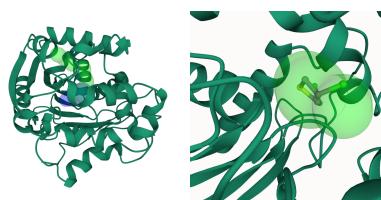


Identifikace proteinových tunelů s využitím molekulárních dynamik

Bc. Petr Kohout*



Abstrakt

Tato práce se zabývá analýzou proteinových struktur. Cílem je vytvořit Caver Web 2.0 – novou verzi webové aplikace, která začlení další vědecké nástroje a uživatelům umožní projít komplikovaný pracovní protokol za poskytnutí relevantních výsledků bez nutnosti hlubší znalosti integrovaných nástrojů. Vše bude zprostředkováno prostřednictvím jednoduchého a interaktivního uživatelského rozhraní. Aplikace rozšiřuje původní aplikaci Caver Web 1.0 o nové vlastnosti. Caver Web 1.0 je webový server vhodný pro identifikaci proteinových tunelů a kanálů, pro které umožňuje spustit analýzy transportu ligandů. Program se vyznačuje intuitivním a uživatelsky přívětivým rozhraním s minimem požadovaných vstupů od uživatele. Server je vhodný i pro výzkumníky bez pokročilých bioinformatických nebo technických znalostí. Jeho současná verze je ve vědecké komunitě dobře zavedená a velmi využívaná (35 000 dokončených výpočtů během dvou let provozu). Nejvýznamnějším omezením současné verze je možnost analyzovat pouze statickou strukturu, což často poskytuje neúplný biologický obraz. Proto jsme se rozhodli nástroj rozšířit o výpočet molekulárních dynamik, které poskytnou ucelený obraz na proměny proteinových struktur. Touto funkcí se Caver Web 2.0 stane prvním webovým nástrojem, který bude poskytovat analýzu tunelů bez nutnosti ručních výpočtů molekulárních dynamik. Grafické uživatelské rozhraní bude navrženo speciálně pro vědeckou komunitu s podporou jednoduchého exportu. Nástroj bude zdarma k dispozici celé vědecké komunitě.

Klíčová slova: bioinformatika, proteinové inženýrství, molekulární dynamika, tunely, dokování

Přiložené materiály: N/A

*xkohou14@fit.vutbr.cz, Fakulta informačních technologií, Vysoké učení technické v Brně

1. Úvod

Jednou z hlavních složek živých organismů na planetě jsou makromolekulární sloučeniny zvané proteiny (bílkoviny). Proteiny jsou tvořeny řetězcem aminokyselin, který je syntetizován na základě informace obsažené v DNA. Sekvence aminokyselin se za působení nekovalentních vazeb, které působí mezi vodíky (vodíkové můstky), ionty (iontové vazby) a mezi molekulami

(van der Waalsovy síly), formuje do 3D prostorové struktury zvané konformace. Posloupnost aminokyselin jednotlivých proteinů zajišťuje jejich unikátní konformaci, která jim také přřazuje jedinečné vlastnosti [1].

Při pohledu na protein bychom však neviděli nehybný propletenec proteinové struktury, ale pulzující a neustále se měnící systém. Výše zmíněné síly v kom-

binaci s vlivy z okolního prostředí (působení rozpouštěla, další molekuly) mají za následek neustálý pohyb atomů v rámci proteinu. Pro tento děj se zavedl pojem molekulární dynamiky a má za následek neustálé proměny proteinových konformací.

Proteiny můžeme nalézt ve vlasech (keratin) a svalech (aktin), kde ovlivňují náš vzhled a umožňují motoriku, nebo v mozku (myelin), kde zajišťují správné a rychlé vedení nervových vznrchů. Proteiny můžeme rozdělit do několika oblastí. Hovoříme o transportních, stavebních, zásobných, pohybových, ochranných a katalytických funkcích [2].

Analýza proteinů je velmi užitečná nejenom k lepšímu pochopení živých organismů, ale také k jejich cílenému vylepšování. Proteiny jsou často velmi zajímavým cílem aplikace léčiv, které dokáží ovlivnit jejich funkci. Další oblastí využití proteinů je průmysl (například syrovátky). Většina z nich však není k využití v průmyslu vhodná, protože se vyvinuly uvnitř buněk a mimo ně nedokáží fungovat. Proto vzniká tlak na nalezení jejich mutací, které jejich vlastnosti zlepší.

Odborníci, kteří se věnují analýze, úpravě a syntéze proteinů, se nazývají *proteinoví inženýři*. Jelikož fyzické laboratorní experimenty jsou časově i finančně náročné, přesouvá se značná část experimentů do oblasti výpočetní techniky, kdy vznikají specializované programy.

Díky tomu zažila bioinformatika explozi a v současné době existuje nepřeberné množství počítačových různých nástrojů a algoritmů, které jsou schopny analyzovat téměř všechny možné aspekty proteinů: jejich funkce, struktury a vlastnosti.

2. Enzymy a proteinové tunely v čase

Z výše zmíněného výčtu funkčních skupin proteinů se posledně jmenovaná skupina (tzv. enzymy) podílí na biochemických procesech, které probíhají v organismu a zpracovávají nebo produkují důležité látky.

Enzymy jsou nedílnou součástí každé buňky, která jich obsahuje velké množství. V buňce plní velice užitečnou funkci katalyzátorů chemických reakcí, kdy snižují potřebnou *aktivaci energii*. Bez jejich přítomnosti by většina chemických dějů v buňce neprobíhala [2].

Od ostatních skupin proteinů se odlišují především principem fungování. Zatímco ostatní skupiny pouze naváží svůj ligand (malou molekulu) a tím je jejich aktivita splněna, pro enzymy je to první potřebný krok k zahájení jejich činnosti, kdy váží jeden nebo více ligandů, zvaných *substráty*. Poté substrát vázaný v tzv. *aktivním místu* (například principem zámku a klíče) přeměňuje na produkt při tvorbě nebo rušení kova-

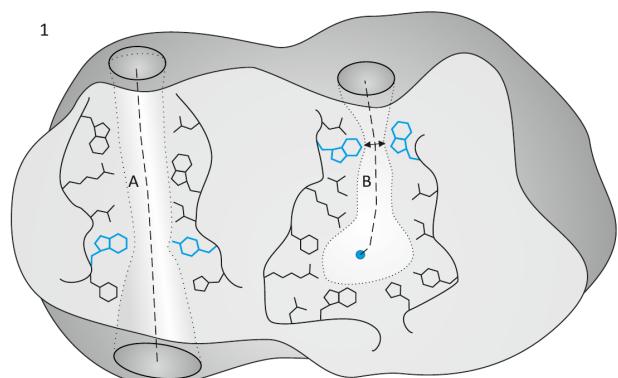
lentních vazeb. Po dokončení reakce je produkt díky pozměněné struktuře, která již nesplňuje podmínky pro vázání v aktivním místu, uvolněn do prostředí [2].

2.1 Proteinové tunely

Ve všech skupinách proteinů můžeme nalézt spousty zákrutů, otvorů a dlouhých slepých chodeb, které se nazývají *tunely*. Problematika těchto tunelů je však důkladně studována ve spojitosti s enzymy, které si v milionech let trvající evoluci vyvinuly a zformovaly svá aktivní místa právě na konci těchto chodeb [3].

Potenciálním důvodem přenosu aktivního místa dovnitř enzymů může být snaha enzymů o plnou kontrolu okolí aktivního místa, která by na povrchu mohla být ovlivněna různými rozpouštědly z okolního prostředí. Aktivní místa jsou spojena s povrchem *tunely*, které zprostředkovávají výměnu látek mezi aktivním místem a vnějším prostředím a rozšiřují tak model *zámků a klíče* o prvek klíčové dírky. *Klíčová dírka* v tomto případě tvořená tunelem určitého tvaru a složení předurčuje enzymatickou funkci celého enzymu [3].

V literatuře se lze setkat s mnoha výklady termínu *tunel* (*tunnel*) a *kanál* (*channel*). V rámci tohoto textu a studované literatury pojednáváme o *tunelu*, označujícím spojení mezi aktivním místem uvnitř enzymu a jeho povrchem nebo spojení mezi aktivními míssty uvnitř enzymu, či v proteinovém komplexu (multimeru). Pojem *kanál* označuje spojení různých částí povrchu proteinu, kterou může látka projít bez nutnosti transformace (viz obrázek 1) [3].



Obrázek 1. Proteinový kanál (A) a tunel (B) (převzato z [4]).

Nejužší místo tunelu se nazývá *hrdlo*. Hrdlo tunelu je často ovládáno tzv. *bránou*, která toto místo může ještě více zužovat (uzavírat) a učinit tak tunel neprůchozí. Brána často otevírá tunel v určitém časovém intervalu, takže nabízí další regulační mechanismus pro přístup látek do aktivního místa [3].

Některé proteiny při prvotní analýze působí dojem, že obsahují neměnné *permanentní tunely*, které

se v čase neuzavírají. Pokud bychom však tyto proteiny podrobili dlouhodobému pozorování s možností identifikace tunelů v reálném čase, zjistili bychom, že některé proteiny disponují dalšími otevřenými tunely jen v určitých okamžících. Tunely, které jsou identifikovány jen ve vybraném čase se nazývají *přechodné* (*transient*) [3].

2.1.1 Počítačová analýza tunelů

Pro detekci tunelů je nutné protein reprezentovat vhodnou strukturou. Protože tunel prochází volným prostorem uvnitř proteinu, musí pro něj platit, že jeho osa se nachází v největší možné vzdálenosti od jeho atomů. K nalezení osy se využívá konstrukce Voroného diagramu [5]. Vrcholy diagramu odpovídají atomům proteinu a jeho hrany identifikují osu tunelu. Ke spuštění výpočtu tunelu je nejprve zapotřebí specifikovat bod uvnitř proteinu, ze kterého je vyhledávání spuštěno. Poté je vyhledána cesta po hranách Voroného diagramu až do vnějšího prostředí bílkoviny. K prohledávání lze využít například Dijkstrova algoritmu [6].

2.1.2 Analýza dynamického chování

Jelikož v současnosti existují v naprosté většině záznamy pouze o statické podobě proteinů, je nutné pro sledování změn v struktuře proteinů v čase využít počítačových programů. Pro sledování změn struktury proteinu v čase se využívá výpočtu *molekulárních dynamik*.

Simulace molekulárních dynamik je metoda určená na analýzu proměnlivosti atomových struktur (a tedy i tunelů). Simulace umožňuje předpovídat chování molekul za určitých podmínek (teploty, tlaku a prostředí). Pro výpočet molekulárních dynamik se používá specializovaný software. Mezi nejpopulárnější nástroje patří AMBER, CHARMM, GROMACS, NAMD a YASARA. Ačkoli se jedná o velice optimalizované nástroje, stále není možné simulovat rozsáhlé a dlouhé dynamiky z důvodu velké náročnosti na hardware a výpočetní čas. Příkladem stroje, který je schopen provést výpočty simulující trvání pozorování stovky nanosekund nebo jednotky mikrosekund za využití velkého množství grafických jednotek, je superpočítáč Anton [3].

Simulace molekulárních dynamik probíhá v silnulačních krocích a k popisu chování systému využívá druhého Newtonova zákona. U každé částice systému (atomu nebo molekuly) je veden záznam o její kinetické energii (směru a zrychlení). V každém kroku simulace je proveden integrační krok a je určena nová souřadnice a energie částice. Pohyb částice je ovlivněn působením okolních částic prostřednictvím elektrostatických sil a kovalentních a nekovalentních vazeb mezi nimi. Parametry jednotlivých částic (hmotnost, veli-

kost a další) jsou definovány pomocí tzv. *silových polí*, jejichž hodnoty byly naměřeny v laboratořích. [7, 8].

2.1.3 Využívání proteinových tunelů

Jak bylo zmíněno výše, současná lidská činnost se zaměřuje především na analýzy možnosti kontroly proteinů. Snahou proteinových inženýrů je navýšit, či blokovat funkčnost jednotlivých proteinů za účelem dosažení kýženého efektu v organismu nebo biologickém systému. V případě záměru blokace činnosti enzymu nabízí proteinové tunely ideální místo pro vložení (navázání) tzv. artefaktů, které aktivní místo proteinu učiní nepřístupným pro objekty z vnějšího prostředí [3]. Naopak v případě snahy navýšení produkce enzymu se vědci zaměřují na způsoby, jak učinit přístupovou cestu do aktivního místa přívětivější pro vybraný ligand. Tohoto efektu lze docílit opakem dříve zmíněné techniky, a to odstraněním případních artefaktů, které by mohly blokovat cestu ligandu k aktivnímu místu. Odstranění blokujících artefaktů lze provést substitucí (mutací) dané sekvence aminokyselin v proteinu. Změny v základní sekvenci aminokyselin však mohou způsobit jiné nežádoucí efekty, a proto je nutné mít možnost tyto změny rychle a jednoduše ověřovat a reportovat. Zmíněné techniky jednoznačně ukazují, jak důležitá místa pro činnost proteinů tunely představují. Proto je potřebné zmínit důležitost a nutnost tvorby aplikací v doméně analýzy proteinových tunelů, které vědecké komunitě tyto možnosti poskytnou.

Za tímto účelem vznikla aplikace Caver Web [9], která bude prezentována v článku. Jejím úkolem je zjednodušit a urychlit práci proteinovým inženýrům bez nutnosti podrobné znalosti integrovaných nástrojů. Uživatelé jsou odstíněni od nastavování a spouštění jednotlivých nástrojů ve správném pořadí a se správnou vstupní konfigurací. Mohou se tak plně zaměřit na analýzu proteinových tunelů a kanálů.

3. Existující řešení

Tato práce je realizována pod záštitou Loschmidtových laboratoří, ve kterých vznikla první verze aplikace Caver Web 1.0 [9]. Aplikaci implementoval vedoucí této práce Mgr. Jan Štourač, a proto byla k rozšíření zvolena právě tato aplikace.

Současně vznikající nástroj Caver Web 2.0 vytvoří novou generaci již existujícího nástroje. Nová verze aplikace přináší rozšíření o molekulární dynamiky. Integrace molekulárních dynamik v rámci webové aplikace vytváří zcela novou dimenzi a v současné době není známa aplikace se stejnými možnostmi.

V rámci tvorby předchozí verze byl proveden průzkum již existujících řešení v oblasti analýzy protei-

nových struktur a proteinových tunelů. Srovnání existujících řešení obsahuje tabulka 1.

Původní verze aplikace se skládá z přímočarého workflow, které uživateli provede celým procesem bez nutnosti jeho větších zásahů. Nová verze aplikace toto workflow rozšiřuje o nové užitečné kroky se zachováním původního postupu. Proto zde bude tento postup zkráceně popsán a následně rozšířen o popis nových funkcionalit v další části textu. Nástroj [9] je dostupný na adrese: <https://loschmidt.chemi.muni.cz/caverweb/>. Pracovní protokol je zachycen na obrázku 5. V první fázi je uživatel vyzván k výběru struktury, se kterou bude program pracovat. Uživatel může vložit svoji proteinovou strukturu ve formátu `pdb` nebo `cif`.

Nejdůležitějším krokem při detekci tunelu je výběr správného výchozího bodu. Pro detekci tunelů je využit software Caver [17], kde výchozí bod definuje společný bod pro všechny detekované tunely. Špatně umístěný bod může významně ovlivnit relevanci detekovaných tunelů. Pro usnadnění výběru počátečního bodu se využívá automatických protokolů, které vyberou výchozí body vhodné pro nejběžnější scénáře. Nástroj umožňuje vybrat počáteční bod jednou ze čtyř možností: (i) výběr katalytické kapsy; (ii) výběr kapsy; (iii) výběr ligandu, a (iv) výběr ze sekvence s možností manuální korekce souřadnic [9].

Posledním (nepovinným) krokem pracovního postupu je analýza transportu ligandů skrze detekované tunely. Pro analýzu transportu se využívá software CaverDock [18]. Požadovaným vstupem pro kalkulace transportu ligandu je nutné zadat dvě informace. První část vstupu tvoří ligand, což je malá molekula, kterou lze nahrát v podobě souboru, textového popisu nebo ji lze nakreslit pomocí přítomného nástroje. Lze zadat i více ligandů současně (obrázek 7E). Za druhou část vstupu je považován výběr alespoň jednoho identifikovaného tunelu, který se použije jako cesta pro transport zvolených ligandů [9].

Po ukončení identifikace proteinových tunelů jsou uživateli zobrazeny výsledky v prohlížeči výsledků. Okno s výsledky obsahuje 3 části. První obsahuje tabulku s identifikovanými tunely (obrázek 7C), ve které jsou zobrazeny veškeré statistiky a informace o tunelu jako: délka, velikost hrdla, členitost a propustnost (kvalita) tunelu. Po otevření podrobného náhledu lze zjistit, ze kterých aminokyselin je tunel zformován a které aminokyseliny vytváří jeho hrdlo (obrázek 7D). V druhé části lze daný tunel graficky zobrazit a v poslední sekci se nachází seznam vypočítaných dokovacích úloh (analýz transportu ligandu). Každý záznam disponuje daty získanými při analýze. Jsou

to data pro animaci dokovacího (transportního) procesu ligandu do/z počátečního (uživatelem zvoleného) místa identifikovaných tunelů.

Tyto kroky a prohlížeč výsledků jsou implementovány i v nové verzi aplikace, aby současní uživatelé nemuseli měnit své pracovní návyky a byla umožněna rychlá analýza struktur. Grafická implementace jednotlivých kroků bude znázorněna v další sekci tohoto textu.

4. Analýza rozšíření

Rozšíření aplikace integruje výpočet nových proteinových konformací, jejichž výsledky je nutné zobrazit ve srozumitelné formě. Níže jsou popsány vybrané nástroje a techniky zobrazení výsledků.

4.1 Molekulární dynamiky

Výpočet molekulárních dynamik spadá do časově náročných výpočtů a provádí se pomocí specializovaných nástrojů. Aplikace využívá nástroj Yasara [19]. Nástroj Yasara je kontinuálně vyvíjen již od roku 1993. Autoři se zaměřují na jednoduché používání tohoto nástroje. Jednou formou usnadnění používání daného nástroje bylo publikování sady skriptů pro spuštění jednotlivých úloh. Součástí je i skript `md_run.mcr` [20] pro výpočet molekulárních dynamik zadané proteinové struktury. V rámci skriptu lze zadat velké množství parametrů simulace, které jsou inicializovány na základní hodnoty. Při používání aplikace se však uvažuje velmi široká skupina uživatelů s různě obsáhlými (a někdy až nulovými) znalostmi v oblasti molekulárních dynamik. Proto jsou v nástroji obsaženy pouze nejnuttnejší a nejčastěji měněné parametry. Jsou jimi: teplota, délka simulované doby, hodnota pH prostředí, velikost kroku simulace a volba parametrů silových polí. Posledně jmenovaná silová pole jsou parametry, kterými můžeme definovat potencionální energii atomů a molekul v rámci simulačního modelu. Hodnoty parametrů jsou získány experimenty v laboratořích současně s aplikací kvantové mechaniky [21].

4.2 tCONCOORD

tCONCOORD [22] umožňuje generování různých prostorových konformací proteinu na základě geometrických transformací. Nejprve je proteinová struktura proložena její geometrickou reprezentací. Tato reprezentace následně slouží pro aplikaci náhodných transformací v počtu několika stovek pokusů, aby bylo dosaženo velkého množství snímků. Snímky jsou následně analyzovány metodou, která na základě laboratorně ověřených podmínek pro jednotlivé transformace zaručí dosažení požadované biologické relevantnosti.

Analýza tunelů a kanálů			Analýza transportu ligandu			Integrace molekulárních dynamik		
Software	Vstup	Podporováno	Výběr startovacího bodu	Podporováno	Zdroj ligandu	Výstup	Podporováno	Reference
Caver Web 2.0	PDB ID ^a , PDB/CIF soubor ^b	Ano	katalytická kapsa, kapsa, ligandy, výběr sekvence, souřadnice katalytická kapsa, kapsa, ligandy, výběr sekvence, souřadnice	Ano	ZINC15, soubor uživatele, kreslení	tunely/kanály, trajektorie proteinu	Ano	tato práce
Caver Web	PDB ID ^b , PDB/CIF soubor ^b	Ano		Ano	ZINC15, soubor uživatele, kreslení	tunely/kanály, trajektorie ligandu, energetický profil		[9]
MolAxis	PDB ID, PDB/CIF soubor	Ano	největší dutina, souřadnice	Ne	- ^d	tunely/kanály	Ne	[10]
MoleOnline	PDB ID ^c , PDB/CIF soubor ^c	Ano	katalytická rezidua, rezidua, kapsa, souřadnice, výběr sekvence	Ne	- ^d	tunely/kanály	Ne	[11]
BetaCavityWeb	PDB ID, PDB soubor	Ano	není vyžadováno	Ne	- ^d	tunely/kanály	Ne	[12]
PoreWalker	PDB soubor	Ano	není vyžadováno	Ne	- ^d	kanály	Ne	[13]
ChExVis	PDB ID, PDB soubor	Ano	katalytická rezidua, heteroatomové záznamy	Ne	- ^d	tunely/kanály	Ne	[14]
MoMA-LigPath ^a	PDB soubor	Ne	- ^d	Ano	součástí PDB souboru	trajektorie ligandu	Ne	[15]

Tabulka 1. Porovnání existujících řešení v doméně analýzy proteinových struktur a transportu ligandu (převzato z [9] a upraveno).

^a Webový server SLITHER pro analýzu transportu ligandu nebyl v době psaní textu přístupný.

^b Biologické jednotky detekovány pomocí skriptu MakeMultimer.

^c Biologické jednotky načteny z databáze PDBe [16].

^d Není možné provést.

V kaž-dém kroku vznikají desítky nových možných konformací, z nichž lze dále počítat další tvary proteinu. tCONCOORD tvoří alternativu k výpočtu molekulárních dynamik, kdy molekulární dynamiky investují výpočetní výkon do přesného určení koncové pozice atomu v rámci simulačního kroku, zatímco metoda tCONCOORD tento výpočetní výkon přesouvá do procesu ověření správnosti náhodně vygenerovaných snímků. Pomocí programu tCONCOORD je možné dosáhnout generování velice různorodých konformací daného proteinu, kterých by při běhu simulace bylo dosaženo až po dlouhém výpočtu. Získáme tak různorodé konformace v poměrně krátkém čase, a proto byl tento nástroj integrován v aplikaci.

4.3 Prezentace výsledků

Zavedením výpočtu molekulárních dynamik významně narůstá množství výsledných dat, proto je nutné zavést nové techniky jejich prezentování. Pracovní postup proteinových inženýrů v této oblasti spočívá ve výběru úzké skupiny kandidátních tunelů pro pozdější podrobné analyzování. Je tedy nutné z celé simulace a stovek identifikovaných tunelů vybrat takové tunely, které jsou v čase stabilní (nejčastěji otevřené) a mají

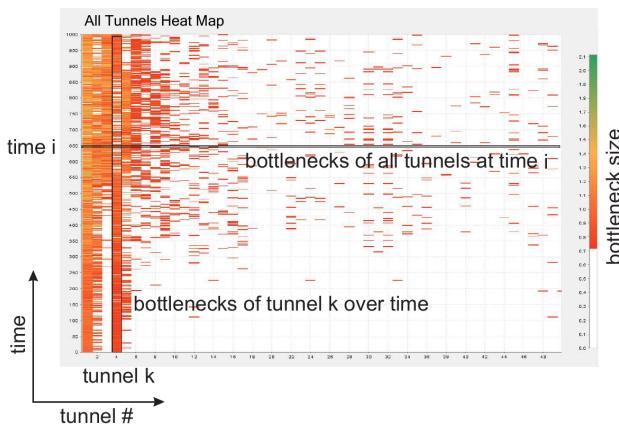
co největší propustnost. Vyfiltrované tunely jsou dále zkoumány v jejich dalších parametrech, jako je zkroucení a proměnlivost šířky v celé délce od aktivního místa až po povrch proteinu. Tunely jsou tak vyfiltrovány do jednotek kandidátů, u kterých se zkoumají možnosti zlepšení jejich vlastností a navrhují se případné mutace v sekvenci proteinu, nebo se volí vhodnější ligand [23]. Velké množství generovaných snímků proteinové trajektorie (tzv. snapshotů) činí tuto manuální analýzu proměnlivosti proteinových tunelů v reálném čase velice obtížnou až přímo nemožnou.

4.3.1 Analýza a volba řešení

Za účelem umožnění analýz trajektorie byl proveden průzkum existujícího řešení zabývajícího se analýzou molekulárních dynamik. K podrobné analýze byl vybrán nástroj MoleCollar [23] od brněnské vědecké skupiny zabývající se vizualizací biologických dat. MoleCollar je desktopová aplikace, která analyzuje trajektorie proteinových struktur. Jednou z nabízených funkcí je identifikace a analýza proteinových tunelů. Tyto analýzy přesně odpovídají požadavkům našeho řešení, a proto bylo rozhodnuto o jejich implementaci v aplikaci Caver Web 2.0.

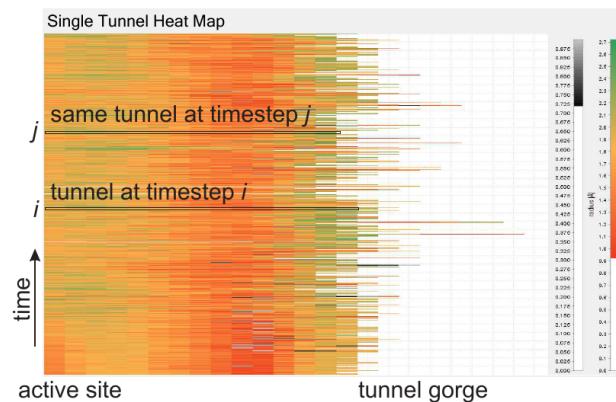
Pro implementaci byly zvoleny 3 metody prezentace výsledných dat: (i) All tunnels Heat Map (tepelná mapa všech tunelů); (ii) Single tunnel Heat Map (tepelná mapa jednoho tunelu v rámci celé simulace), a (iii) Contours (tunelové kontury).

Tepelná mapa všech tunelů (obrázek 2) slouží k zobrazení všech tunelů, které byly během běhu simulace identifikovány. Tunely jsou approximovány velikostí jejich hrdla, takže jeden tunel v konkrétním čase je reprezentován jedním bodem (barevným obdélníkem). Barevná výplň bodů je škálována podle šířky tunelu, přičemž bílá barva označuje, že tunel v čase ve vybraném snímku nesplnil podmínky identifikace (hodnota velikosti jeho hrdla byla menší, než uživatelem definovaný parametr) [23]. Získaná tepelná mapa usnadňuje výběr kandidátů pro podrobnější analýzu, které jsou v průběhu času nejčastěji otevřeny (jsou stabilní). Tyto tunely jsou vhodné k následné analýze transportu ligandů do aktivních míst proteinů. K analýze je vhodné vybrat snímek tunelu v čase, ve kterém disponuje nejmenší velikostí hrdla, protože velikost hrdla významně ovlivní propustnost celého tunelu.



Obrázek 2. Tepelná mapa všech tunelů zobrazuje variabilitu velikostí hrdel jednotlivých tunelů napříč celou dobou simulace. Bílá barva znamená, že tunel nebyl v daném snímku se zadánou hodnotou minimálního poloměru hrdla klasifikován (převzato z [23]).

Po výběru skupiny kandidátů z tepelné mapy všech tunelů lze jednotlivé tunely zobrazit v tepelné mapě jednoho tunelu (obrázek 3). Tato datová reprezentace na rozdíl od předchozí neapproximuje tunely jednou hodnotou, ale zobrazuje proměnlivost jednoho tunelu v celé jeho délce od aktivního místa po povrch proteinu [23]. Tento přehled nám umožňuje identifikovat úseky tunelu, které mohou být modifikovány. Například při zámkru zvýšit výtěžek z proteinu se snažíme rozšířit užší místa tunelu a při snaze tento protein blokovat v jeho činnosti hledáme způsoby, jak toto místo zablokovat nebo ještě více zúžit.



Obrázek 3. Tepelná mapa jednoho tunelu zobrazuje proměnlivost šířky tunelu v celé jeho délce (od počátečního místa až po dosažení povrchu) napříč celou dobou simulace. Bílá barva znamená, že tunel nebyl v daném snímku se zadánou hodnotou minimálního poloměru hrdla klasifikován (převzato z [23]).

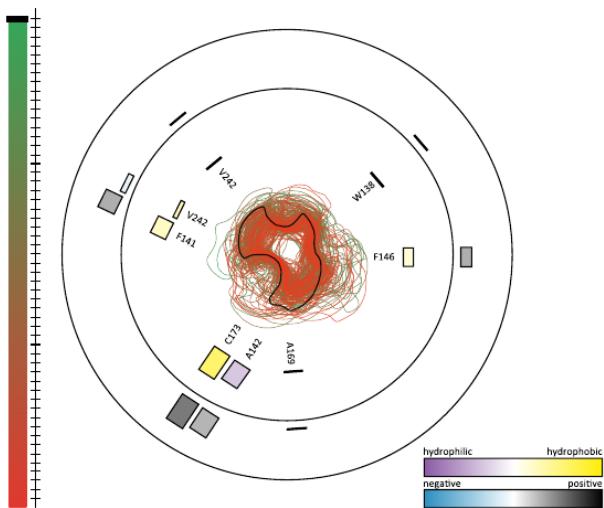
Po výběru oblasti tunelu chceme tuto oblast studovat v souvislosti se složením okolních aminokyselin, které tunel v tomto místě tvarují. Identifikování aminokyselin usnadňuje následnou volbu samotné úpravy proteinu. Při nalezení aminokyseliny, která negativně ovlivňuje požadované vlastnosti tunelu, můžeme navrhnut vhodné mutace v primární sekvenci proteinu a docílit tak našeho záměru. Pro tento typ analýzy byl vytvořen přehled kontur tunelu (obrázek 4). Kontury jsou vytvořeny jako obrys průřezu vedeného kolmo vůči ose tunelu ve vybrané vzdálenosti od počátečního místa. Společně s konturami jsou v přehledu zobrazeny i aminokyseliny včetně četnosti jejich výskytu ve snímcích a je je tak možné vybrat pro případné návrhy mutací proteinové sekvence [23].

5. Implementace

Předchozí verze aplikace byla vytvořena ve framework GWT. V rámci modernizace nástrojů bylo rozhodnuto převést implementaci do novějších technologií. Frontend aplikace je implementován nově pomocí knihovny React a backend je vytvořen ve frameworku Spring Boot.

Na obrázku 5 je graficky znázorněné workflow aplikace. Workflow bylo převzato z původní verze aplikace a bylo rozšířeno o nové funkcionality, které jsou zvýrazněny pomocí červené barvy.

Pracovní postup je stejně jako v předchozí verzi rozdělen do čtyř logických fází. První je vstupní fáze, ve které uživatel specifikuje data, se kterými bude dále pracovat. Uživatel si může zvolit ze čtyř způsobů vložení vstupních dat: (i) vložení jedné struktury proteinu (stejně jako v původní verzi), kdy uživatel na-



Obrázek 4. Kontury tunelu zobrazují ortogonální průřez tunelu k jeho ose. Tato analýza zobrazuje, které aminokyseliny vymezují tvar tunelu v dané vzdálenosti od počátečního místa. V náhledu kontur jsou zobrazeny průřezy konkrétního tunelu ve všech snímcích simulace (převzato z [23]).

hraje soubor obsahující proteinovou strukturu ve formátu `pdb` a `cif` nebo ji identifikuje pomocí `pdb` identifikátoru, který slouží jako vyhledávací klíč v proteinové databázi (Protein Data Bank); (ii) nahrání více souborů stejným způsobem jako v bodě (i); (iii) nahrání archivu obsahujícího až 50 souborů ve formátu `pdb` nebo `cif` a (iv) nahrání archivu obsahujícího již vypočtené trajektorie molekulárních dynamik.

V případě nahrání jednoho souboru dochází ke předzpracování vstupní struktury. Struktury jsou obvykle uloženy ve formě asymetrických jednotek, které nemusí odrážet jejich přirozeně se vyskytující kvartérní formy (biologické jednotky) a provedené analýzy těchto struktur mohou vést k nesprávným výsledkům prostřednictvím detekce neexistujících tunelů. Za tímto účelem je spuštěn software MakeMultimer [24], který našle odpovídající biologické jednotky. Uživatel poté může vybrat biologickou jednotku, se kterou chce pracovat, nebo pokračovat v práci s původní strukturou. Následně vstupuje do procedury, ve které specifikuje očekávaný výstup (zpracování souboru, obrázek 6). Uživatel volí opět mezi několika možnostmi: (i) jednoduchá analýza (single), pro analýzu statické struktury bez hledání podobných struktur; (ii) zarovnání struktury s vybranými podobnými strukturami; (iii) spuštění molekulárních dynamik na dané struktuře a následná analýza těchto dynamik, a (iv) generování více prostorových konformací proteinu pomocí nástroje tCONCOORD.

Jednoduchá analýza byla ponechána v aplikaci za účelem umožnění rychlých analýz, protože výpočet

konformací proteinu je časově náročný a není vždy nutnou podmínkou pro získání relevantních informací.

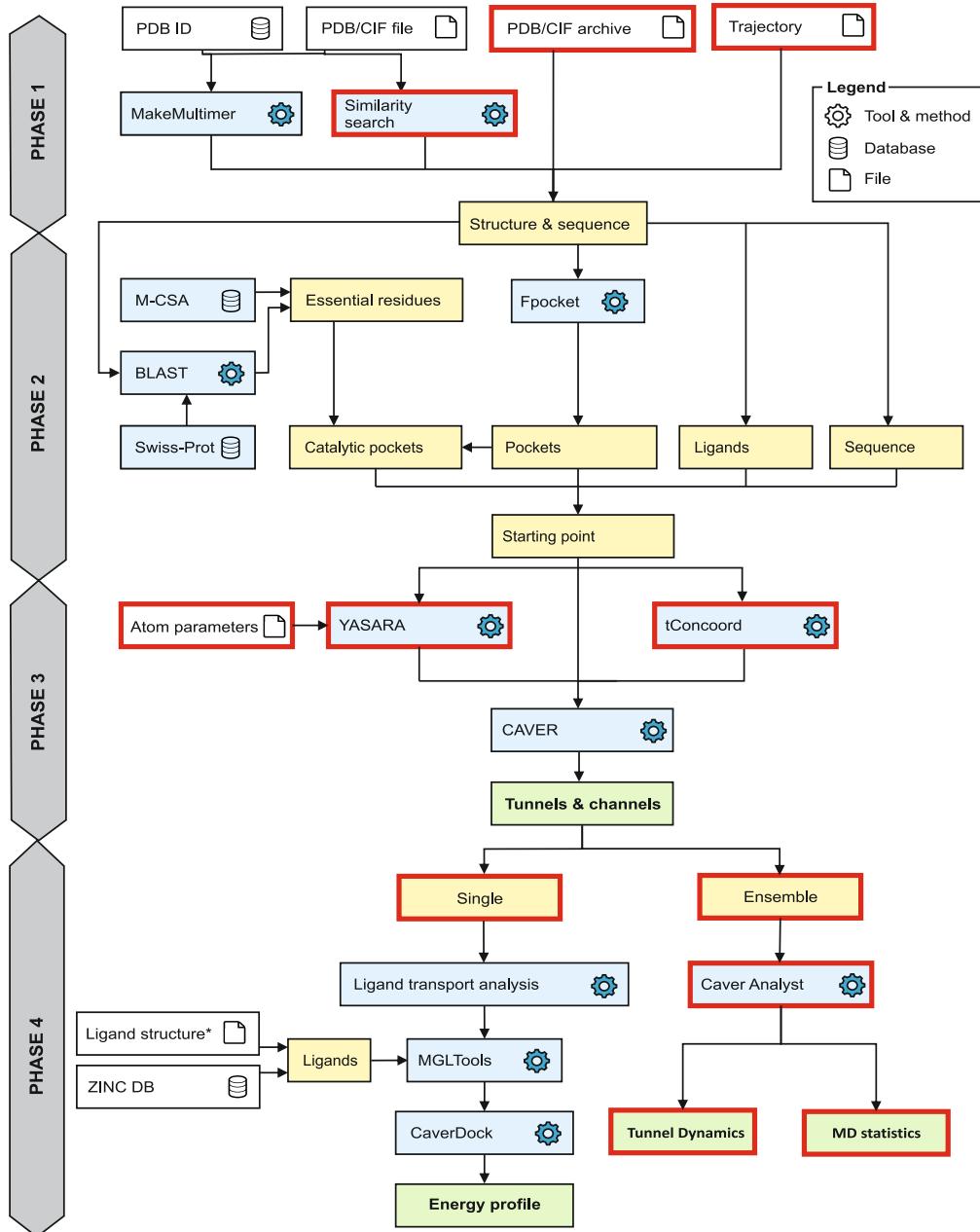
Vyhledání podobné struktury probíhá za účelem obecné analýzy více struktur v rámci jedné úlohy. Pro jejich hledání je využit program BLAST s požadavkem na 100% schodu v sekvenci s délkou odpovídající 50 až 100 % délky původní sekvence.

Další možností je volba výpočtu konformací pomocí molekulárních dynamik (obrázek 6) nebo nástroje tCONCOORD.

Následující kroky jsou pro uživatele prezentovány stejně jako v předchozí verzi: (i) definice startovacího bodu pro identifikaci tunelů; (ii) zadání vstupních parametrů pro nástroj Caver [17]. Po nastavení parametrů pro nástroj Caver je spuštěna sekvence výpočtů. V případě požadavku pro výpočet molekulárních dynamik je zahájen výpočet trajektorie. Po jeho ukončení je nad každým snímkem simulace spuštěna identifikace tunelů pomocí nástroje Caver.

Výpočet molekulárních dynamik provádí program Yasara [19], který je spuštěn prostřednictvím autory před-připraveného skriptu `md_run.mcr` [20]. Vstupem pro program je struktura proteinu ve formátu `pdb` a sada parametrů (nastavitelné parametry uživatelem jsou popsány v sekci 4.1). Běh molekulárních dynamik vyprodukuje velké množství snímků (konformací), přičemž každý snímek je uložen v samostatném `pdb` souboru. Snímky jsou poté vloženy do archivu, který je předán jako vstup nástroji Caver. Nástroj Caver již umí pracovat s trajektoriemi z molekulárních dynamik. Program postupně analyzuje jeden snímek za druhým a identifikuje v nich tunely. Tunely ve snímcích jsou následně přiřazeny identifikátory tzv. shluků. Identifikátor shluku označuje jeden tunel napříč snímkem, takže jsme tak schopni pozorovat proměny jednoho tunelu v čase. Výsledky jsou poté automaticky analyzovány a je vygenerován soubor, který obsahuje popis všech nalezených tunelů v simulaci a jejich celkové statistiky. Soubor je následně zpracován a jsou z něj extrahovány potřebné informace pro jednotlivé komponenty uživatelského rozhraní.

Aplikace se musí vyrovnat s velkým objemem zobrazených dat, protože jednorázové načtení všech dat do prohlížeče uživatele by ve většině případů způsobilo zahlcení jeho operační paměti a značnému zpomalení aplikace. Z tohoto důvodu jsou implementovány specializované koncové body API, které poskytují nejkonkrétnější výsledky na vyžádání klienta. Tyto koncové body jsou zodpovědné za předzpracování a transformaci všech dat z výsledného souboru a vážou se na jednotlivé komponenty. Díky tomuto přístupu se v operační paměti uživateli udržují data pouze o cel-



Obrázek 5. Původní workflow rozšířené o nové funkcionality (červeně zvýrazněné).

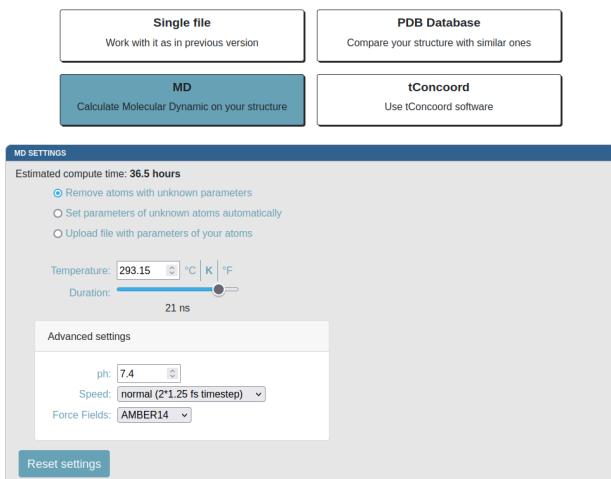
kové statistice simulace a jednom vybraném snímku.

5.0.1 Konečné zobrazení výsledků

Oproti původní verzi bylo zapotřebí umožnit analýzu transportu ligandu u jednotlivých snímků simulace a zobrazit souhrnné statistiky z běhu simulace. Konečný prohlížeč výsledků je složený ze dvou logických částí: celkové statistiky a analýzy jedné konformace. Logické části jsou reprezentovány pomocí 2 záložek.

První záložka obsahuje celkové statistiky a je zpřístupněná pouze v případě, kdy byly vytvořeny nebo

načteny trajektorie proteinu. Záložka celkových statistik obsahuje 4 komponenty. Sémantika těchto komponent je převzata z nástroje MoleCollar [23], ale v rámci aplikace byly tyto komponenty nově implementovány. První zobrazenou komponentou je teplotní mapa všech tunelů (obrázek 7H), ze které získá uživatel okamžitý přehled o trendech v systému jako jsou počet identifikovaných tunelů a jejich stabilita. Data o jednom tunnelu z jednotlivých snímků se sdružují do shluků (clusterů) a jsou analyzovány jako celky. Další in-



Obrázek 6. Volba očekávaného výstupu uživatelem při zpracování jedno souboru (s příkladem očekávaného spuštění molekulárních dynamik).

formací je proto zobrazení statistik identifikovaných shluků tunelů (obrázek 7J). Statistiky shluků obsahují četnost výskytu tunelů, prů-měry a směrodatné odchylky hrdla, členitosti a propustnosti tunelu. V další části může uživatel provést analýzu shluku tunelu v teplné mapě jednoho tunelu (obrázek 7I). Toto zobrazení je reprezentováno grafem, ve kterém je vykreslen průběh vybraného tunelu v jednotlivých snímcích od aktivního místa až k povrchu proteinu. Aplikace umožňuje vybrat průřez, ve kterém se provede výpočet kontur. Kontury se následně zobrazí v další komponentě (obrázek 7F).

Druhá záložka slouží k analýze statických struktur včetně jednotlivých snímků z molekulárních dynamik. Rozložení a funkcionalita komponent jsou stejné jako v předchozí verzi. Novou komponentou je výběr snímku molekulární dynamiky, která se zobrazí pouze při práci s trajektorií a umožňuje vybrat snímek, pro který lze provést všechny analýzy statické struktury.

6. Závěr

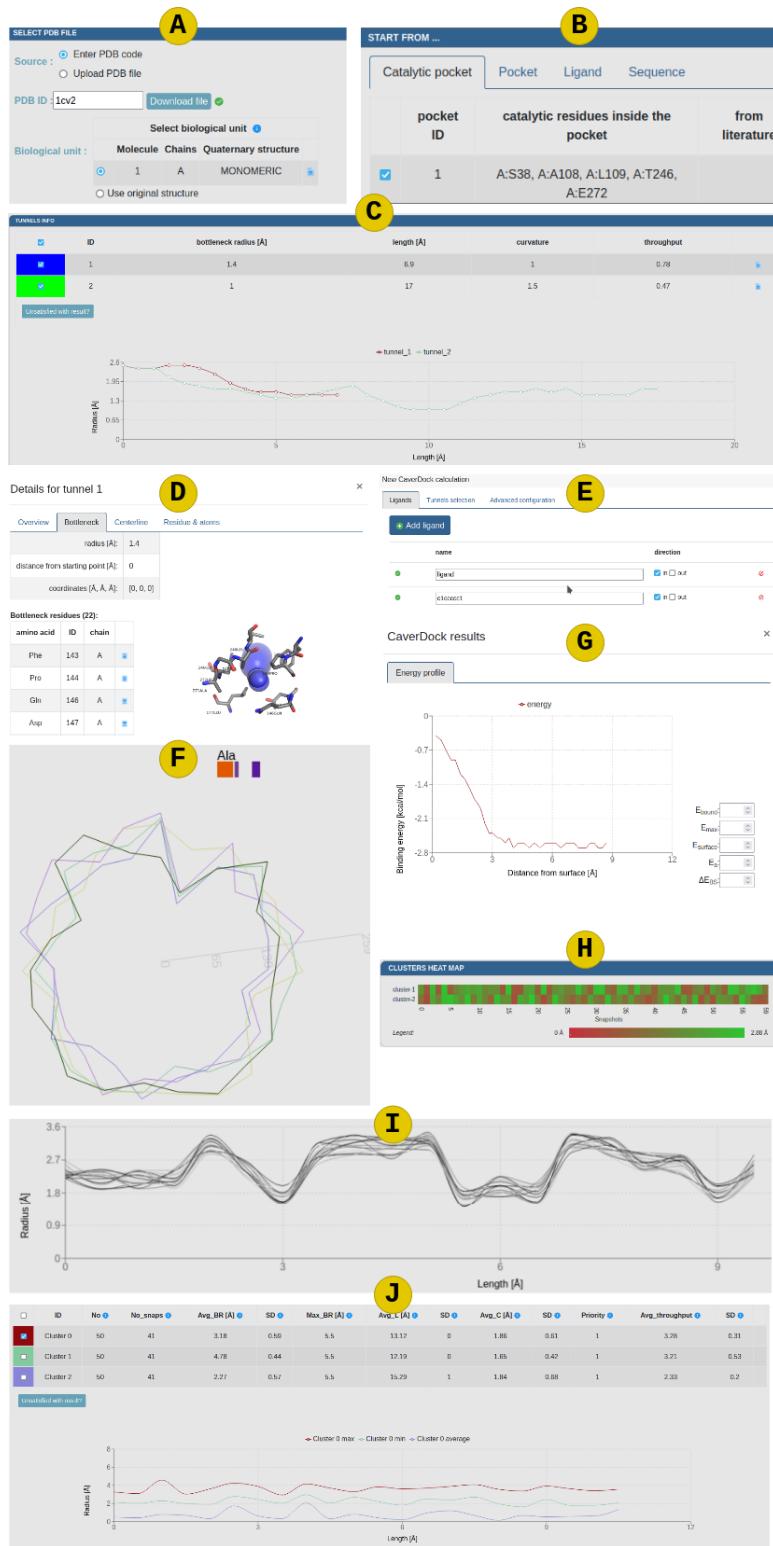
Caver Web 2.0 je nová verze webové aplikace, která se zabývá analýzou proteinových tunelů. Hlavním přínosem nové verze je integrace výpočtu a analýz molekulárních dynamik, které zvýší relevantnost získaných výsledků. V současnosti neexistuje obdobné řešení. Nástroj bude sloužit vědecké komunitě při vývoji nových nebo vylepšování stávajících proteinů.

7. Poděkování

Za odborné vedení a cenné rady bych chtěl poděkovat svému vedoucímu Mgr. Janu Štouračovi a celému týmu Loschmidtových laboratoří.

Literatura

- [1] Wikipedia. Protein structure — Wikipedia, the free encyclopedia. <http://en.wikipedia.org/w/index.php?title=Protein%20structure&oldid=995899200>, 2020. [Online; navštíveno 29. 12. 2020].
- [2] Bruce Alberts. *Základy buněčné biologie : Úvod do molekulární biologie buňky*. Espero Publishing, Ústí nad Labem, 1998.
- [3] Allan Svendsen. *Understanding Enzymes: Function, Design, Engineering, and Analysis*. CRC Press, 2016. Chapter 12: Marques S. M., Brezovsky J., Damborsky J.. *Role of tunnels, channels and gates in enzymatic catalysis*.
- [4] Jan Brezovsky, Eva Chovancova, Artur Gora, Antonin Pavelka, Lada Biedermannova, and Jiri Damborsky. Software tools for identification, visualization and analysis of protein tunnels and channels. *Biotechnology advances*, 31(1):38–49, 2013.
- [5] Georges Voronoi. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. premier mémoire. sur quelques propriétés des formes quadratiques positives parfaites. *Journal für die reine und angewandte Mathematik (Crelles Journal)*, 1908(133):97–102, 1908.
- [6] Edsger W Dijkstra et al. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271, 1959.
- [7] Hugosson W Håkan and Hans Ågren. Quantum mechanics/classical mechanics modeling of biological systems. In *Multiscale Modeling and Simulation in Science*, pages 291–294. Springer, 2009.
- [8] Reinhard Hentschke. Integrating the equations of motion. In *Classical Mechanics*, pages 123–153. Springer, 2017.
- [9] Jan Stourac, Ondrej Vavra, Piia Kokkonen, Jiri Filipovic, Gaspar Pinto, Jan Brezovsky, Jiri Damborsky, and David Bednar. Caver web 1.0: identification of tunnels and channels in proteins and analysis of ligand transport. *Nucleic acids research*, 47(W1):W414–W422, 2019.
- [10] Eitan Yaffe, Dan Fishelovitch, Haim J Wolfson, Dan Halperin, and Ruth Nussinov. Molaxis: a server for identification of channels in macromolecules. *Nucleic acids research*, 36(suppl_2):W210–W215, 2008.



Obrázek 7. Grafické rozhraní nástroje Caver Web 2.0. **(A)** Komponenta pro výběr biologické jednotky. **(B)** Panel pro výběr startovacího bodu obsahuje 4 možnosti výběru. **(C)** Tabulka detekovaných tunelů. **(D)** Podrobné informace o tunelu jsou umístěny ve vyskakovacím okně. **(E)** Okno pro zadání nových výpočtů transportu ligandu. **(F)** Tunelové kontury s vlivem jedné aminokyseliny. **(G)** Výsledky výpočtů transportu ligandu – energetické profily. **(H)** Teplotní mapa všech tunelů (byly identifikovány 2 tunely a trajektorie obsahují 50 snímků). **(I)** Zvolená reprezentace teplotní mapy pro jeden tunel (průběh šířky tunelu je zachycen v grafu). **(J)** Statistiky shluků tunelů v celé simulaci.

[11] Lukáš Pravda, David Sehnal, Dominik Toušek, Veronika Navrátilová, Václav Bazgier, Karel

Berka, Radka Svobodová Vařeková, Jaroslav Koča, and Michal Otyepka. Moleonline: a web-

- based tool for analyzing channels, tunnels and pores (2018 update). *Nucleic acids research*, 46(W1):W368–W373, 2018.
- [12] Jae-Kwan Kim, Youngsong Cho, Mokwon Lee, Roman A Laskowski, Seong Eon Ryu, Kokichi Sugihara, and Deok-Soo Kim. Betacavityweb: a webserver for molecular voids and channels. *Nucleic acids research*, 43(W1):W413–W418, 2015.
- [13] Marialuisa Pellegrini-Calace, Tim Maiwald, and Janet M Thornton. Porewalker: a novel tool for the identification and characterization of channels in transmembrane proteins from their three-dimensional structure. *PLoS computational biology*, 5(7):e1000440, 2009.
- [14] Talha Bin Masood, Sankaran Sandhya, Nagasuma Chandra, and Vijay Natarajan. Chexvis: a tool for molecular channel extraction and visualization. *BMC bioinformatics*, 16(1):1–19, 2015.
- [15] Juan Cortés, Duc Thanh Le, Romain Iehl, and Thierry Siméon. Simulating ligand-induced conformational changes in proteins using a mechanical disassembly method. *Physical Chemistry Chemical Physics*, 12(29):8268–8276, 2010.
- [16] Aleksandras Gutmanas, Younes Alhroub, Gary M Battle, John M Berrisford, Estelle Bochet, Matthew J Conroy, Jose M Dana, Manuel A Fernandez Montecelo, Glen van Ginkel, Swanand P Gore, et al. Pdbe: protein data bank in europe. *Nucleic acids research*, 42(D1):D285–D291, 2014.
- [17] Eva Chovancova, Antonin Pavelka, Petr Benes, Ondrej Strnad, Jan Brezovsky, Barbora Kozlikova, Artur Gora, Vilem Sustr, Martin Klvana, Petr Medek, et al. Caver 3.0: a tool for the analysis of transport pathways in dynamic protein structures. *PLoS Comput Biol*, 8(10):e1002708, 2012.
- [18] Ondrej Vavra, Jiri Filipovic, Jan Plhak, David Bednar, Sergio M Marques, Jan Brezovsky, Jan Stourac, Ludek Matyska, and Jiri Damborsky. Caverdock: a molecular docking-based tool to analyse ligand transport through protein tunnels and channels. *Bioinformatics*, 35(23):4986–4993, 2019.
- [19] Henrik Land and Maria Svedendahl Humble. Yasara: a tool to obtain structural guidance in biocatalytic investigations. In *Protein Engineering*, pages 43–67. Springer, 2018.
- [20] Pre-made script in yasara written by elmar krieger for molecular dynamics simulation. http://www.yasara.org/md_run.mcr. Accessed: 2022-03-21.
- [21] Deborah Leckband and Jacob Israelachvili. Intermolecular forces in biology. *Quarterly reviews of biophysics*, 34(2):105, 2001.
- [22] Daniel Seeliger, Jürgen Haas, and Bert L de Groot. Geometry-based sampling of conformational transitions in proteins. *Structure*, 15(11):1482–1492, 2007.
- [23] Jan Byška, Adam Jurčík, M Eduard Gröller, Ivan Viola, and Barbora Kozlikova. Molecular and tunnel heat map visualizations for conveying spatio-temporo-chemical properties across and along protein voids. In *Computer Graphics Forum*, volume 34, pages 1–10. Wiley Online Library, 2015.
- [24] Makemultimer script. <http://watcut.uwaterloo.ca/tools/makemultimer/>. Accessed: 2022-03-21.