C (PKC), malý ligand diaglycerolu (DAG) umožňuje enzýmu uchytiť sa na

[1] *Redoxné reakcie* sú chemické reakcie, pri ktorých sa menia oxidačné čísla atómov
[2] Messenger v literatúre nazývaný aj posol, bol zvolený anglický z dôvocu jeho zaužívania.

plazmovú membránu, aktivovať sa a zúčastniť sa v hlavnom procese prenosu signálu. Podobne nukleotidová sekvencia na začiatku génu (promotor) umožňuje proteínom fungovať ako transkripčné faktory na rozpoznávanie polohy a aktiváciu špecifického génu. Ďalšie nukleotidové sekvencie sú rozpoznané proteínmi, ktoré sa podieľajú na replikácii DNA a spracovaní RNA.

- **Protetické skupiny**
  Vybrané ligandy sa pevne naväzujú na proteíny a pomáhajú im vykonávať určité funkcie. Napríklad hemo skupina obsahujúca železo, ktorá sa napája na kyslík v hemoglobíne a myoglobíne, pokiaľ v cytochrómoch naväzuje elektróny. Ďalšia protetická skupina, retinálna, umožňuje svetlom aktivovať proteín rhodopsínu a bacteriorhodopsínu.

- **Obrana a útok**
  Niektoré ligandy sa správajú ako toxíny, útočia na ďalšie bunky. Pri baktériách sú vylučované a použité pri obrane proti ďalším baktériám alebo útoku proti hostiteľovi. Toxíny sú taktiež produkované vyššími organizmami (rastliny, hmyz, hady), v ktorých prípade môžu byť použité na odstrašenie predátorov alebo chytenie obete.

3.3.2 Činnosť ligandov

V ľudskom tele sa ligand prenáša pomocou tekutín (spoločne s krvou, v tkanivách alebo v samotnej bunke) nachádzajúcich sa v organizme. Po pripojení ligandu na proteín nastane zmena v konformácii. Ak sa doposiaľ nevytvorila alebo nezrušila chemická väzba, ligand naväzujúci sa na proteín zmení tvar celej štruktúry. Reverzibilita väzby medzi ligandom a proteínom je kľúčový aspekt pre všetky formy života. Pokiaľ by sa ligandy napojili opačne, nemuseli by slúžiť ako mediátory, a tým by značná časť biologických procesov prestala fungovať. Ligandy majú veľmi široké spektrum využitia.

- **Kyslík**
  V krvnom obehu a v tkanivách je nutné zabezpečiť kyslík pre všetky mitochondrie, aby mohol organizmus prežiť. Organizmi musia obsahovať systém, ktorý bude zaisťovať distribúciu kyslíka v tele. Väčšina využíva na jeho distribúciu obehový systém. Na prenos kyslíka slúžia konkrétne proteíny, človek a ostatné cicavce na to využívajú hemoglobín (hlavný proteín krvi zodpovedný za prenos kyslíka).

- **Dopamín**
  Je ligand využívaný najmä v mozgu. Dopamín nám signalizuje pôžitok z úspechu, čiže sa naväzuje na vnem motivácie. Pokiaľ sú ním recepory vyplnené, človek začne pociťovať eufóriu. Efektivita týchto ligandov sa môžeme zvýšiť pomocou drog ako kokaín a metamfetamín. Aj tento efekt vedie k závislosti na látkach, ktoré danú eufóriu vyvolajú.

- **Inzulín**
  Umožňuje aby sa glukóza získaná z potravy dostala do vnútra bunky, a tam bola premenená na energiu. V pečeni zvyšuje zachytávanie glukózy z krvi a podporuje tvorbu zásobného glykogénu.

- **Acetylcholín**
  Mozog využíva na prenos nervových impulzov.

Ligandy sú nevyhnutné pre kontrolovanie metabolizmu a ďalších komplexných procesov v organizmoch Regulačné ligandy majú za úlohu aktivovať enzýmy. Pokiaľ nebudú aktivované, nebudú mať správny tvar na transformáciu molekúl, na ktoré pôsobia. [26, 27, 29].

Kapitola 4

Problematika molekulárneho dokovania

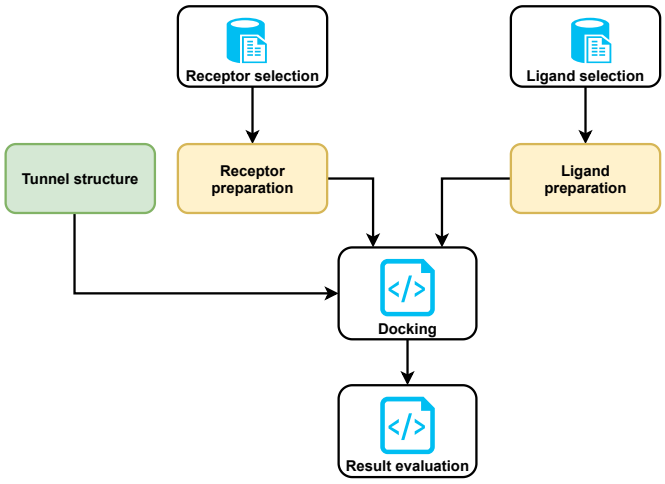
Zaoberá sa modelovaním interakcie medzi dvoma chemickými látkami, najčastejšie medzi proteínom a nízkomolekulárnym ligandom. Skúmanie týchto interakcií, je hlavnou metódou štruktúrovaného návrhu liečiv. Schopnosť presne a rýchlo vypočítať zmenu voľnej väzbovej energie by znamenala revolúciu. Otvorila by možnosť testovať milióny zlúčením ešte pred finančne a časovo náročnými laboratórnymi experimentami. Táto vymoženosť je však veľmi vzdialená. Aj napriek nedostatkom nástrojov, ktoré vykonávajú danú analýzu sa toto odvetvie rozvíja čím zefektívňuje vývoj liečiv a ďalších látok.

V diagrame 4.1 sú naznačené základné kroky pri molekulárnom dokovaní, ktoré sú spoločné pre všetky protokoly. Molekulárne dokovanie sa zaoberá vyhľadávaním najvhodnejších väzbových módov ligandu a receptoru. Režim viazania ligandu vzhľadom na receptor môže byť jedinečne definovaný jeho stavovými premennými. Tie pozostávajú z pozície (reprezentovanej x,y,z koordinátmi), orientácie (Eulerove uhly alebo osové uhly), v prípade flexibilného ligandu aj jeho konformácie (torzné uhly pre každú z otočných väzieb). Každá z týchto premenných popisuje jeden *stupeň voľnosti*[1] vo viacrozmernom vyhľadávacom priestore a ich hranice popisujú rozsah hľadania. Dokovanie tuhého telesa je rýchlejšie ako počítanie s flexibilitou ligandu. Veľkosť vyhľadávacieho priestoru je oveľa menšia pri dokovaní tuhého telesa ako flexibilného. To môže spôsobiť irelevantnosť získaných výsledkov, pokiaľ konformácia ligandu nie je správna. Dokovanie ligandu ako tuhého telesa znižuje pravdepodobnosť identifikácie komplementárnej zhody väzobného miesta a ligandu [33].

4.1 Príprava proteínu a ligandu

Procesu samotného dokovanoia predchádza príprava vhodnej reprezentácie receptoru a ligandu, vrátane tautomerov, stereoisomerov pri fyziologickom pH. Voľba vhodnej reprezentácie proteínu predstavuje zásadnú úlohu, ktorá môže skresliť výsledok dokovania. Základná príprava proteínu začína odobratím všetkých ligandov a molekúl vody, s výnimkou nevyhnutných kofaktorov a molekúl vody, ktoré sa konzistentne nachádzajú v aktívnych častiach proteínu a sú zahrnuté v naväzovaní ligandu. Pokiaľ proteín funguje ako monomér, sú odobrané všetky ďalšie proteínové reťazce vyskytujúce sa v asymetrickej jednotke PDB súboru. Molekula ligandu je konvertovaná do 3D podoby. Následne sa vykoná energická minimalizácia pre zaistenie korektnej dĺžky väzieb, usporiadania väzieb a väzbových uhlov.

[1] *Stupne voľnosti* určujú počet parametrov definujúcich stav fyzikálneho systému.



Obr. 4.1: Pracovný postup pri molekulárnom dokovaní pozostáva z niekoľkých častí. Prvým krokom je výber receptoru a ligandu. Následne príprava vhodnej reprezentácie receptoru a ligandu, niektoré nástroje môžu požadovať súbor reprezentujúci tunel vedúci k aktívnemu miestu receptora. Ďalší krok je spustenie dokovacieho algoritmu. Posledný krok je evaluácia výsledkov dokovania pomocou skórovacej funkcie.

4.1.1 Chýbajúce slučky

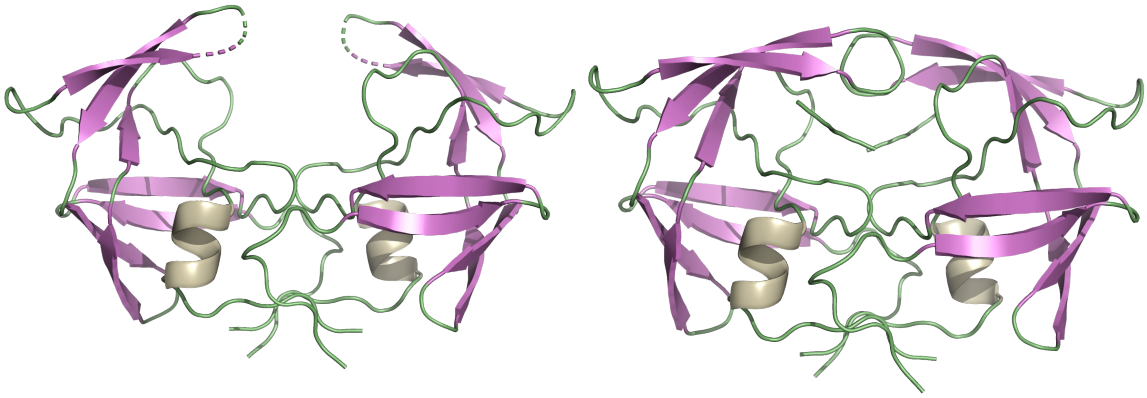
Pokiaľ sú v proteíne identifikované medzery, začína sa procedúra dostavby chýbajúcich slučiek. Tieto medzery sú vytvárané pri Röntgenovej kryštalografii, ktorá sa spolieha na získavanie kryštálov z veľkého množstva proteínov v identických pozíciach, čo spôsobuje, že flexibilné časti proteínov spôsobujú problém, ktorý vidíme na obrázku 4.2a. Regióny, ktoré sa pohybujú, nie sú vo všeobecnosti zachytené. Čo znamená, že sú vynechané z pdb záznamu proteínu. V ďalšom kroku sú pridané atómy vodíka a optimalizovaná sieť vodíkových väzieb.

4.1.2 Asymetrické jednotky a biologické zhluky

V kryštáloch použitých pri Röntgenovej kryštalografií je niekoľko kópií proteínu alebo/a nukleovej kyseliny naskladaných za sebou v poli. Zvyčajne len najmenšia časť bez duplikátov zvaná asymetrická jednotka je uložená v PDB archíve. Biologicky relevantný zhluk asymetrických jednotiek, reprezentujúci molekulu môže byť kompletne odlišný od asymetrickej jednotky. Operácie symetrie pre dogenerovanie biologicky relevantnej molekuly sú poskytnuté v PBD zázname, alebo je možné stiahnuť koordináty pre biologický zhluk z archívu PDB [4].

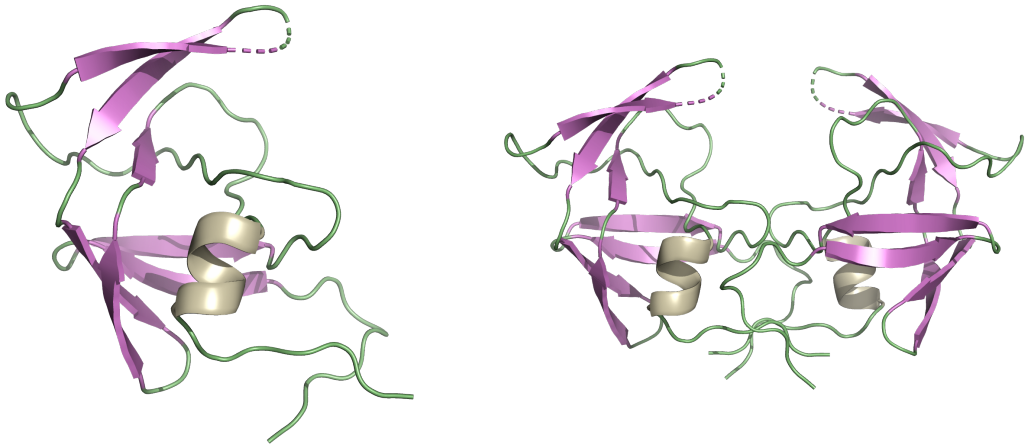
4.1.3 Identifikácia aktívneho miesta

Identifikácia aktívneho miesta je dôležitou súčasťou molekulárneho dokovania. Pri chybnej anotácii aktívneho miesta sú výsledky dokovania irelevantné. Je niekoľko možností identifikácie aktívnych miest. Pre vybraný enzým avšak nemusia byť dostupné všetky možnosti.



(a) 1AZ5 (b) 1YTI

Obr. 4.2: Štruktúra SIV proteázy vyriešená bez aktívnych miest (PDB záznam 1az5) má dve slučky, ktoré sú veľmi flexibilné a teda neboli zachytené experimentom. Avšak keď bol proteín kryštalizovaný pomocou inhibitora, slučka nadobudla stabilnú štruktúru, ktorú je možné vidieť v PDB zázname 1YTI



(a) 1AZ5 - asymetrická jednotka (b) 1AZ5 - biologický zhluk

Obr. 4.3: PDB záznam 1AZ5 obsahuje len jeden reťazec, ktorý nie je biologicky aktívny, nazývaný asymetrická jednotka je zobrazený vo figúre 4.3a, avšak vo figúre 4.3b vidíme biologicky aktívny dimér, ktorý sa skladá z dvoch asymetrických jednotiek.

- **Nástroj na detekciu aktívnych miest** Je možné použiť softvérový nástroj, ktorý identifikuje aktívne miesta v štruktúre proteínu. Jedným z možných riešení je nástroj GASS-WEB [30], ktorý využíva evolučný algoritmus na hľadanie podobností medzi aktívnymi miestami v proteínoch.
- **Databáza aktívnych miest** Jedna z najpoužívanejších databáz je "Catalytic Site Atlas" (CSA), ktorá dokumentuje aktívne miesta enzýmov a katalytické reziduá v 3D štruktúre enzýmov. Klasifikuje iba katalytické reziduá, ktoré sa priamo podieľajú nejakým aspektom na katalýze chemickej reakcie enzýmom. CSA obsahuje 2 typy záznamov:
  - ručne anotované záznamy, odvodené z primárnej literatúry. Referencie k tejto literatúre sú poskytnuté.
  - homologické záznamy, nájdené pomocou porovnávania sekvencií s jedným z originálnych záznamov. Sú ekvivalentné reziduá, ktoré sú zarovnané ku katalytickým reziduám v sekvencii z originálneho záznamu.
- **Slepé dokovanie** Je typ dokovania, ktorý je používaný na detekciu možných väzbových miest peptidových ligandov pomocou skenovania celého povrchu cieľového proteínu. Metóda je taktiež používaná pre objektívne mapovanie väzbových vzorov ligandov pri návrhu liečiv [16].

4.2 Flexibilita proteínu

Naväzovanie ligandu na proteín obvykle zahŕňa konformačné zmeny v štruktúre proteínu zvané indukované prispôsobenie, ktoré je možno vidieť v obrázku 3.1, ktoré sa pohybujú od lokálnych preusporiadaní bočných reťazcov až po pohyby veľkých domén. Kvôli veľkosti proteínov a veľa stupňov voľnosti proteínov, ich flexibilita je významný problém pri molekulárnom dokovaní. Aktuálne prístupy k flexibilite môžu byť rozdelené do štyroch kategórií.

4.2.1 Mäkké dokovanie ("Soft Docking")

Mäkké dokovanie je najjednoduchšia metóda, ktorá berie flexibilitu proteínu implicitne. Dovoľuje malý stupeň prekrytia medzi ligandom a receptorom pomocou zmenšovania vplyvu interatomických Van der Waalsvých interakcií počas výpočtu dokovania. Výhodou mäkkého dokovania je jeho výpočtová efektivita a jednoduchosť implementácie. Avšak mäkké dokovanie berie v úvahu len malé konformačné zmeny [13].

4.2.2 Flexibilita bočného reťazca ("Side-Chain Flexibility")

Prístup sa zameriava iba na flexibilitu bočných reťazcov. Kostra receptoru je ponechaná fixná a konformácie bočných reťazcov sú vzorkované. Približne 90% aktívnych miest pri vytváraní väzby s ligandom podstúpi aspoň jednu transformáciu. Zanedbanie flexibility bočných reťazcov vedie k výsledkom, kde pri skoro tretine prípadov sterické strety zabránia naviazaniu ligandu na aktívne miesto [13].

4.2.3 Molekulárna relaxácia ("Molecular Relaxation")

Molekulárna relaxácia začína dokovaním ligandu do aktívneho miesta receptoru bez ohľadu na flexibilitu. Následne sú kostra proteínu a okolité bočné reťazce uvoľnené. Počiatočné

dokovanie do fixnej štruktúry dovoľuje kolízie atómov medzi proteínom a umiestneným ligandom, podľa ktorých je následne možné zvážiť konformačné zmeny. Sú vytvorené komplexy relaxované a minimalizované pomocou metódy Monte Carlo, molekulárnej dynamickej simulácie alebo iných metód.

Výhoda metódy molekulárnej relaxácie je zahrnutie určitej flexibility kostry receptoru k flexibilite bočných reťazcov. V porovnaní s metódou Flexibility bočného reťazca je metóda molekulárnej relaxácie výpočetne náročnejšia pre skórovaciu funkciu, keďže zahŕňa nielen pohyb bočných reťazcov ale aj komplikované vzorkovanie kostry receptoru. To môže spôsobiť artefakty pri nepresnostiach skórovacej funkcie [13].

4.2.4 Dokovanie Súboru proteínov ("ensemble docking")

Je najpoužívanejší prístup k začleneniu flexibility proteínu, pri ktorom je použitá skupina proteínových štruktúr na reprezentáciu rozdielnych konformačných zmien, ktoré sa môžu vyskytnúť. Jedna z prvých štúdii bola vykonaná Knegtelom a jeho tímom [22], v ktorej bola zostrojená spriemerovaná energetická sieť pomocou spájania energetických sietí vygenerovaných z individuálnych experimentálne analyzovaných proteínových štruktúr. S využitím váhovej schémy, nasledovanej štandardným molekulárnym dokovaním. Osterberg s tímom rozšírili metódu a vytvorili nástroj AutoDock [31] s početnejšou skupinou pozostávajúcou z 21 rozdielnych konformácii HIV-1 proteázy. Priemerovací potenciál metódy môže strácať určitú presnosť pri geometrii proteínu.

Claussen s tímom vyvinuli nástroj FlexE na dokovanie ligandov do zoskupenia proteínových štruktúr [10]. V nástroji sú podobné segmenty proteínovej štruktúry zarovnané, zatiaľ čo odlišné segmenty sú použité na kombinatorické vytvorenie nových možných konformácií pre dokovanie. V algoritme vyvinutom Wei-om s tímom, je proteín rozložený na fixnú časť a niekoľko flexibilných častí podľa kryštalografických štruktúr súboru proteínov. Pre dané umiestnenie ligandu sa ukladá iba najlepšia konformácia pre každú flexibilnú časť proteínu s predpokladom, že flexibilné regióny sa pohybujú nezávisle. Vybrané lokálne konformácie sú spojené s rigidnou časťou a vytvoria takzvanú *best-fit*[2] konformáciu proteínu. V porovnaní s FlexE je tento algoritmus signifikantne rýchlejší a časovo sa škáluje lineárne na rozdiel od exponenciálneho škálovania v algoritme FlexE. Avšak kvalita výsledkov algoritmu závisí od kvality konfomačného vzorkovania ligandu, či bolo vybrané správne počiatočné umiestnenie ligandu.

Huang a Zou vyvinuli rýchly súborový dokovací algoritmus, v ktorom berú konformačný súbor ako dodatočnú dimenziu k tradičným 6 stupňom voľnosti (3 translácie a 3 rotácie) pre optimalizáciu pohybu ligandu [18, 23]. Algoritmus má porovnateľnú rýchlosť ako jednotkové dokovanie a zachováva presnosť postupného dokovania.
Algoritmus dokovania súboru proteínov nie je použitý na generovanie nových proteínových štruktúr, ale slúži k výberu takzvaného "induced-fit" štruktúry z poskytnutého súboru štruktúr. Abagyan s tímom rozšírili Huang a Zou algoritmus na vytvorenie ICM súborového dokovacieho algoritmu, referovaného ako štvor-dimenzionálne (4D) dokovanie [5]. Narozdiel od využitia len NMR štruktúr alebo kryštalografických štruktúr, algoritmus využíva súbory proteínových konformácií generovaných pomocou molekulárnych dynamických simulácií, Monte Carlo simulácií alebo predikciou štruktúry. [19]

[2] *Best-fit* je konformácia, ktorá je najlepšie prispôsobená na dokovanie ligandu

4.3 Dokovanie ligandu do proteínu

V tejto fáze sa snažia dokovacie algoritmy predpovedať väzobnú konformáciu proteínu a ligandu. Sú zavedené rôzne pravidlá a heuristické obmedzenia, keďže nie je možné vyskúšať všetky konformácie molekúl. Pre vytvorenie množiny *póz*[3] existuje mnoho prístupov, ktoré ale môžu byť rozdelené do systematických a stochastických metód.

4.3.1 Zhoda s tvarom

Je jedna z najjednoduchších metód vzorkovacích algoritmov, ktorá je často využívaná v počiatočných fázach dokovacieho procesu alebo v prvom kroku pokročilejších metód vzorkovania ligandu. Umiestňuje ligand s kritériom, ktoré hovorí, že molekulárny povrch umiestňovaného ligandu musí byť komplementárny k povrchu väzbového miesta proteínu. Šesť stupňov voľnosti (tri translácie a tri rotácie) dovoľuje mnoho možností uloženia. Úlohou algoritmu zhody s tvarom je teda čo najrýchlejšie vyhľadať najvhodnejšiu komplementárnu polohu pre ligand vo väzobnom mieste. Príklady dokovacích programov využívajúcich túto metódu sú DOCK, MDock, LigandFit a mnoho ďalších. Hlavnou výhodou algoritmu zhody tvarom je jeho výpočtová efektivita. Avšak, konformácia ligandu je väčšinou fixná počas výpočtu. Z toho dôvodu, flexibilné dokovanie ligandu je obvykle vykonávané pomocou dokovania celého súboru predgenerovaných konformácií ligandu v proteíne, nasledované spájaním póz zo všetkých behov dokovacieho algoritmu podľa ich energetického skóre.

4.3.2 Systematický prístup

Sú využívané pri flexibilnom dokovaní ligandu, počas ktorého sa generujú všetky možné väzbové konformácie ligandu pomocou skúmania všetkých stupňov jeho voľnosti. Metódy môžeme rozdeliť do troch kategórii:

- **Konformačné metódy** sa snažia systematicky rotovať všetky rotovateľné väzby, pokiaľ nie sú vygenerované všetky existujúce kombinácie. Nevýhodou týchto metód je enormný nárast počtu vygenerovaných štruktúr so zvyšujúcim sa počtom rotovateľných väzieb. Tento fenomén je známy ako kombinačná explózia. Aplikácia konformačných metód je teda značne obmedzená a k redukcii konformačného priestoru ligandu sú použité ďalšie aproximačné metódy.
- **Fragmentačné metódy** redukujú konformačný priestor ligandu tak, že dokujú len jeho fragmenty do aktívneho miesta receptoru a tam ich kovalentne spájajú do iniciálnej štruktúry ligandu, alebo rozdelia ligand na rigidné fragmenty, ktoré sú dokované ako prvé. Flexibilné časti sú pripájané tak, aby vytvárali štruktúru ligandu.
- **Databázové metódy** využívajú databázy s definovanými konformačnými stavmi ligandu, ktoré sú následne dokované ako rigidné štruktúry do aktívneho miesta receptoru.

4.3.3 Stochastický prístup

Stochastické algoritmy prehľadávajú konformačný priestor pomocou náhodných modifikácií stavových premenných. V súčasnosti sa využívajú Monte Carlo metódy, genetické algoritmy, Tabu metódy ako aj metódy optimalizácie rojom.

[3] *Póza* opisuje vzájomnú pozíciu proteínu a ligandu.

- **Monte Carlo metódy** Replica-exchange Monte Carlo (REM) simulačná metóda, ktorá definuje proteín-ligand ako rigidnú štruktúru, z ktorej sa v každom kroku metódy vytvorí markovovský reťazec dokovacích konformácií pomocou náhodného pohybu počiatočnej konformácie ligandu. V každom kroku je daná konformácia dostatočne akceptovaná s pravdepodobnosťou:

$$P_{local} = min\left[1, exp\left(-\frac{\Delta E}{kT}\right)\right] \tag{4.1}$$

  Kde:

  - ΔE je rozdiel energie medzi novou a pôvodnou konformáciou
  - kT reprezentuje teplotný parameter

  V danom prístupe je zrejmé, že by sa samostatná metóda Monte Carlo mohla jednoducho zastaviť na lokálnom minime na čo sa využije protokol REM, ktorý tieto lokálne minimá prekoná [39].

- **Genetické algoritmy** Aplikujú idey mendelovskej genetiky a Darwinovej evolučnej teórie na problém dokovania. Každý GA začína s populáciou, ktorá reprezentuje množinu možných riešení problému. Špecifická konformácia ligandu reprezentuje jedinca v populácii a je definovaná súborom stavových premenných (gény), ktoré popisujú aspekty ako translácia, orientácia a konformácia ligandu vzhľadom na proteínový receptor. Celkový súbor stavových premenných ligandu definuje *genotyp*[4], pokiaľ atómové koordináty definujú *fenotyp*[5]. Schopnosť fenotypu prežiť v určitom prostredí závisí od jeho stavu, ktorý je daný celkovou energiou interakcie ligandu a receptoru, vyhodnotenou príslušnou skórovacou funkciou. Správny výber jedincov z generácie na základe vyhodnotenia nám zaistí, že populácia bude konvergovať ku globálnemu optimu. Pri dokovaní, genetický algoritmus hľadá globálne minimum v konformačnom priestore ligandu. Z vybraných jedincov sú náhodne vytvorené dvojice, ktoré si s určitou pravdepodobnosťou vymenia gény a vytvoria nového jedinca. Tento proces sa nazýva "crossover". ďalším procesom, ktorý prebieha pri vytváraní nasledujúcej generácie je mutácia. Mutácia je náhodne vyvolaný proces, počas ktorého sú gény jedinca náhodne zmenené a vytvoria nového jedinca. Globálne prehľadávanie pri stochastických algoritmoch však nemusí vždy nájsť energeticky najvýhodnejší väzbový mód, preto bol vyvinutý hybridný globálno-lokálny prehľadávací algoritmus, v programe AutoDock 3.0 nazvaný ako lamarckovský genetický algoritmus [15].

- **Tabu metódy** Môže sa jednať o samostatnú implementáciu tabu metódy ako je popísané v článku [3] kedy:

  1. Vytvorí sa náhodné riešenie a uloží sa ako aktuálne
  2. Vyhodnotí sa, pokiaľ je najlepšie tak sa nahrá
  3. Obnoví sa tabu list
     - Ak tabu list nie je plný (<25 členov) tak sa pridá aktuálne riešenie.
     - V inom prípade sa vymení najstaršie riešenie za aktuálne.

[4] *Genotyp* predstavuje genetickú informáciu, ktorá sa vyskytuje v určitom organizme vo forme DNA.
[5] *Fenotyp* je súhrn všetkých vonkajších znakov a vlastností jedinca, realizácia genotypu.

4. Vygeneruje a vyhodnotí sa N (100) možných pohybov z aktuálneho riešenia. Využívaná Cauchyova mutácia s δ 0.075.
5. Riešenia sa zoradia vzostupne v závislosti na energii.
6. Vyhodnotia sa usporiadané riešenia:
   - Ak má pohyb menšiu energiu ako zatiaľ najlepší, akceptuje sa a pokračuje sa v bode 7.
   - Ak pohyb nie je tabu (>0.75 RMS v citovanej publikácii sa uvádza 0.75Å medzi dvoma porovnávanými riešeniami) tak sa akceptuje a pokračuje na bod 7.
   - Ak nie sú identifikované žiadne akceptovateľné pohyby, tak sa algoritmus terminujeme.
7. Ak iteračný limit bol dosiahnutý (1000), algoritmus je ukončený s najlepším nájdeným riešením. Ak riešenie nebolo zmenené posledných x (100) iterácii tak sa algoritmus reštartuje (bod 1.) Inak pokračujeme bodom 2.

V novších publikáciách [17] je možné vidieť aj kombináciu s inými metódami ako sú napríklad generické algoritmy.

- **Metódy optimalizácie rojom ("Swarm optimisation")** sa pokúšajú nájsť optimálne riešenie vo vyhľadávacom priestore modelovaním inteligencie roju. V tejto metóde sa pohyby ligandu v priestore riadia informáciami o najlepších pozíciach susedov. Medzi nástroje, ktoré využívajú túto metódu patrí napríklad SODOCK [7] a Tribe-PSO [8].

4.4 Skórovanie

Skórovacie funkcie patria do triedy výpočtových metód aplikovaných pri evaluácii proteín-ligand interakcií. Sú náhradou za fyzikálne výpočty, ktoré sú časovo veľmi náročné. Snažia sa využiť rôzne aproximácie, aby dosiahli kompromis medzi časom nutným na výpočet a presnosťou výsledkov. V literatúre ich môžeme nájsť rozdelené do sledujúcich tried:

4.4.1 Založené na silových poliach ("Force-field-based")

Väčšina stanovuje energiu z konkrétnej vzájomnej pozície ligandu a proteínu, takzvanej pózy. V literatúre boli často uvádzané ako funkcie založené na silových poliach. Jedno z najpoužívanejších polí, AMBER, má predpis:

$$E = \Delta E_v + \Delta E_u + \Delta E_d + \Delta E_w + \Delta E_e \tag{4.2}$$

kde E je celková energia pózy a sumy energetických príspevkov:

- ΔE_v z dĺžok jednotlivých väzieb.
- ΔE_u veľkosti väzbových uhlov.
- ΔE_d rotácie okolo dihedrálnych uhlov.
- ΔE_w van der Waalsových interakcí mezi jednotlivými pármi atómov.
- ΔE_e elektrostatických interakcí mezi jednotlivými atómovými pármi.

V dnešnej dobe je možné nájsť mnoho variácií fyzikálnych funkcií, v ktorých môžu byť zahrnuté úpravy pre vodíkové väzby alebo ďalšie energetické príspevky.

4.4.2 Empirické ("Empirical")

Empirické skórovacie funkcie predstavujú aproximácie zmeny gibbsovej voľnej energie (ΔG), ktorá je vyjadrená vzťahom:

$$\Delta G = \sum_i W_i \times \Delta G_i \tag{4.3}$$

kde ΔG_i sú jednotlivé príspevky k zmene voľnej energie a W_i sú ich váhy, ktoré sú nastavené pomocou viacnásobnej lineárnej regresie na základe experimentálnych dát získaných skúmaním afinity už známych ligandov. Typickým príkladom empirickej skórovacej funkcie je ChemScore:

$$ChemScore = S_H + S_m + S_l + P_r + P_s + Pc + [P_{cov} + P_{con}] \tag{4.4}$$

kde S značí pozitívne skóre za vodíkové väzby, koordinované väzby s iónmi kovov, lipofilné kontakty. Negatívne skóre sú označené P a zahŕňajú zmrazené otočné väzby, energiu vnútorného napätia ligandu, sterické strety medzi proteínom a ligandom. Najväčším problémom empirických metód je ich závislosť na experimentálnych dátach, ktoré sú často nevyvážené vzhľadom k rôznym proteínovým rodinám a pokrývajú iba obmedzenú časť chemického priestoru.

4.4.3 Založené na znalostiach o potenciáloch ("knowledge-based potential")

Všetky skórovacie funkcie zdieľajú charakteristiku sčítania po pároch atómov, štatistické potenciály medzi proteínom a ligandom:

$$A = \sum_i^{lig} \sum_j^{prot} \omega_{ij}(r) \tag{4.5}$$

Potenciál závisí na vzdialenosti medzi atómovým párom $i-j$, $\omega_{ij}(r)$, je odvodený z Boltzmanovej analýzy:

$$\omega_{ij}(r) = -k_B \times T \times ln\left[g_{ij}(r)\right] = -k_B \times T \times ln\left[\frac{\rho_{ij}(r)}{\rho^*_{ij}}\right] \tag{4.6}$$

kde $\rho_{ij}r$ je numerická hustota atómového $i-j$ páru vo vzájomnej vzdialenosti r. ρ^*_{ij} je numerická hustota rovnakého atómového páru v referenčnom stave, kde interatomické interakcie sú pokladané za nulové. S týmto prístupom sa frekvencia kontaktu daného páru meria z jeho energetickej kontribúcie pri navezovaní proteínu na ligand.

Ak sa konkrétna dvojica vyskytuje častejšie ako tá v referenčnom stave, pre príklad náhodná distribúcia, je indikáciou, že interakcia je energeticky výhodná. V opačnom prípade, keď je výskyt danej dvojice menší ako v referenčnom stave, tak je interakcia nepravdepodobná. Na odvodenie párového potenciálu sa vyberie dostatočne veľká množina proteín-ligand komplexov, napríklad z databázy PDB. Následne sú vypočítané potenciály závislé na vzdialenosti pre každý možný atomický pár z frekvencie výskytu danej atomickej dvojice v tréningovej sade pomocou vzorca 4.6. Nedokonalosť daného prístupu je uvažovanie, že ligand alebo proteín sú náhodné usporiadania atómov ako je tomu v tekutinách. Narozdiel od tekutín sú atómy usporiadané kovalentnými väzbami v určitom poradí, tým pádom referenčný stav vo vzorci 4.6 je v rozpore s reálnym referenčným stavom. V porovnaní s empirickou skórovacou funkciou sa funkcie založené na znalostiach snažia zachytiť všetky energetické faktory v proteín-ligand interakciach implicitne s párovými potenciálmi.

4.4.4 Založené na strojovom učení ("Machine-learning based")

Využívajú kvantitatívne vzťahy štruktúry a aktivity alebo QSAR ("quantitative structure-activity relationship") pri analýze proteín-ligand interakcií. QSAR analýza bola zaužívaná pri modelovaní rôznych fyzikálno-chemických, biologických a farmaceutických vlastností malých molekulárnych zlúčenín od počiatku počítačovo navrhovaných liečiv. Ak je možné vlastnosti ligandu a proteínu ako aj vzory ich interakcií zaznamenať do deskriptorov. Potom sofistikované techniky strojového učenia zahrnuté v QSAR analýze môžu byť aplikované na odvodenie štatistických modelov, ktoré vypočítajú skóre proteín-ligand interakcií. Skórovacia funkcia založená na strojovom učení je napríklad NNScore, ktorá využíva niekoľko desiatok neurónových sietí, z ktorých následne spriemeruje výsledky a vytvorí NNScore [12, 11]. SFCscore [40], ID-Score [25] sú ďalšie skórovacie funkcie založené na strojovom učení [28, 24].

Kapitola 5

Nástroj Caver Web

Je webový server, umožňujúci analýzu proteínových tunelov a kanálov ako aj transport ligandov týmito tunelmi. Nástroj je jednoduchý na používanie a poskytuje užívateľovi množstvo prednastavených parametrov, ktoré mu uľahčia prácu a poskytnú validné výsledky analýz.

5.1 Nástroje

Server je postavený na softvérových nástrojoch Caver 3.x a CaverDock 1.x umožňujúcich štúdium transportu ligandov. Nástroj Caver Web ako aj nástroje CaverDock a Caver sú vyvíjané výskumnou skupinou Loschmidt laboratories.

5.1.1 Caver

Caver je softvérový nástroj slúžiaci na analýzu a vizualizáciu tunelov v štruktúre proteínov. Najkritickejším krokom pri detekcii tunelu je výber relevantného počiatočného bodu. Pozícia vybraného bodu vymedzuje kalkuláciu nástroja Caver a definuje spoločný počiatočný bod pre všetky detekované tunely. Zle umiestnený bod môže výrazne ovplyvniť dôležitosť detekovaných tunelov a viesť k irelevantným tunelom. Aby sa tomu predišlo, bolo vytvorených niekoľko automatizovaných protokolov. Tie zaisťujú správne umiestnenie počiatočných bodov pre pokrytie čo najväčšieho okruhu možných prípadov riešeného problému. Čo najviac zaujíma užívateľov pri enzýmoch sú prístupové cesty ligandov vedúce k aktívnym alebo väzbovým miestam. Najlepším počiatočným bodom je miesto v katalytickom vačku (catalytic pocket), ktorý obsahuje základné reziduá. Katalytické vačky sú detekované pomocou nástroju Fpocket 2, založeným na hľadaní alfa sfér. Základné reziduá sú získané z "Mechanism and Catalytic Site Atlas" a databázy SwissProt. Manuálne upravená databáza SwissProt sa prehľadáva pomocou nástrooja BLAST, s podmienkou 30% identity sekvencie a medzi 90% až 110% dĺžky sekvencie. Po identifikácii základných reziduií, katalytické vačky sú spojené s týmto reziduami. Vačky, obsahujúce minimálne jedno katalytické reziduum, sú označené ako katalytické. Ak základné reziduá chýbajú, Caver Web ponúkne niekoľko možností pomoci. Ako prvé Caver vypíše všetky detekované vačky a zoradí ich podľa odhadovaného skóre druggability. Ako druhú možnosť Caver Web umiestni počiatočný bod do centra masy ľubovoľného ligandu, nachádzajúceho sa v štruktúre. Ako poslednú možnosť Caver Web ponúkne vypočítať pozíciu počiatočného bodu, založeného na reziduách vybraných užívateľom v proteínovej sekvencii, ktorá neskôr môže byť manuálne nastavená.

Detekcia tunela sa vykonáva pomocou Caver 3.02, ktorý vyhľadáva cesty s daným minimálnym polomerom a najnižšími nákladmi pomocou Voronoiovho mozaikovania proteínovej štruktúry. Tento algoritmus analyzuje tunel a počíta geometrie, štatistické vlastnosti, vypisuje reziduá tvoriace tunel a formuje *bottleneck*[1]. Používatelia môžu upravovať niekoľko dôležitých parametrov v nastavení ovplyvňujúcich vlastnosti detekovaných tunelov [9]. Výsldekom výpočtu je tunel reprezentovaný sekvenciou sfér, ktorý je vizualizovaný vo figúre 5.1.

- **Možné reziduá pre kalkuláciu tunela**
  Časti štruktúry, ktoré Caver zváži pre analýzu, aby sa dali vylúčiť ligandy, ióny a molekuly vody.
- **Minimálny polomer sody**
  Minimálna veľkosť guľovitej sondy, ktorá prejde tunelom tak, aby bol správne detekovaný.
- **Hĺbka obalu**
  Maximálna hĺbka - plytké vrcholy zamedzujúce zbytočnému vetveniu tunela.
- **Polomer obalu**
  Špecifikuje polomer sondy využívajúcej sa na definovanie, ktoré časti Voronoiho diagramu reprezentujú objemové rozpúšťadlo.
- **Prah podobnosti**
  Definuje level podobnosti, pri ktorej sa tunely budú považovať za rovnaké a spolu zoskupené.
- **Maximálna vzdialenosť**
  Určuje limit ako ďaleko môže byť Voronoiov vrchol vzdialený od počiatočného bodu zvoleného užívateľom.
- **Žiadúci polomer**
  Špecifikuje nutnú vzdialenosť medzi počiatočným bodom a atómov proteínovej štruktúry.

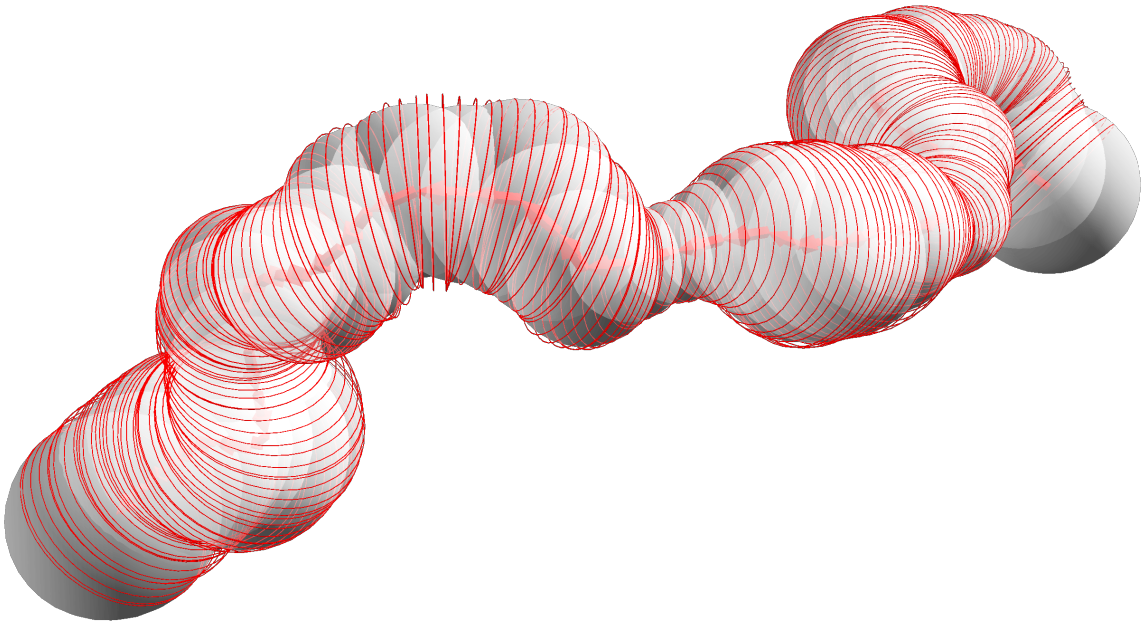
5.1.2 CaverDock

Posledným voliteľným krokom je analýza transportu ligánd detekovaným tunelom pomocou nástroja CaverDock, ktorý umožňuje rýchly, presný a plne automatizovaný výpočet tunelov v statických a dynamických šruktúrach. Molekuly vhodné na analýzu CaveruDocku zahŕňajú proteíny, nukleové kyseliny alebo inorganické materiály. Caver Dock je nástroj na báze molekulárneho dokovania slúžiaci na analýzu transportu ligandov tunelmi a kanálmi proteínov. Celá metóda je postavená na postupnom pohybe ligandov pozdĺž tunelmi.

Užívateľ musí poskytnúť minimálne jednu molekulu slúžiacu ako ligand a minimálne jeden tunel, ktorý sa bude analyzovať ako cesta pre transport ligandov. V tomto bode sa začína kalkulácia. Užívateľ má možnosť upraviť dva najpodstatnejšie parametre.

- **Diskretizačná delta**
  Určuje vzdialenosť medzi centrami dvoch prierezov tunela.

[1] Bottleneck reprezentuje najužšie miesto tunelu.



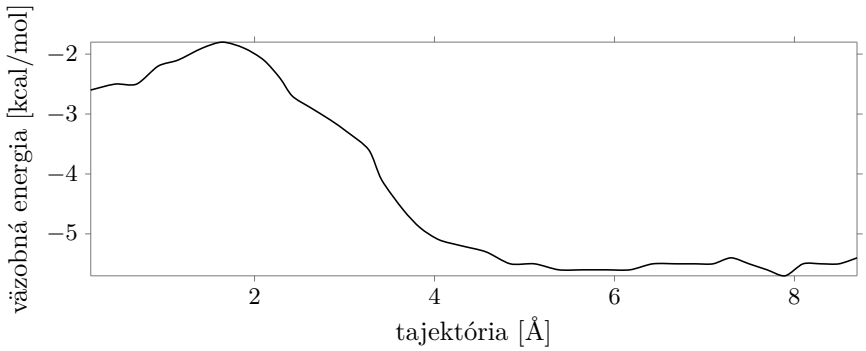
Obr. 5.1: Diskretizácia tunela získaného pomocou nástroja Caver pre toluén, červené kružnice predstavujú disky, červené šípky reprezentujú smer tunela. Šedé gule reprezentujú trav tunela získaného pomocou nástroja Caver [36].

- **Kalkulačný mód**
  Definuje, vyžadované obmedzenia ligandu.
  - **Spodná hranica (“lower-bound“)**
    Vyžaduje priestorové obmedzenie. Tento mód je veľmi rýchly aj keď môže vynechať nejaké bottlenecky kvôli možnosti preklopenia ligandu, čo vedie k nesúvislému pohybu.
  - **Vrchná hranica (“upper-bound“)**
    Využíva maximálnu rotáciu ligandu spojenú so spätným sledovaním, ktoré zaručuje neustály pohyb. Aj keď stály pohyb je viac realistickejší, analýza je výpočtovo signifikantne intenzívnejšia a kvôli limitovanej kapacite spätného sledovania môže preceniť energie alebo úplne zlyhať pri hľadaní akejkoľvek možnej cesty. Tým pádom je dolná hranica trajektórie predvolená. Užívateľom je odporúčané používať energetické profily vypočítané v tomto móde.

Ako vstup tunela využíva tunelovú geometriu modelovanú guľovými sekvenciami. Guľové sekvencie môžeme získať z nástrojov, ktoré poskytujú PDB súbor tunela reprezentovaný guľami. Jedným príkladom je Caver 3.02, pre ktorý sa výsledný formát súboru CaverDocku optimalizuje.

1. Guľové sekvencie sú rozdelené do sekvencií diskov (prierez plátkov maximálnej hrúbky určenej používateľom). Ako prvé sú zvolené atómy ligandu umiestnené na disk pomocou vzájomného obmedzenia pozície.

2. CaverDock s využitím skórovacej funkcie z nástroja AutoDock Vina minimalizuje konformáciu ligandov a vyhodnotí ich (binding) voľnú energiu.



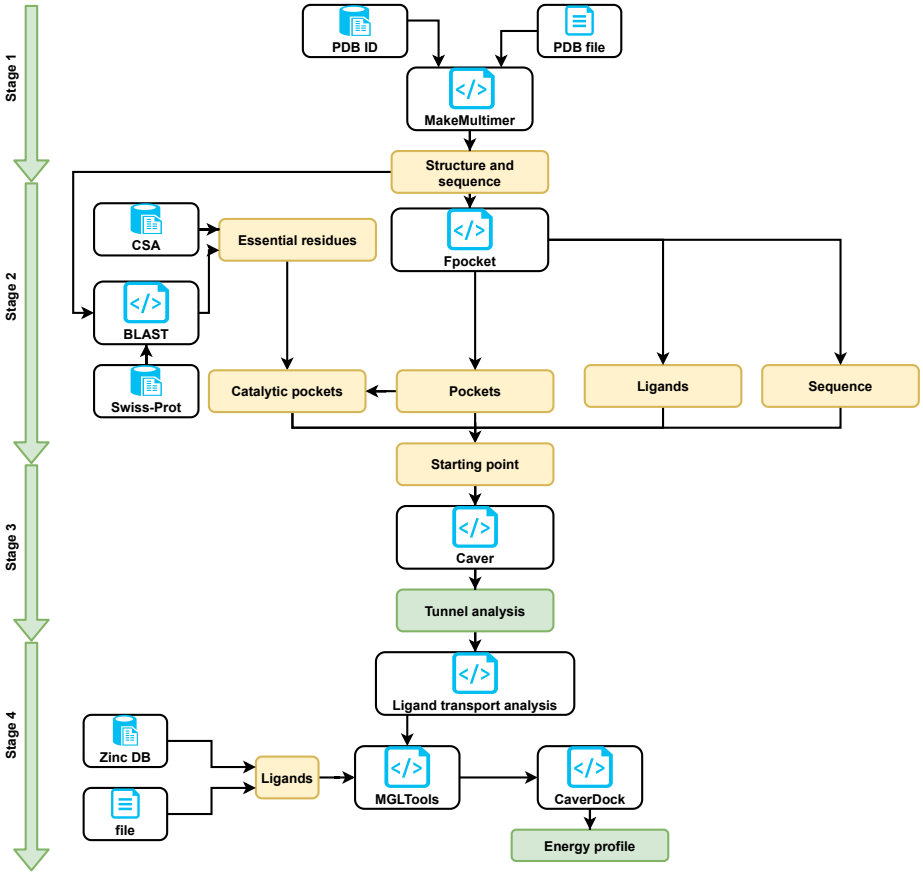
Obr. 5.2: Energetický profil vytvorený na základe dokovania ligandu do receptora pomocou nástroja Caver Web.

3. V treťom kroku sa vytvára trajektória ligandu pomocou zhromaždenia dokovaných pozícií ligandu na každom nasledujúcom disku. Takáto trajektória dôkladne vzorkuje tunel, kde pohyb ligandu môže byť nekontinuálny. Táto nekontinuálna dolná hranica trajektórie je používaná ako aj najnižšia možná hranica energetického profilu transportu ligandu cez tunel. Zvýšenie energie je možné dosiahnuť pokiaľ sa nekontinuálny pohyb náhlymi zmenami v orientácii alebo konformácii ligandu vyhne malým prekážkam. Finálne sa použije obmedzenie vzorom na výpočet kontinuálnej trajektórie. Výsledkom sú hodnoty z ktorých je možné získať graf 5.2.

V každom kroku je ligand uložený v blízkosti predchádzajúcej pozície, čo nám zaručuje len malé zmeny v konformácii ligandu. Počet možných kontinuálnych trajektórií stúpa exponenciálne spolu s počtom diskov, dôvodom je každá zmena vedúca k zmene pozície, orientácie alebo konformácie ligandu. Preto sa používa heuristická metóda na hľadanie kontinuálnej trajektórie. Spätné sledovanie ("Backtracking") je aktivované pokiaľ je voľná energia danej konformácie výrazne vyššia ako voľná energia konformácie získanej z dolnej hranice trajektórie. Nakoľko nie je istotu, že výsledná kontinuálna trajektória je optimálna, nazýva sa ako horná hranica trajektórie. Dôvodom je možnosť, že skutočná energia môže byť nižšia ako vypočítaná energia. Niekoľko základných rozdielov medzi hornou hranicou trajektórie a dolnou hranicou trajektórie: dolná hranica trajektórie je schopná urobiť kompletný vzor trajektórie ligandu. Informácie získané z dolnej hranice trajektórie sú dostačujúce na porovnávanie, jedinou limitáciou je možnosť vynechania malej prekážky z dôvodu náhlych zmien v orientácii ligandu [37].

5.2 Pracovný postup

Celý pracovný postup pri používaní Caver Webu pozostáva zo štyroch fáz. Štruktúry ligandov môžu byť nahraté vo všetkých formátoch, ktorý podporuje Open Babel [34]. Ako prvý krok výpočtu je výber štruktúry proteínu a jej predspracovanie. V druhom kroku sa volí začínajúca pozícia pre detekciu tunelov. Identifikácia a analýza proteínových tunelov je tretím krokom. Vo štvrtom, voliteľnom kroku, sa študuje transport zvolených ligandov.



Obr. 5.3: Graf pracovného postupu nástroja Caverweb.

5.3 Detekcia tunelov

5.3.1 Vstup

Vstup pre Caver Web je Terciárna štruktúra proteínu. Tá môže byť špecifikovaná identifikátorom Protein Data Banky, nahraná v podobe PDB súboru alebo v CIF formáte. nahraté štruktúry sú automaticky konvertované do PDB formátu pomocou *RCSB MAXIT*[2] nástroja, nakoľko CIF formát nie je doposiaľ podporovaný nástrojom Caver. Štruktúry sú väčšinou uložené ako asymetrické časti, ktoré neovplyvňujú ich kvaternálnu formu. Analýza asymetrických jednotiek alebo nesprávnej konformácii biologicky aktívnych zhlukov môže viesť k irelevantným výsledkom. Na získanie týchto zhlukov sa využíva nástroj MakeMultimer, ktorý automaticky po zadaní štruktúry detekuje jej biologické zhluky [36].

5.3.2 Výstup

Používateľ môže špecifikovať prioritnú *úlohy*[3] na jednoduchšiu orientáciu medzi zvolenými úlohami. Upozornenia o statuse výpočtov môžu byť zaslané na zvolenú e-mailovu adresu. Všetky úlohy sú uschované a prístupné kedykoľvek na vygenerovanej adrese. Akonáhle je úloha zvolená, nástroj Caver vypočíta tunely a konečná analýza je ukázaná na stránke. V sekcii informácie o úlohe sú užívateľovi poskytnuté informácie ako identifikátor a názov spolu možnosťou priameho stiahnutia niekoľkých súborov [36].

- **PyMol session**
  sťahovanie predgenerovaného súboru pre vizualizáciu pomocou nástroja *PyMol*[4]. Obsahuje nahranú štruktúru proteínu a všetky detekované tunely ponúkajúce užívateľovi podrobnejšiu vizuálnu analýzu alebo kvalitné obrázky.

- **Zip výsledky**
  sťahovanie archívu obsahujúceho nespracované dáta vygenerované nástrojom Caver počas kalkulácie. Dáta môžu byť použité na pokročilé analýzy alebo môžu byť priamo importované do nástroja Caver Analyst [20].

- **Caver konfigurácia**
  otvorí pop-up okno s kompletnou konfiguráciou súboru použitého na kalkuláciu

- **Caver log**
  otvorí pop-up okno s nespracovaným textovým výstupom nástroja Caver, ktorý poskytuje detaily o procese kalkulácie

V sekcii, informácie o tuneli sa nachádza list všetkých identifikovaných tunelov a ich vybraných vlastností.

5.3.3 Profil tunelu

Pop-up okno profilu tunelu nám umožňuje porovnávať analýzy rôznych ďalších profilov. Používatelia môžu zvoliť jeden alebo viac tunelov z tabuľky. V takomto prípade sú grafy

[2] RCSB MAXIT je nástroj na spracovanie makromolekulárnych štruktúr dostupný na stránke https://sw-tools.rcsb.org/apps/MAXIT/index.html.

[3] *Úloha* reprezentuje jednu zadanú kalkuláciu. V databáze a zdrojových textoch je úloha označovaná ako job.

[4] *PyMol* je nástroj umožňújúci vizualizáciu molekulárnych štruktúr dostupný na webovej stránke https://pymol.org/2/.

automaticky vygenerované. Každý údajový bod je interaktívny a umožňuje nám výber správnej sféry z vizualizácie tunelu. Grafy s informáciami sa dajú stiahnuť v CSV súbore alebo ako PNG obrázok [36].

5.3.4 Vizualizácia tunelu

Proteín a všetky detekované tunely môžu byť vizualizované priamo vo webovom prehliadači pomocou *JSmol applet*[5]. Užívateľ má niekoľko možností vizualizácie, aby mu boli výsledky ľahko a efektívne interpretované. Súčasťou výstupu sú aj jeho dôležité charakteristiky ako sú polomer najužšieho miesta, dĺžka, zakrivenie a priepustnosť. Priepustnosť je vypočítaná ako 5.1, kde e je Eulerovo číslo a cost je funkcia 5.2 [36].

$$e^{-cost} \tag{5.1}$$

$$\int_0^L r\,(l)^{-2}\,dl \tag{5.2}$$

Kde:

- L značí dĺžku tunelu.
- $r\,(l)$ je funkcia definujúca polomer najväčšej gule, ktorá sa nijako nezrazí s atómami štruktúry.

Každý tunel je vidieť vo vizuálnej podobe pokiaľ je označené korešpodujúce políčko a obraz priblížený pomocou ikony lupy. Keď je vybraná ikona knihy a grafu, sú otvorené detaily o tuneli a profil tunelu sa zobrazí ako pop-up okno.

5.4 Dokovanie proteínu

5.4.1 Vstup

Posledná sekcia výstupnej stránky z kalkulácie tunelu je zameraná na (voliteľnú) analýzu transportu ligandov cez tunely. CaverWeb umožňuje ligand nahrať ako mol2 alebo PBD súbor, vložiť ligand ako text vo formáte PDB, PDBQT, MOL alebo ho nakresliť.

5.4.2 Výstup

Výstup nástroja CaverDock je vygenerovaný formou dvoch PDBQT súborov. Jeden súbor obsahuje hornú hranicu trajektórie, zatiaľ čo druhý súbor nám ponúka dolnú hranicu trajektórie ligandu. Informácie o energiách väzieb a polomeroch tunela sú na riadkoch označených "REMARK" trajektórií ligandu. Tie môžu byť extrahované a vykreslené pomocou skriptov uložených v balíčku.

[5] *JSmol applet* je dostupný na webovej stránke http://jmol.sourceforge.net/

Kapitola 6

Implementácia a výsledky

Táto kapitola je zameraná na opis implementácie modulu hromadného dokovania ako aj použitého datasetu a jeho predpríprava 6.1. Upravená časť pracovného postupu je ukázaná na obrázku 6.1. V sekcii venovanej výsledkom 6.2 sú opísané pre užívateľa dôležité informácie súvisiace s modulom ako aj demonštrovaný formát výstupu vykonanej analýzy.

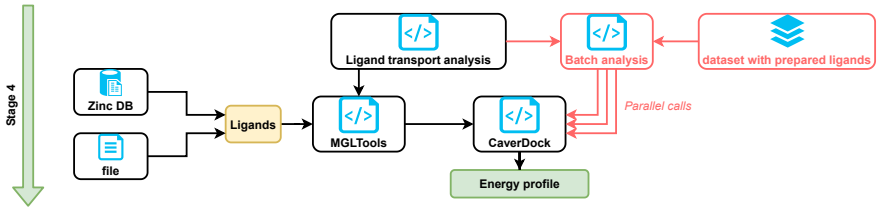
6.1 Implementácia

Nástroj je pozostáva z niekoľko častí ako sú výpočetné jadro, frond-end a back-end. Back-end nástroja Caver web komunikuje s Výpočtovým jadrom. Jadro vytvára príkaz spustenia skriptu a následne čaká na jeho ukončenie. Po jej vytvorení je úloha, spustená a pridaná do čakacej rady. Keď sa uvolnia výpočetné prostriedky vyžadované ulohou tak sa začína proces výpočtu, prebiehajúci na *cloude*[1].

6.1.1 Použitý dataset

Ako dataset ligandov, ktoré reprezentujú schválené liečivá (minimálne jednou zo svetových liekových agentúr) bol vybraný dataset "world" z verejne dostupnej databázy ZINC 15 [35]. ZINC obsahuje databázu a súbor nástrojov, ktoré poskytujú zlúčeniny používané pre virtuálne testovanie. Zásadnou výhodou ZINC 15 je množstvo datasetov, ktoré je možno jednoducho získať a integrovať do softwarového riešenia. Modul teda bude ľahké prispôsobiť podľa potreby aj pre ďalšie datasety. Použitý dataset bol stiahnutý a predspracovaný tak, aby pred hromadným výpočtom nebolo nutné stiahnutie a príprava všetkých ligandov, čím

[1] *Cloud* je server, ku ktorému je pristupované pomocou internetového pripojenia, ako aj software a databázy na tomto serveri.



Obr. 6.1: Upravená 4. časť pôvodného pracovného postupu nástroja Caverweb 5.3

sa zefektívni využívanie procesorového času ako aj sieťový prenos. Predspracovanie ligandov je vykonávané pomocou skriptu prepare_ligand4.py, ktorý je súčasťou MGL tools[32], Rodzdiel medzi stiahnutou molekulou z databazy a predspracovanou molekulou je možné vidieť na obrázku 6.3.

- **Molekulárne vlastnosti**
  Vlastnosti sú pripisované molekulám na základe členstva v niektorom z katalógov. Sú rozdelené do 4 biogenetických anotačných tried: biogenetické (prírodné produkty), metabolity, endogénne ľudské metabolity a neznáme. Ďalej sú zaradené do 8 bioaktívnych tried, ktoré zahŕňajú lieky: FDA approved, world drugs, v priebehu skúmania, v človeku, in vivo, v bunkách, in vitro, a neznáme. ZINC taktiež podporuje agregátory ako anotácie [1].

- **ZINC15 aplikačné rozhranie**
  Je skoro identické v porovnaní so štruktúrou URL webových stránok a poskytuje možnosť jednoduchej integrácie do aplikácií tretích strán. API podporuje 11 formátov ako sú PDBQT, MOL2, ... z 18 zdrojov [2]. Každý zdroj podporuje do 10 endpointov, tie dovoľujú definovať populárne podmnožiny zlúčenín, ktoré následne zjednodušujú syntax požiadaviek. Tieto rozdelenia do podmnožín sú dostupné na podstránke "subsets" [1].

6.1.2 Front-end

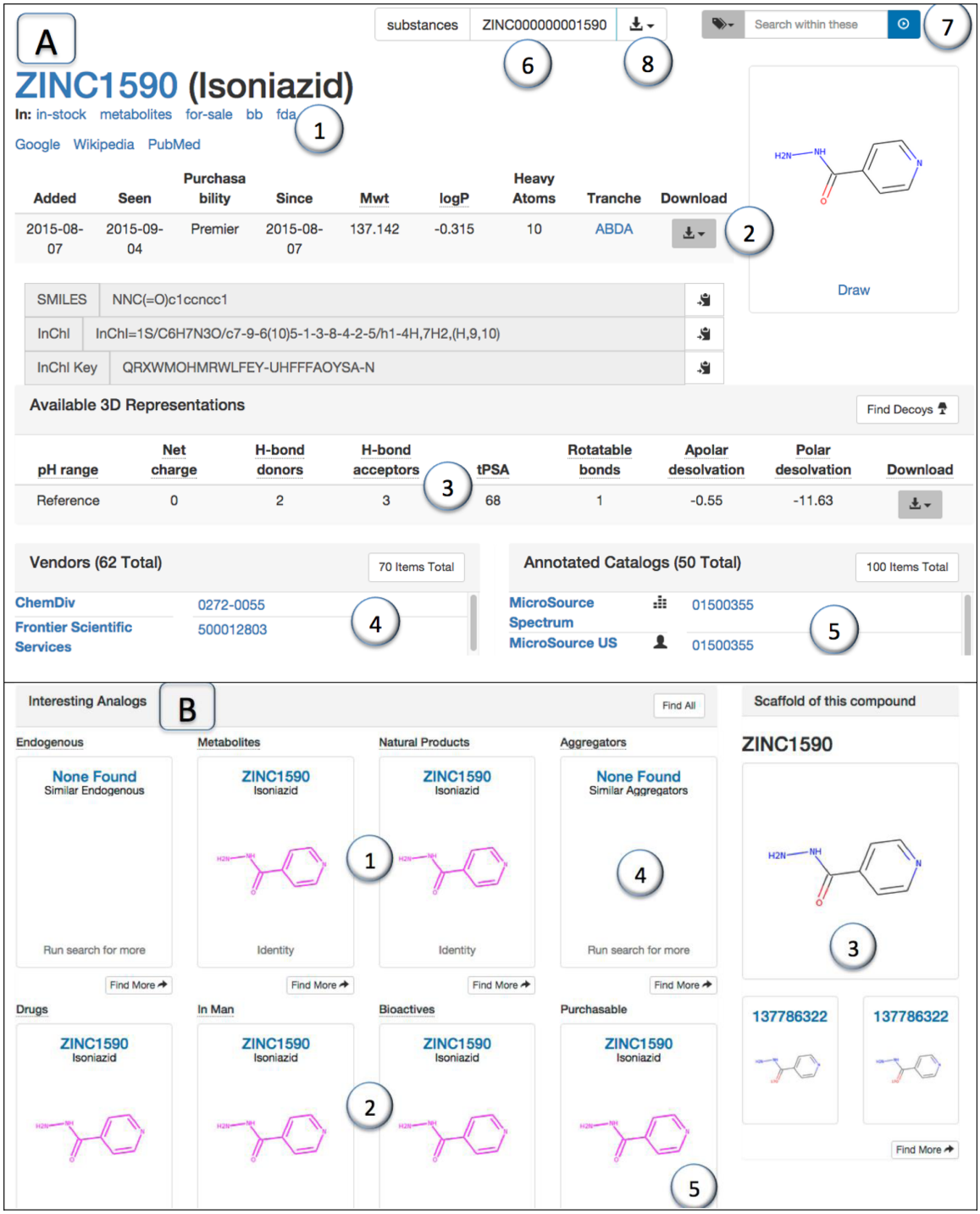
Bol implementovaný pomocou frameworku *Smart Google Web Toolkit (Smart GWT)*[2] a integrované do nástroja Caver Web. V tejto fáze analýzy bude mať užívateľ spracovaný receptor a v ňom identifikované tunely ako aj aktívne miesta. Je nutné zvoliť jeden z tunelov ako aj smer ligandov ako je ukázané na obrázku 6.4. Následne je požiadavka na výpočet zaslaná na spracovanie pre back-end.

Po vykonaní výpočtu si užívateľ môže prehliadať všetky ligandy v datasete zoradené podľa názu, energie pri prechode najužšiou časťou tunelu, priemernej energie pri prechode celým tunelom alebo názvu ligandu. Užívateľ má možnosť pokračovať v analýze energetických profilov vypočítaných pre jednotlivé ligandy.

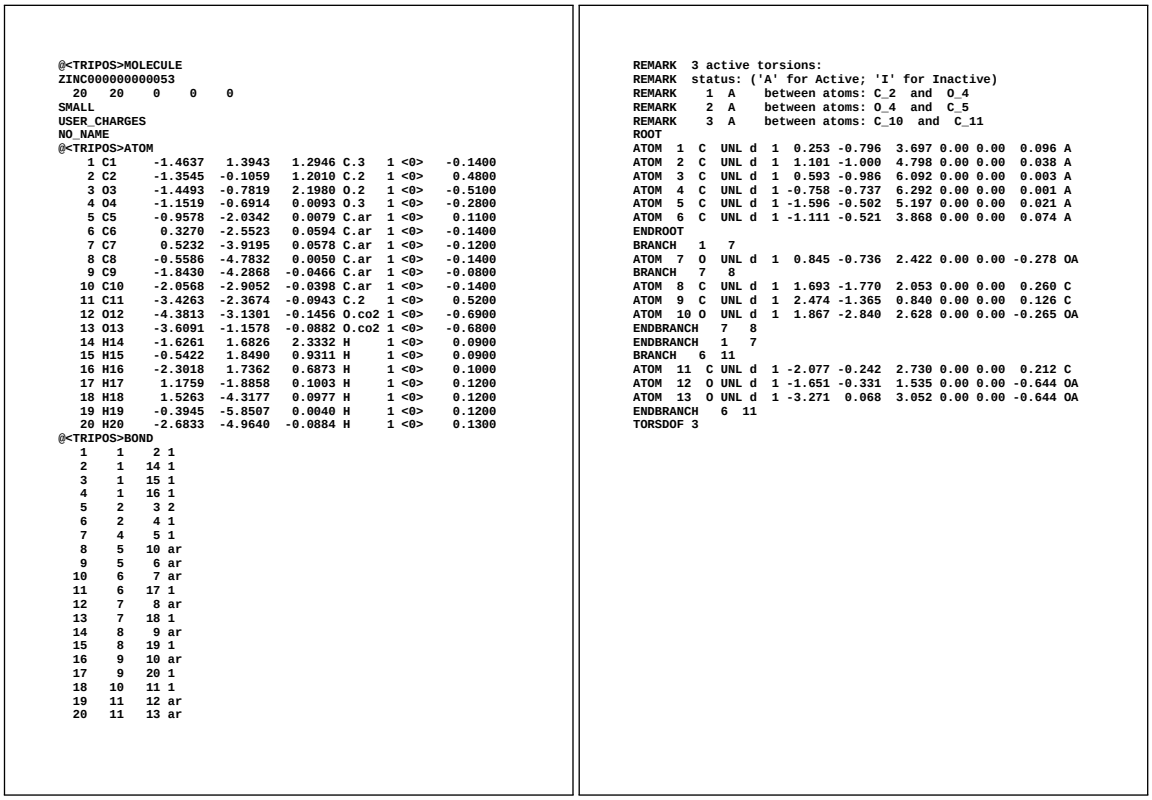
6.1.3 Back-end

Je tak ako front-end implementovaný v Smart GWT a slúži na komunikáciu medzi Front-endom a výpočtovým jadrom projektu. V back-ende bola pridaná funkcionalita, ktorá identifikuje úlohu ako hromadnú analýzu. Úlohe je pridelený jedinečný identifikátor. V MySQL databáze do tabulky úloh je pridaný záznam obsahujúci medzi inými vygenerovaný identifikátor úlohy, typ, cestu ku konfiguračnému súboru pre úlohu ako aj status kód. V tabulke jobs_meta sú uložené informácie ako identifikátor rodičovskej úlohy, názov proteínu, smer analýzy, ... Typy uložených metadát sa odlišujú na základe typu úlohy. Stĺpec job_id špecifikuje v tabulke job_meta konkrétnu úlohu, ku ktorej bude záznam pridaný. Meta_key je atribút, ktorý chceme danej úlohe priradiť a meta_value značí hodnotu atribútu. V prípade ukladania značiek v energetických profiloch pre jednotlivé ligandy sú záznamy dynamicky pridané keď užívateľ zvolí možnosť uložiť pri označovaní energetických profilov. V

[2]Smart GWT je framework programovacieho jazyka java poskytujúci jednoduchší vývoj webových aplikácií, domovská stránka frameworku sa nachádza na adrese https://www.smartclient.com/product/smartgwt.jsp.



Obr. 6.2: Detail molekuly vo webovom rozhraní databázy Zinc15. A. je zobrazené (1) ZINC ID, názov, a členstvo v podskupinách, (2) vlastnosti a 2D zobrazenie, (3) 3D reprezentácia, (4) informácie o možnosti kúpy, (5) anotovaný katalóg členstva, (6) aktuálna lokácia, (7) vyhľadávanie vo webovom portáli, (8) možnosť stiahnutia. B. Sekcia bioaktívnych a biogenických analógov molekuli (1) podobné biogenetické zlúčeniny, (2) podobné bioaktívne zlúčeniny, (3) zlúčeniny s rovankou kostrou, (4) podobné agregátory a (5) podobné zlúčeniny, ktoré je možno zakúpiť [35].



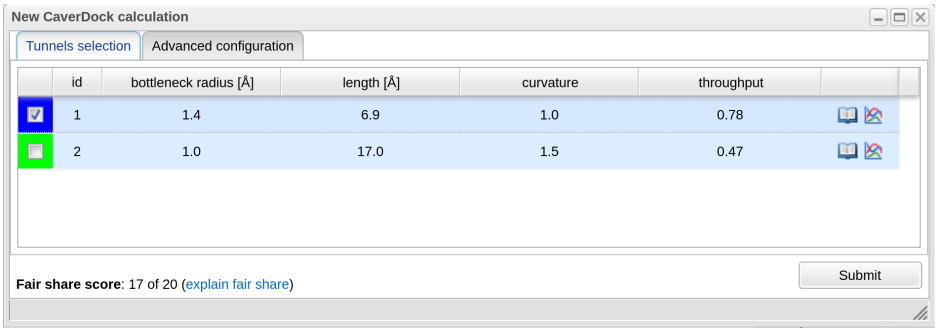
```
@<TRIPOS>MOLECULE
ZINC000000000053
  20   20    0    0    0
SMALL
USER_CHARGES
NO_NAME
@<TRIPOS>ATOM
    1 C1      -1.4637   1.3943   1.2946 C.3   1 <0>   -0.1400
    2 C2      -1.3545  -0.1059   1.2010 C.2   1 <0>    0.4800
    3 O3      -1.4493  -0.7819   2.1980 O.2   1 <0>   -0.5100
    4 O4      -1.1519  -0.6914   0.0093 O.3   1 <0>   -0.2800
    5 C5      -0.9578  -2.0342   0.0079 C.ar  1 <0>    0.1100
    6 C6       0.3270  -2.5523   0.0594 C.ar  1 <0>   -0.1400
    7 C7       0.5232  -3.9195   0.0578 C.ar  1 <0>   -0.1200
    8 C8      -0.5586  -4.7832   0.0050 C.ar  1 <0>   -0.1400
    9 C9      -1.8430  -4.2868  -0.0466 C.ar  1 <0>   -0.0800
   10 C10     -2.0568  -2.9052  -0.0398 C.ar  1 <0>   -0.1400
   11 C11     -3.4263  -2.3674  -0.0943 C.2   1 <0>    0.5200
   12 O12     -4.3813  -3.1301  -0.1456 O.co2 1 <0>   -0.6900
   13 O13     -3.6091  -1.1578  -0.0882 O.co2 1 <0>   -0.6800
   14 H14     -1.6261   1.6826   2.3332 H     1 <0>    0.0900
   15 H15     -0.5422   1.8490   0.9311 H     1 <0>    0.0900
   16 H16     -2.3018   1.7362   0.6873 H     1 <0>    0.1000
   17 H17      1.1759  -1.8858   0.1003 H     1 <0>    0.1200
   18 H18      1.5263  -4.3177   0.0977 H     1 <0>    0.1200
   19 H19     -0.3945  -5.8507   0.0040 H     1 <0>    0.1200
   20 H20     -2.6833  -4.9640  -0.0884 H     1 <0>    0.1300
@<TRIPOS>BOND
   1    1    2 1
   2    1   14 1
   3    1   15 1
   4    1   16 1
   5    2    3 2
   6    2    4 1
   7    4    5 1
   8    5   10 ar
   9    5    6 ar
  10    6    7 ar
  11    6   17 1
  12    7    8 ar
  13    7   18 1
  14    8    9 ar
  15    8   19 1
  16    9   10 ar
  17    9   20 1
  18   10   11 1
  19   11   12 ar
  20   11   13 ar
```

(a) Mol2 formát molekuly získaný z databázy Zinc ako súčasť datasetu pred prípravou.

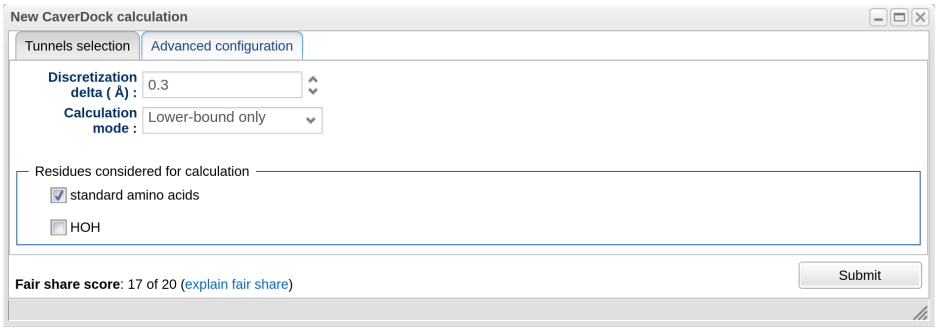
```
REMARK  3 active torsions:
REMARK  status: ('A' for Active; 'I' for Inactive)
REMARK    1  A    between atoms: C_2  and  O_4
REMARK    2  A    between atoms: O_4  and  C_5
REMARK    3  A    between atoms: C_10  and  C_11
ROOT
ATOM  1  C  UNL d  1  0.253 -0.796  3.697 0.00 0.00  0.096 A
ATOM  2  C  UNL d  1  1.101 -1.000  4.798 0.00 0.00  0.038 A
ATOM  3  C  UNL d  1  0.593 -0.986  6.092 0.00 0.00  0.003 A
ATOM  4  C  UNL d  1 -0.758 -0.737  6.292 0.00 0.00  0.001 A
ATOM  5  C  UNL d  1 -1.596 -0.502  5.197 0.00 0.00  0.021 A
ATOM  6  C  UNL d  1 -1.111 -0.521  3.868 0.00 0.00  0.074 A
ENDROOT
BRANCH   1   7
ATOM  7  O  UNL d  1  0.845 -0.736  2.422 0.00 0.00 -0.278 OA
BRANCH   7   8
ATOM  8  C  UNL d  1  1.693 -1.770  2.053 0.00 0.00  0.260 C
ATOM  9  C  UNL d  1  2.474 -1.365  0.840 0.00 0.00  0.126 C
ATOM  10 O  UNL d  1  1.867 -2.840  2.628 0.00 0.00 -0.265 OA
ENDBRANCH   7   8
ENDBRANCH   1   7
BRANCH   6  11
ATOM  11  C UNL d  1 -2.077 -0.242  2.730 0.00 0.00  0.212 C
ATOM  12  O UNL d  1 -1.651 -0.331  1.535 0.00 0.00 -0.644 OA
ATOM  13  O UNL d  1 -3.271  0.068  3.052 0.00 0.00 -0.644 OA
ENDBRANCH   6  11
TORSDOF 3
```

(b) Predpripravená molekula na dokovanie vo formáte pdbqt.

Obr. 6.3: Rozdiel medzi molekulou získanou z databázy ZINC15 vo formáte MOL2 a molekulou predpripravenou na dokovanie vo formáte PDBQT.



(a) Pri výbere tunelu sú užívateľovi zobrazené informácie získané pomocou nástroja Caver v predchádzajúcom kroku.



(b) Pokročilé nastavenia umožňujú užívateľovi vybrať atribúty výpočtu ako sú diskretizačná delta a mód výpočtu, kde sú dostupné varianty spodný alebo vrchnú medz.

Obr. 6.4: Pred spustením kalkulácie si užívateľ vyberie tunel, v ktorom chce ligandy analyzovať, zvyšné nastavenia sú predvyplnené odporúčanými hodnotami.

tomto prípade je nutné odlíšiť značky pre jednotlivé ligandy v datasete a z toho dôvodu je Meta_key zložený z názvu značky a názvu ligandu, pre ktorý sú ukladané. Následne je vytvorené HTTP spojenie s výpočtovým jadrom. Podľa špecifikácii úlohy je ďalej vytvorený korešpondujúci príkaz, ktorý bude vykonaný na vzdialenom serveri.

6.1.4 Výpočtové jadro

Je časť nástroja, ktorá sa stará o správu úloh. Monitoruje databázu a pokiaľ je pridaný záznam novej úlohy tak si na jeho základe stiahne konfiguračný súbor a PDB súbor s receptorom. Pomocné nástroje, ktoré sú určené na výpočet úlohy sú zabalené v *singularity*[3] kontajneroch. Výpočtové jadro zasiela požiadavok na cloud, ktorý následne spúšťa skripty na pridelenie výpočtových prostriedkov pre výpočet pomocou singularity image-u. Tie spracovávajú skupiny ligandov z predpripraveného datasetu. Dataset je rozdelený aktuálne do priečinkov po 20 ligandov. Úlohy pre jednotlivé priečinky s ligandmi sú vytvorené naraz. Tieto úlohy sú pre každý priečinok reprezentované skriptom, ktorý sériovo spracováva ligandy v určenom priečinku sériovo. Predtým sú a zaradené do rady, kde budú čakať na pridelenie požadovaných výpočtových prostriedkov. Tieto prostriedky sú poskytnuté ICRC cloudom, kde majú nástroje pod správou Loschmdt laboratories rezervovaných 1 500 jadier. Skript bude pokračovať v behu pokiaľ nebudú všetky úlohy spracované. Status úloh je zisťovaný každú minútu pomocou príkazu *qstat*[4].

6.2 Výsledky

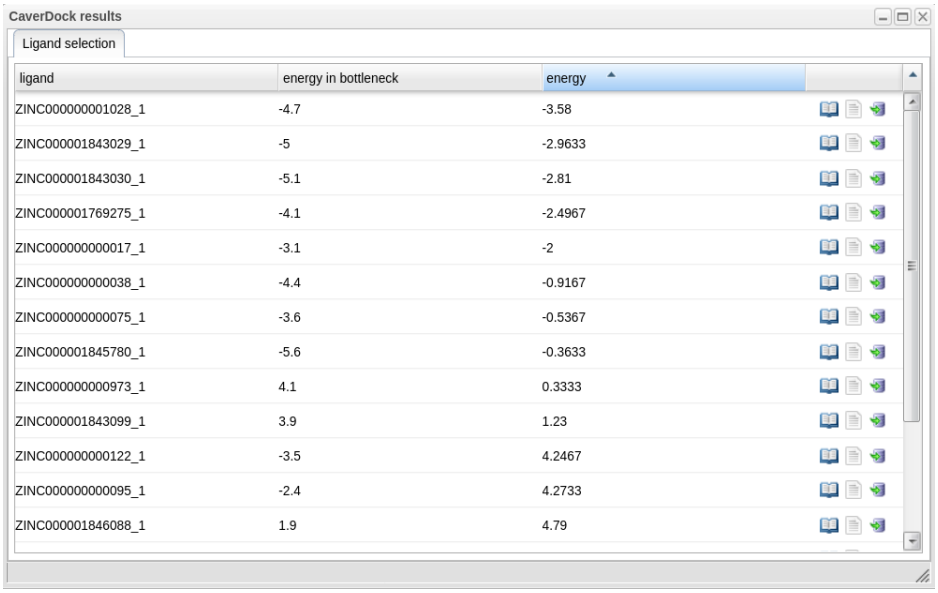
Modul bol integrovaný do nástroja CaverWeb a umožňuje paralelnú analýzu viac ako 3 000 ligandov na vybranom proteíne. Čas potrebný na spracovanie podľa testovania je približne 30–40 hodín, čo ale záleží na vyťažení výpočtových prostriedkov. Užívateľ môže výsledky hromadnej kalkulácie stiahnuť ako zip archív obsahujúci výsledky všetkých ligandov z datasetu, alebo zip archív, ktorý obsahuje kalkuláciu týkajúcu sa len konkrétneho ligandu. Ďalšou možnosťou je porovnávanie energetických profilov jednotlivých kalkulácií vyobrazené na vo figúre 6.5b priamo vo webovom rozhraní CaverWebu. Nástroj Caver Web je verejne dostupný na stránke https://loschmidt.chemi.muni.cz/caverweb/.

6.2.1 Výsledky analýzy

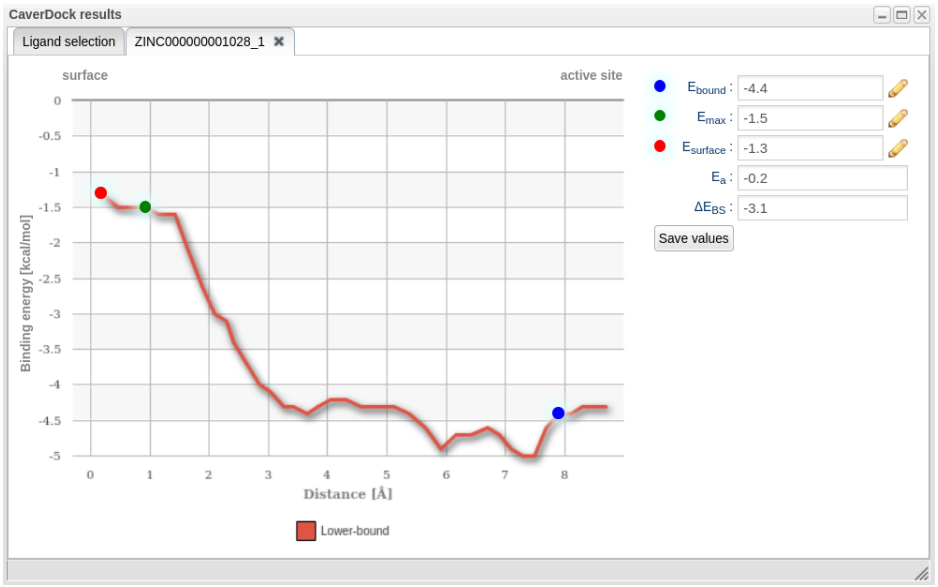
Výsledky analýzy sú užívateľovi sprostredkované vo forme listu ligandov 6.5a pre, ktoré bolo dokovanie vykonané. Počiatočne sa ligandy usporiadajú podľa najnižšej priemernej energie ale užívateľ ich má možnosť usporiadať taktiež podla ich názvu alebo energie v najužšom mieste tunelu. Výsledky zostávajú na pridelenej fixnej adrese a je možné k nim pristúpiť neskôr pomocou vygenerovanej webovej adresy. Po stiahnutí je možné proces dokovania analyzovať krok po kroku pomocou súboru profile-lb. Názorný príklad je možno vidieť na obrázku 6.6.

[3] Singularity je počítačový program, ktorý vykonáva virtualizáciu na úrovni operačného systému, nazývanú ako kontajnerizácia. Jedným z hlavných využití produktu Singularity je priniesť kontajnery a reprodukovateľnosť do vedeckých výpočtov a do sveta superpočítačov. Podrobnejší opis je dostupný na webstránke `https://sylabs.io/singularity/`.

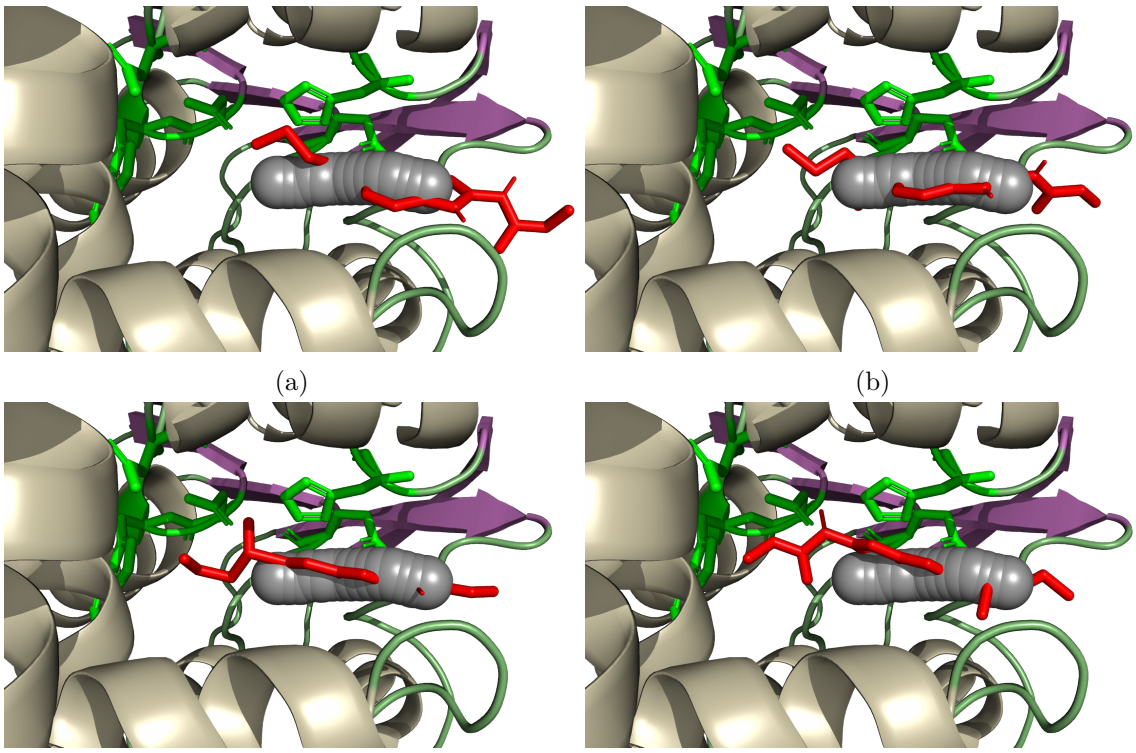
[4] *Qstat* je príkaz, pomocou ktorého je možné zistiť informácie o úlohách, manuál je dostupný na stránke `https://linux.die.net/man/1/qstat`



(a) Ligandy sú zoradené vzostupne, od najnižšej priemernej energie po najvyššiu. Užívateľ môže zvoliť taktiež zoradenie na základe energie v najužšom mieste ako aj podľa názvu ligandu.



(b) Po rozkliknutí ligandu je užívateľovi poskytnutý energetický profil, v ktorom môže označiť a uložiť v databáze dôležité energetické body, z ktorých sa automaticky vypočíta celková zmena energie.



(a) (b) (c) (d)

Obr. 6.6: Vizualizovaná lower-bound trajektória, ktorá je výsledkom dokovania pomocou nástroja nástroja CaverDock.

Kapitola 7

Záver

V úvode teoretickej časti sú opísané proteíny a ich dôležitosť v živých organizmoch. Nasledované chemickou štruktúrou, rozdeleniami na základe vybraných kritérií a odôvodnením ich dôležitosti pri návrhu liečiv.

Druhý diel teoretickej časti vysvetľuje termíny neoddeliteľné od problematiky dokovania ako sú aktívne miesta, tunely a ligandy. Kde aktívne miesta sú časti proteínov, v ktorých sa naväzujú ligandy. Tunely sú cesty, ktorými sa ligandy dostávajú k aktívnym miestam. Ligandy sú molekuly slúžiace na aktiváciu alebo deaktiváciu funkcie proteínov.

Po objasnení spomínaných termínov je čitateľovi vysvetlený základný proces molekulárneho dokovania, metódy pre štruktúrovaný návrh liečiv. Ďalej je práca zameraná na prezentáciu rôznych prístupov využívaných pri riešení problematiky molekulárneho dokovania. Proces dokovania je rozdelený do troch hlavných fáz. Prvá časť sa zaoberá prípravou a reprezentáciou proteínu a ligandu. Druhá časť je samotný proces dokovanania ligandu. Tretia časť je zameraná na hodnotenie póz vygenerovaných dokovacím algoritmom pomocou skórovacej funkcie.

V kapitole venovanej nástroju Caver Web sú opísané jeho súčasti a aké čiastkové nástroje využíva. Taktiež sú v nej demonštrované výsledky spomínaných podnástrojov ako aj vstupy a výstupy nástroja Caver Web, do ktorého bola pridaná nová funkcionalita.

V nasledujúcej kapitole venovanej praktickej časti 6 sú opísané použité nástroje na spracovanie molekúl, databáza a dataset, ktorý bol vybraný pre implementáciu modulu. Pre získanie datasetu je využitá databáza ZINC. Konkrétne *subset*[1] databázy ZINC s označením "world" bol vybraný, ako vhodný pre analýzu interakcií liečiv s proteínmi.

Poslednou časťou práce je vyhodnotenie výsledkov a stručný opis použitia modulu. Použitie modulu má dve kľúčové časti. Nastavenie a spustenie kalkulácie. Následná analýza výsledkov s možnosťou vizualizácie 3D štruktúr použitých pri dokovaní ako aj pozície ligandu pre každý krok kalkulácie.

Výsledky práce budú vložené do nástroja Caver Web, ktorý bol publikovaný v časopise *Nucleic Acids Research*[2] z faktorom dopadu 11.5. Caver Web je verejne dostupný na webovej adrese https://loschmidt.chemi.muni.cz/caverweb/.

[1] *Subset je zaužívaný anglický ekvivalent pre podmnožinu v kontexte dát.*

[2] *Nucleic Acids Research* je časopis dostupný na adrese `https://academic.oup.com/nar` zaoberajúci sa prelomovými výskumami zaoberajúcimi sa fyzickými, chemickými, biochemickými a biologickými aspektami nukleových kyselín a proteínov.

Literatúra

[1] *ZINC15 Subsets* [online]. [cit. Apríl 14, 2021]. Dostupné z: http://zinc15.docking.org/catalogs/subsets.

[2] *ZINC15 Resources wiki page* [online]. Február, 2017 [cit. Apríl 14, 2021]. Dostupné z: http://wiki.docking.org/index.php/ZINC15:Resources.

[3] Baxter, C., Murray, C., Clark, D., Westhead, D. a Eldridge, M. Flexible docking using Tabu search and an empirical estimate of binding affinity. *Proteins*. November 1998, zv. 33, č. 3, s. 367—382. DOI: 10.1002/(sici)1097-0134(19981115)33:3<367::aid-prot6>3.0.co;2-w. ISSN 0887-3585. Dostupné z: https://doi.org/10.1002/(SICI)1097-0134(19981115)33:3<367::AID-PROT6>3.0.CO;2-W.

[4] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N. et al. The Protein Data Bank. *Nucleic Acids Research*. Január 2000, zv. 28, č. 1, s. 235–242. DOI: 10.1093/nar/28.1.235. ISSN 0305-1048. Dostupné z: https://doi.org/10.1093/nar/28.1.235.

[5] Bottegoni, G., Kufareva, I., Totrov, M. a Abagyan, R. Four-dimensional docking: a fast and accurate account of discrete receptor flexibility in ligand docking. *Journal of medicinal chemistry*. Jan 2009, zv. 52, č. 2, s. 397–406. DOI: 10.1021/jm8009958. ISSN 1520-4804. Dostupné z: https://pubmed.ncbi.nlm.nih.gov/19090659.

[6] Boundless. *Enzyme Active Site and Substrate Specificity*, 03. Jan 2021. Dostupné z: https://bio.libretexts.org/go/page/8811.

[7] Chen, H.-M., Liu, B.-F., Huang, H.-L., Hwang, S.-F. a Ho, S.-Y. SODOCK: Swarm optimization for highly flexible protein–ligand docking. *Journal of Computational Chemistry*. 2007, zv. 28, č. 2, s. 612–623. DOI: https://doi.org/10.1002/jcc.20542. Dostupné z: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.20542.

[8] Chen, K., Li, T. a Cao, T. Tribe-PSO: A novel global optimization algorithm and its application in molecular docking. *Chemometrics and Intelligent Laboratory Systems*. 2006, zv. 82, č. 1, s. 248–259. DOI: https://doi.org/10.1016/j.chemolab.2005.06.017. ISSN 0169-7439. Selected Papers from the International Conference on Chemometrics and Bioinformatics in Asia. Dostupné z: https://www.sciencedirect.com/science/article/pii/S0169743905001346.

[9] Chovancova, E., Pavelka, A., Benes, P., Strnad, O., Brezovsky, J. et al. CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures. *PLOS Computational Biology*. Public Library of Science. Október 2012, zv. 8, č. 10, s. 1–12. DOI: 10.1371/journal.pcbi.1002708. Dostupné z: https://doi.org/10.1371/journal.pcbi.1002708.

[10] Claussen, H., Buning, C., Rarey, M. a Lengauer, T. FlexE: efficient molecular docking considering protein structure variations1 1Edited by J. Thornton. *Journal of Molecular Biology*. 2001, zv. 308, č. 2, s. 377–395. DOI: https://doi.org/10.1006/jmbi.2001.4551. ISSN 0022-2836. Dostupné z: https://www.sciencedirect.com/science/article/pii/S0022283601945516.

[11] Durrant, J. a McCammon, J. NNScore 2.0: A Neural-Network Receptor–Ligand Scoring Function. *Journal of Chemical Information and Modeling*. 2011, zv. 51, s. 2897 – 2903.

[12] Durrsant, J. D. a McCammon, J. A. NNScsore: A Neural-Network-Based Scoring Function for the Characterization of P.-L. C. *Josurnal of Chemical Information and Modeling*. 2010, zv. 50, č. 10, s. 1865 – 1871. Dostupné z: https://doi.org/10.1021/ci100244v.

[13] Gaudreault, F., Chartier, M. a Najmanovich, R. Side-chain rotamer changes upon ligand binding: common, crucial, correlate with entropy and rearrange hydrogen bonding. *Bioinformatics*. September 2012, zv. 28, č. 18, s. i423–i430. DOI: 10.1093/bioinformatics/bts395. ISSN 1367-4803. Dostupné z: https://doi.org/10.1093/bioinformatics/bts395.

[14] Gromiha, M. M. *Protein Bioinformatics:From Sequence to Function*. 1. vyd. Elsevier, A Division of Reed Elsevier India, 2004. ISBN 978-81-312-2297-3.

[15] Guan, B., Zhang, C. a Ning, J. Genetic algorithm with a crossover elitist preservation mechanism for protein-ligand docking. Springer Berlin Heidelberg. Sep 2017, zv. 7, č. 1, s. 174–174. DOI: 10.1186/s13568-017-0476-0. ISSN 2191-0855. Dostupné z: https://pubmed.ncbi.nlm.nih.gov/28905320.

[16] Hassan, N. M., Alhossary, A. A., Mu, Y. a Kwoh, C.-K. Protein-Ligand Blind Docking Using QuickVina-W With Inter-Process Spatio-Temporal Integration. *Scientific Reports*. Nov 2017, zv. 7, č. 1, s. 15451. DOI: 10.1038/s41598-017-15571-7. ISSN 2045-2322. Dostupné z: https://doi.org/10.1038/s41598-017-15571-7.

[17] Hou, T., Wang, J., Chen, L. a Xu, X. Automated Docking of Feptides and Proteins by Using a Genetic Algorithm Combined with a Tabu Search. *Protein engineering*. September 1999, zv. 12, s. 639–48. DOI: 10.1093/protein/12.8.639.

[18] Huang, S.-Y. a Zou, X. Efficient molecular docking of NMR structures: application to HIV-1 protease. *Protein science : a publication of the Protein Society*. 2006/11/22. Cold Spring Harbor Laboratory Press. Jan 2007, zv. 16, č. 1, s. 43–51.

[19] Huang, S.-Y. a Zou, X. Advances and challenges in protein-ligand docking. *International journal of molecular sciences*. Molecular Diversity Preservation International (MDPI). Aug 2010, zv. 11, č. 8, s. 3016–3034. DOI:

10.3390/ijms11083016. ISSN 1422-0067. Dostupné z: https://pubmed.ncbi.nlm.nih.gov/21152288.

[20] Jurcik, A., Bednar, D., Byska, J., Marques, S. M., Furmanova, K. et al. CAVER Analyst 2.0: analysis and visualization of channels and tunnels ... *Bioinformatics*. Máj 2018, zv. 34, č. 20, s. 3586–3588. DOI: 10.1093/bioinformatics/bty386. ISSN 1367-4803. Dostupné z: https://doi.org/10.1093/bioinformatics/bty386.

[21] Kessel, A. *Introduction to proteins : structure, function, and motion*. Boca Raton, FL: CRC Press, 2018. ISBN 9781498747172.

[22] Knegtel, R. M., Kuntz, I. D. a Oshiro, C. Molecular docking to ensembles of protein structures11Edited by B. Honig. *Journal of Molecular Biology*. 1997, zv. 266, č. 2, s. 424–440. DOI: https://doi.org/10.1006/jmbi.1996.0776. ISSN 0022-2836. Dostupné z: https://www.sciencedirect.com/science/article/pii/S0022283696907767.

[23] Knegtel, R. M., Kuntz, I. D. a Oshiro, C. Molecular docking to ensembles of protein structures11Edited by B. Honig. *Journal of Molecular Biology*. 1997, zv. 266, č. 2, s. 424–440. DOI: https://doi.org/10.1006/jmbi.1996.0776. ISSN 0022-2836. Dostupné z: https://www.sciencedirect.com/science/article/pii/S0022283696907767.

[24] Kunka, A., Damborsky, J. a Prokop, Z. Chapter Seven - Haloalkane Dehalogenases From Marine Organisms. In: *Marine Enzymes and Specialized Metabolism - Part B*. Academic Press, 2018, sv. 605, s. 203–251. Methods in Enzymology. DOI: https://doi.org/10.1016/bs.mie.2018.03.005. ISSN 0076-6879. Dostupné z: https://www.sciencedirect.com/science/article/pii/S0076687918301344.

[25] Li, G.-B., Yang, L., Wang, W.-J., Li, L.-L. a Yang, S.-Y. ID-Score: A New Empirical Scoring Function Based on A Comprehensive Set of Descriptors Related to Protein-Ligand Interactions. *Journal of chemical information and modeling*. Február 2013, zv. 53. DOI: 10.1021/ci300493w.

[26] Encyclopaedia Britannica, T. E. of. *Ligand* [online]. Encyclopædia Britannica, August 12, 2010 [cit. January 10, 2021]. Dostupné z: https://www.britannica.com/science/ligand.

[27] Editors, B. *Ligand* [online]. Apríl, 2018 [cit. January 10, 2021]. Dostupné z: https://biologydictionary.net/ligand.

[28] Liu, J. a Wang, R. Classification of Current Scoring Functions. *Journal of Chemical Information and Modeling*. 1. vyd. 2015, zv. 55, č. 3, s. 475–482. DOI: 10.1021/ci500731a. PMID: 25647463. Dostupné z: https://doi.org/10.1021/ci500731a.

[29] Lodish, H. *Molecular Cell Biology*. 1. vyd. W. H. Freeman and Company, 2016. ISBN 978-1-4641-8339-3.

[30] Moraes, J., Pappa, G. L., Pires, D. a Izidoro, S. C. GASS-WEB: a web server for identifying enzyme active sites based on genetic algorithms. *Nucleic Acids Research*. Apríl 2017, zv. 45, W1, s. W315–W319. DOI: 10.1093/nar/gkx337. ISSN 0305-1048. Dostupné z: https://doi.org/10.1093/nar/gkx337.

[31] Morris, G., Goodsell, D., Halliday, R. S., Huey, R., Hart, W. et al. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* 1998, zv. 19, s. 1639–1662.

[32] Morris, G. a Huey, R. AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J. Comput. Chem.* Január 2009, zv. 48, s. 443–453.

[33] Morris, G. M. a Lim Wilby, M. *Molecular Docking*. Totowa, NJ: Humana Press, 2008. 365–382 s. ISBN 978-1-59745-177-2. Dostupné z: https://doi.org/10.1007/978-1-59745-177-2_19.

[34] O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T. et al. Open Babel: An open chemical toolbox. *Journal of Cheminformatics*. Oct 2011, zv. 3, č. 1, s. 33. DOI: 10.1186/1758-2946-3-33. ISSN 1758-2946. Dostupné z: https://doi.org/10.1186/1758-2946-3-33.

[35] Sterling, T. a Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling*. 2015, zv. 55, č. 11, s. 2324–2337. DOI: 10.1021/acs.jcim.5b00559. PMID: 26479676. Dostupné z: https://doi.org/10.1021/acs.jcim.5b00559.

[36] Stourac, J., Vavra, O., Kokkonen, P., Filipovic, J., Pinto, G. et al. Caver Web 1.0: identification of tunnels and channels in proteins and analysis of ligand transport. *Nucleic Acids Research*. 1. vyd. Máj 2019, zv. 47, W1, s. W414–W422. DOI: 10.1093/nar/gkz378. ISSN 0305-1048. Dostupné z: https://doi.org/10.1093/nar/gkz378.

[37] Vavra, O., Filipovic, J., Plhak, J., Bednar, D., Marques, S. M. et al. CaverDock: a molecular docking-based tool to analyse ligand transport through protein tunnels and channels. *Bioinformatics*. Máj 2019, zv. 35, č. 23, s. 4986–4993. DOI: 10.1093/bioinformatics/btz386. ISSN 1367-4803. Dostupné z: https://doi.org/10.1093/bioinformatics/btz386.

[38] Verschueren, K., Seljée, F., Rozeboom, H., Kalk, K. a Dijkstra, B. Crystallographic analysis of the catalytic mechanism of haloalkane dehalogenase. *Nature*. June 1993, zv. 363, č. 6431, s. 693—698. DOI: 10.1038/363693a0. ISSN 0028-0836. Dostupné z: https://doi.org/10.1038/363693a0.

[39] Zhang, W., Bell, E. W., Yin, M. a Zhang, Y. EDock: blind protein–ligand docking by replica-exchange monte carlo simulation. *Journal of Cheminformatics*. May 2020, zv. 12, č. 1, s. 37. DOI: 10.1186/s13321-020-00440-9. ISSN 1758-2946. Dostupné z: https://doi.org/10.1186/s13321-020-00440-9.

[40] Zilian, D. a Sotriffer, C. SFCscore(RF): A Random Forest-Based Scoring Function for Improved Affinity Prediction of Protein-Ligand Complexes. *Journal of chemical information and modeling*. Máj 2013, zv. 53. DOI: 10.1021/ci400120b.

Príloha A

Obsah SD karty

Na priloženej SD karte je možné nájsť:

- bash skript pre vygenerovanie a monitorovanie úloh **monitor_batches.sh**,
- bash skript vytvárajúci konkrétnu úlohu **job_batch.sh**,
- bash skript pre stiahnutie a spracovanie ZINC datasetu **prepare_dataset.sh**,
- predspracovaný dataset ligandov **dataset**,
- python skript z MGL tools pre prípravu ligandu **prepare_ligand4.py**.