ELEKTROTEHNIŠKI VESTNIK 91(4): 210–216, 2024
IZVIRNI ZNANSTVENI ČLANEK

Nadgradnja cenilne funkcije molekulskega sidranja po vzoru orodja PLANTS

Primož Zidanšek [1], **Črtomir Podlipnik** [2], **Davor Sluga** [1], **Nejc Ilc** [1]

[1] *Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Večna pot 113, 1000 Ljubljana, Slovenija*
[2] *Univerza v Ljubljani, Fakulteta za kemijo in kemijsko tehnologijo, Večna pot 113, 1000 Ljubljana, Slovenija*
E-pošta: pz8920@student.uni-lj.si

Povzetek. Nadgradili smo odprtokodno orodje za izvajanje molekulskega sidranja CmDock z izboljšano cenilno funkcijo CHEMPLP, ki je razširjena različica obstoječe funkcije PLP. To smo dosegli z dodatkom vodikovih vezi po vzoru orodja PLANTS. Kljub dodatkom kompleksnejših interakcij natančnost cenilke CHEMPLP ni pokazala bistvenih izboljšav, saj sta tako PLP kot CHEMPLP dosegli povprečni RMSD okoli 9 Å na testni množici DUD-E. Najnatančnejša je ostala cenilka VDW, ki je privzeta cenilka orodja CmDock, s povprečnim RMSD okoli 7 Å. A čeprav je VDW natančnejša, sta PLP in CHEMPLP pokazali prednost v hitrosti, saj je čas sidranja pri njuni uporabi približno pol krajši. To nakazuje njuno potencialno uporabnost pri časovno potratnih procesih, kot je visokozmogljivo virtualno rešetanje.

Ključne besede: CmDock, cenilka PLP, cenilka CHEMPLP, PLANTS

Upgrade of the scoring function of molecular docking modeled after the PLANTS tool

We have upgraded the open-source molecular docking tool CmDock with an improved CHEMPLP scoring function, which is an extended version of the existing PLP function. We achieved this by adding hydrogen bonds following the example of the PLANTS tool. Despite these additional complex interactions, the CHEMPLP estimation accuracy did not show significant improvement, with both PLP and CHEMPLP achieving an average RMSD of around 9 Å on the DUD-E test set. The VDW scoring function remained the most accurate, with an average RMSD of around 7 Å. Although VDW is more accurate, PLP and CHEMPLP showed a speed advantage, with their docking times approximately halved. This suggests their potential utility in time-consuming processes such as high-throughput virtual screening.
Keywords: CmDock, PLP scoring function, CHEMPLP scoring function, PLANTS

1 UVOD

Molekulsko sidranje je metoda, ki simulira proces vezave majhnih molekul, kot so ligandi, na specifična vezavna mesta v beljakovinah. Beljakovine so ključne biološke makromolekule, odgovorne za številne celične procese, kot so encimske reakcije, signalizacija in transport. Vezava liganda na beljakovino lahko močno vpliva na njeno strukturo in funkcijo, kar posledično vpliva tudi na fiziologijo celice, v kateri je beljakovina prisotna. Ta interakcija je bistvenega pomena v številnih bioloških procesih, predvsem pa v farmakologiji, kjer je pogosto cilj razviti učinkovit ligand, ki bi moduliral funkcijo določene beljakovine, kar bi privedlo do terapevtskega učinka [1], [2]. Pomembno je, da se ligand dobro prilega beljakovini, saj to vodi do boljšega prenosa signala v celico in s tem večjega celičnega odziva.

Orodje za molekulsko sidranje, ki ga razvija slovenska ekipa, nosi ime CmDock* (Curie Marie Dock). To orodje poskuša za podani par ligand-beljakovina najti položaj liganda, ki se optimalno prilega beljakovini. Za preverjanje točnosti orodij za sidranje primerjamo rezultate z referenčnim položajem, v katerega bi se ligand postavil v realnem okolju, kar določajo fizikalni eksperimenti, kot sta rentgenska kristalografija in jedrska magnetna resonanca. Ti referenčni položaji so na voljo v bazi podatkov beljakovin (Protein Data Bank, PDB)†, prav tako pa lahko uporabimo referenčne nabore kompleksov ligand-beljakovina za medsebojno primerjavo uspešnosti različnih orodij za molekulsko sidranje.

Postopki molekulskega sidranja se med seboj razlikujejo po štirih ključnih vidikih:

1) **Fleksibilnost receptorja**. Sidranje lahko obravnava receptor kot povsem fleksibilen, z gibljivimi le določenimi funkcionalnimi skupinami ali povsem rigiden. CmDock privzeto uporablja rigidni receptor.
2) **Fleksibilnost liganda**. V CmDocku je ligand fleksibilen, in sicer mu lahko spreminjamo:
   a) položaj težišča v prostoru (tri koordinate),
   b) orientacijo molekule (trije Eulerjevi koti) in
   c) orientacijo okoli vsake vrtljive vezi (kar vpliva na torzijo in končno obliko liganda).

   Število prostostnih stopenj liganda je torej enako

Prejet 25. september, 2024
Odobren 10. oktober, 2024

*Orodje je dostopno na https://gitlab.com/Jukic/cmdock.
†Dostopna je na http://www.rcsb.org/pdb/home/home.do.

6 + število vrtljivih vezi.

3) **Cenilne funkcije**. Te funkcije ocenjujejo položaj liganda glede na receptor in mu priredijo številsko vrednost, ki je približek vezavne energije. Manjša je vrednost, močnejša naj bi bila vezava. CmDock podpira več cenilk, mi pa se osredotočamo na PLP (angl. piecewise linear potential), CHEMPLP (angl. chem piecewise linear potential) in VDW (Van der Waals).
4) **Optimizacijski algoritem**. Algoritem spreminja parametre liganda in išče položaj, pri katerem je vrednost cenilke najnižja. CmDock v glavnem uporablja genetski algoritem, ki začenja z naključno generirano populacijo ligandov ter nato prek križanja in mutacij išče optimalne rešitve.

Osredotočamo se predvsem na cenilke, in ne na optimizacijske algoritme, saj je naš cilj nadgraditi že obstoječo cenilko PLP s kompleksnejšo različico CHEMPLP. Pri tem nam je bila v veliko pomoč izvorna koda orodja PLANTS [5], ki vsebuje implementacijo cenilke CHEMPLP, ki jo želimo integrirati v CmDock.

2 CENILKA CHEMPLP

Rezultat molekulskega sidranja je toliko boljši, kolikor bolj se položaj sidrane molekule približa referenčnemu. Ključni dejavnik, ki vpliva na kakovost sidranja oziroma podobnost referenčnemu položaju, je sposobnost cenilke, da pravilno simulira dejanske fizikalne sile med ligandom in receptorjem. Vendar pa cenilka ne sme biti preveč kompleksna, da jo je mogoče dovolj hitro izračunati.

Cenilki, ki nas zanimata, sta PLP in CHEMPLP. Cenilka PLP je definirana kot vsota štirih členov:

$$\mathrm{PLP} = f_{\mathrm{plp}} + f_{\mathrm{trki}} + f_{\mathrm{torzija}} + f_{\mathrm{kazen}} \,. \tag{1}$$

Prvi člen (f_{plp}) modelira različne privlačne in odbojne sile med atomi liganda in receptorja, kot so vodikove, kovinske, nepolarne in prikrite interakcije.

Člen f_{trki} zagotavlja, da se atomi liganda med seboj ne približajo preveč, kar se lahko zgodi z obračanjem vrtljivih vezi. Težišče in orientacija liganda ne vplivata na njegovo obliko.

Člen f_{torzija} išče kot okrog vsake vrtljive vezi, ki omogoča energijsko najugodnejše medsebojne položaje teh atomov okoli vezi.

Člen f_{kazen} kaznuje ligande, ki se znajdejo zunaj votline beljakovine, kjer naj bi bili sidrani.

Podrobnejšo razlago PLP-ja sta v svojih diplomskih nalogah podala Tine Erent [8] in Miha Kovač [7].

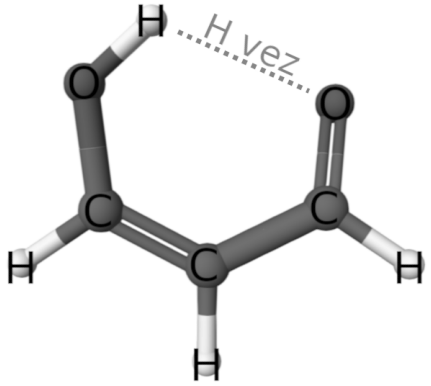
Cenilka CHEMPLP poleg osnovnih štirih vključuje še sedem dodatnih členov, ki natančneje opisujejo vodikove privlačne sile ter privlačne sile med ligandi in kovinami:

$$\begin{aligned}\mathrm{CHEMPLP} = \mathrm{PLP} \\ + f_{\mathrm{HB}} + f_{\mathrm{HB\text{-}nab}} + f_{\mathrm{HB\text{-}CHO}} \\ + f_{\mathrm{kov}} + f_{\mathrm{kov\text{-}nab}} + f_{\mathrm{kov\text{-}koord}} + f_{\mathrm{kov\text{-}koord\text{-}nab}} \,.\end{aligned} \tag{2}$$

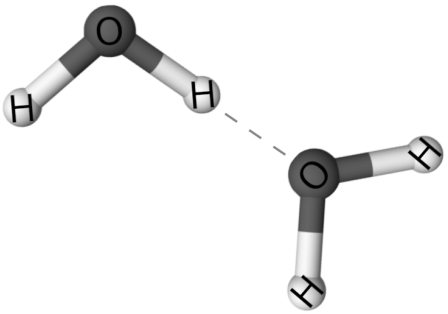
V nadaljevanju se bomo osredotočili na teorijo in implementacijo vodikovih vezi, zlasti na člene f_{HB}, $f_{\mathrm{HB\text{-}nab}}$ in $f_{\mathrm{HB\text{-}CHO}}$, medtem ko bomo obravnavo kovinskih vezi (členi f_{kov}, $f_{\mathrm{kov\text{-}nab}}$, $f_{\mathrm{kov\text{-}koord}}$ in $f_{\mathrm{kov\text{-}koord\text{-}nab}}$) prepustili prihodnjim nadgradnjam.

3 VODIKOVE SILE

Vodikova vez je specifična vrsta medmolekulske ali notranje molekulske interakcije, ki nastane med vodikovim atomom, vezanim na zelo elektronegativen atom (npr. fluor, kisik ali dušik), in neveznim elektronskim parom drugega atoma elektronegativnega elementa [9]. To je prikazano na sliki 2. Vodik, ki odda svoj proton, deluje kot donor, medtem ko drugi elektronegativni element deluje kot akceptor. V simulaciji smo obravnavali samo medmolekulske vodikove vezi, kjer sta donor in akceptor prisotna v različnih molekulah. Razliko med notranjimi in medmolekulskimi vodikovimi vezmi ponazarjata sliki 1 in 2.



Slika 1:: Notranje molekulska vodikova vez v malondialdehidu.

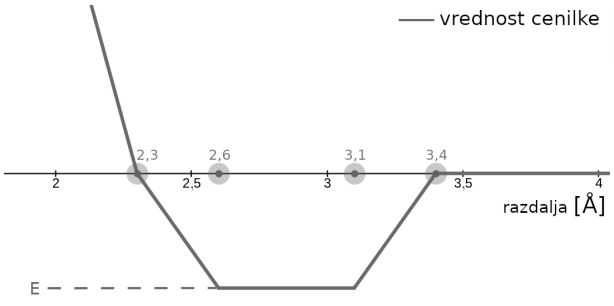


Slika 2:: Medmolekulska vodikova vez med molekulama vode.

3.1 Vodikove interakcije v cenilki PLP

V prvem členu, f_{plp}, ki predstavlja različne medmolekulske interakcije med atomi, so prisotne tudi vodikove vezi. Njihov vpliv je prikazan na grafu na sliki 3. Če je razdalja med donorjem in akceptorjem prevelika,

vodikove vezi k cenilki ne prispevajo nič. Če sta donor in akceptor na optimalni razdalji med 2,3 in 3,4 Å*, vodikova vez doseže energijsko najugodnejše stanje (prispevek je negativen). Če sta preblizu, pa prevlada odbojni učinek.



Slika 3:: Graf člena f_{plp} vodikove vezi v odvisnosti od razdalje med donorjem in akceptorjem vodikove vezi.

3.2 Vodikove interakcije v cenilki CHEMPLP

Cenilka CHEMPLP natančneje simulira vodikove in kovinske vezi v primerjavi s PLP, saj upošteva več parametrov. Člen za vodikove vezi (f_{HB}) v enačbi (2) se od člena f_{plp} razlikuje po tem, da poleg razdalje vključuje kot, ki ga tvori akceptor s svojimi sosedi,† in kot, ki ga tvori donor s svojim sosedom (donor vodik ima vedno le enega soseda).

3.3 Členi vodikovih vezi

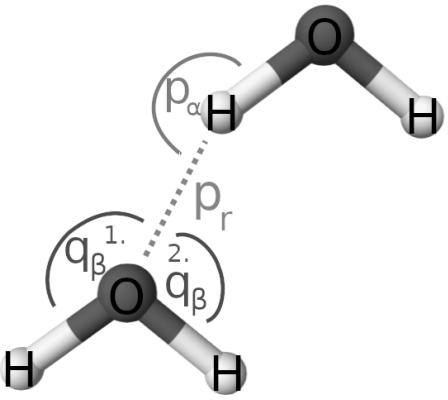
Člen f_{HB} se izračuna po enačbah (3) in (4):

$$\begin{aligned} f_{\text{HB}} = w_{\text{HB}} \cdot \sum_{p \in P_{\text{HB}}} \big[& f(|p_{\text{r}} - 1{,}85|, 0{,}25, 0{,}65) \\ & \cdot f(|p_\alpha - 180|, 30, 80) \\ & \cdot \prod_{q \in Q_{\text{akc-sos}}} f(|q_\beta - 180|, 80, 100) \big] \,, \end{aligned} \tag{3}$$

$$f(x, x_1, x_2) = \begin{cases} 1 & x \leq x_1 \\ (x_2 - x)/(x_2 - x_1) & x_1 < x \leq x_2 \\ 0 & x_2 < x \,. \end{cases} \tag{4}$$

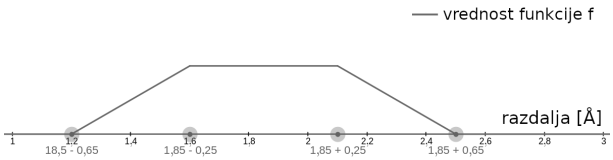
P_{HB} predstavlja vse potencialne pare atomov (donorji in akceptorji), ki tvorijo vodikove vezi, $Q_{\text{akc-sos}}$ pa vključuje vse sosednje atome akceptorja. Na sliki 4 je prikazano, da je p_{r} razdalja med donorjem in akceptorjem, p_α kot z vrhom v donorju, q_β pa kot z vrhom v akceptorju. Molekule, kot je voda, imajo več sosednjih atomov, kar pomeni, da obstajajo različni koti q_β.

Za vsakega izmed treh delov enačbe 3 znotraj vsote velja oblika grafa, kot je vidna na sliki 5. Ker so uteži vodikovih vezi (npr. w_{HB}) negativne, si lahko predstavljamo graf, obrnjen okoli osi x. Zato člen nagradi tiste pare donor-akceptor, ki imajo ravno prave vrednosti



Slika 4:: Medmolekulska vodikova vez med dvema atomoma z označeno razdaljo in koti.

razdalj in kotov z vrhom v donorju in akceptorju. Kazni ali odbojnih sil ta člen ne simulira, saj za to skrbi že člen f_{plp}.



Slika 5:: Graf funkcije f v odvisnosti od razdalje med donorjem in akceptorjem vodikove vezi, če ji podamo argumente $|p_{\text{r}} - 1{,}85|$, $0{,}25$ in $0{,}65$, kot je v prvem delu enačbe 3.

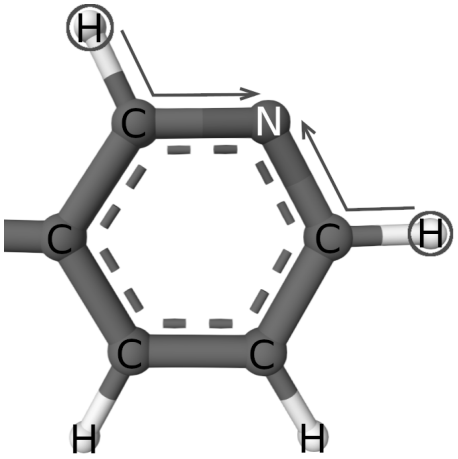
Člen $f_{\text{HB-nab}}$ se izračuna enako kot f_{HB}, le da se na koncu njegova vrednost pomnoži še z dodatno utežjo (z vrednostjo okoli 2 ali 3), s čimer se simulira večja moč teh vodikovih vezi. Kandidati za stvaritev nabite vodikove vezi sta negativno nabiti akceptor in pozitivno nabiti donorjev sosed. V takšnih primerih akceptor še močneje od sebe oddaja elektrone oziroma nase veže protone, medtem ko donorjev sosed dodatno stabilizira interakcijo s tem, da krepi privlak med pozitivno nabitim vodikovim atomom in negativno nabitim akceptorjem.

Člen $f_{\text{HB-CHO}}$ je posebna vrsta vodikovih vezi, ki se računa podobno kot f_{HB}, le da ima malo drugačne argumente v funkciji f (glej enačbo 5). Ta vrsta vezi želi bolj oddaljene pare donor-akceptor in dovoli večje kote z vrhom v donorju. Za donorja vodikove vezi v tem primeru štejemo vodik, ki ima vezan ogljik, ta pa je vezan na dušik. Ogljik in dušik morata biti v aromatičnem obroču kot na sliki 6. Akceptor vodikove vezi je v tem primeru lahko le kisik [5].

*Ángstrem je enota za dolžino, ki se uporablja za opisovanje razdalj na atomarnem nivoju. 1 Å znaša 10^{-10} metra ali 0,1 nanometra.

†Sosed je atom, s katerim si deli vez.

$$f_{\text{HB-CHO}} = w_{\text{HB-CHO}} \cdot \sum_{p \in P_{\text{HB-CHO}}} \big[f(|p_{\text{r}} - 2{,}35|, 0{,}25, 0{,}65) \cdot f(|p_\alpha - 180|, 50, 100) \cdot \prod_{q \in Q_{\text{akc-sos}}} f(|q_\beta - 180|, 80, 100) \big] . \quad (5)$$



Slika 6:: Obkrožena sta donorja vodikove vezi za $f_{\text{HB-CHO}}$. S puščicama sta označena sosednji ogljikov atom in nato še dušikov atom, zaradi katerih sta ta dva vodikova atoma donorja.

4 IMPLEMENTACIJA CENILKE CHEMPLP

Za učinkovito delovanje genetskega algoritma je ključno, da cenilno funkcijo izvedemo čim hitreje, saj se izračuna za vsak osebek v vsaki generaciji algoritma. To dosežemo z vnaprejšnjim izračunom in shrambo vseh vrednosti, ki se med postopkom ne spreminjajo, kar imenujemo priprava liganda in receptorja. Med procesom ustvarimo podatkovne strukture, ki hranijo te vnaprej izračunane podatke.

4.1 Priprava receptorja in liganda

Najprej je treba identificirati vse atome v molekuli in jih razvrstiti v kategorije, kot so vodik, donor, akceptor in nepolarni atom. Nato se ponovno sprehodimo skozi molekulo, da bi poiskali vse donorje in akceptorje ter zanje ustvarili ustrezne ovojne razrede, ki vključujejo vektorje do njihovih sosednjih atomov, kot tudi začetne vrednosti za interakcije vodikovih vezi in druge potrebne parametre. S tem zaključimo pripravo liganda.

Pripravo receptorja še nadaljujemo. Da bi bilo iskanje bližnjih atomov receptorja med samim izračunom cenilke učinkovitejše, za receptor zgradimo 3D-mrežo. Vanjo razporedimo donorje na receptorju in jih vstavimo v mrežne točke, ki ustrezajo njihovim lokacijam, ter v sosednje točke mreže. Tako mreža v vsaki točki vsebuje donorje, ki lahko vplivajo na to območje z vodikovimi vezmi. Enak postopek izvedemo za akceptorje receptorja, vendar v drugi mreži.

4.2 Izračun cenilke

Ko začnemo izračun cenilne funkcije, pregledamo vse donorje liganda. S pomočjo njihove lokacije v mreži receptorja poiščemo bližnje akceptorje. Vsi akceptorji v bližini donorja so upoštevani, vendar k vrednosti cenilke prispeva le najmočnejša vez glede na rezultat enačbe 3. To vrednost nato dodamo skupnemu seštevku cenilne funkcije. Enak postopek uporabimo tudi za vse akceptorje liganda.

Algoritma ne ločimo na del za normalne, nabite in CHO-vodikove vezi, vendar vse obravnavamo enako. Pri pripravi molekul atom klasificiramo kot donor, če bi bil donor katerekoli vrste vodikovih vezi, in enako za akceptor. Na koncu računanja vodikove vezi le preverimo, ali je donor vezan na ogljik. Če je, uporabimo utež za CHO, drugače za normalno vodikovo vez. V zadnjem primeru še pomnožimo vrednost z dodatno utežjo, če sta donorjev sosed in akceptor nabita. Tako razlikujemo med različnimi tipi vodikovih vezi znotraj istih zank.

Algoritem ima časovno zahtevnost, ki je linearna glede na število donorjev in akceptorjev v ligandu, čeprav vključuje trojno zanko. Notranji dve zanki se izvajata v konstantnem času, saj je druga zanka odvisna od števila receptorjevih atomov v bližini posameznega donorja ali akceptorja, kar je omejeno na končno število atomov znotraj določenega območja. Notranja zanka pa je odvisna od števila sosednjih atomov akceptorja, ki jih je lahko tudi le končno mnogo.

5 REZULTATI

Vse meritve so bile izvedene na prenosniku, opremljenem z glavnim procesorjem Intel i7-12700H, ki omogoča obdelavo z 20 strojnimi nitmi. Za izvajanje eksperimentov smo uporabili izvorno kodo orodja CmDock [3], ki je napisana v programskem jeziku C++. Za primerjavo smo uporabili tudi izvorno kodo orodja PLANTS (v jeziku C++), ki pa ni javno dostopna.

5.1 Primerjava z orodjem PLANTS

Naš cilj je bil integrirati cenilko CHEMPLP v CmDock z uporabo izvorne kode PLANTS, pri čemer bi morale biti vrednosti cenilk za enake položaje liganda v obeh orodjih primerljive. Da bi to preverili, smo analizirali štiri različne komplekse, podrobneje opisane v tabeli 1. Da bi ocenili delovanje cenilke na bolj raznoliki množici, smo generirali dva naključna položaja liganda za vsak kompleks s spremembo njegovih parametrov. Skupaj smo pridobili 12 različnih položajev, ki smo jih uporabili za primerjavo rezultatov cenilke CHEMPLP med CmDock in PLANTS. Obe orodji sta za vse vrste vodikovih vezi skupaj dobili povprečno vrednost enako -3,99, kar pomeni da smo uspešno implementirali vodikove vezi v orodju CmDock, da so v skladu z orodjem PLANTS.

kompleks	ligand	receptor
3CS9	nilotinib	ABL kinaza
3NJG	proteaza bakterije Shewanella oneidensis	beljakovina SO1698
3PTB	benzamidin	tripsin
6O0K	venetoklaks	beljakovina BCL-2

Tabela 1:: Osnovni podatki o štirih uporabljenih kompleksih, s katerimi smo primerjali orodji CmDock in PLANTS.

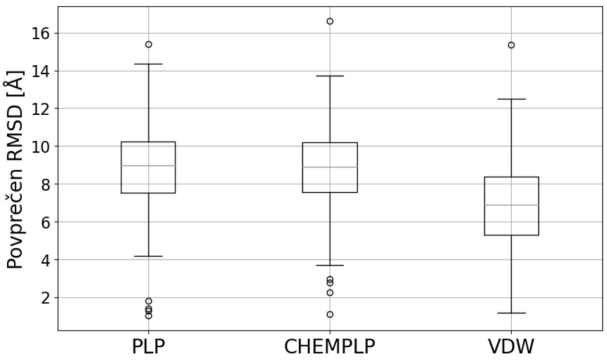
5.2 Točnost napovedanih položajev

Za oceno točnosti sidranja v orodju CmDock smo uporabili zbirko kompleksov DUD-E (Directory of Useful Decoys – Extended) [10], ki je prosto dostopna na spletu*. Obsega 102 različna kompleksa ligand-beljakovina, pri čemer za vsako beljakovino vključuje tudi vabe (angl. decoys), molekule, ki so strukturno podobne ligandom, vendar se ne vežejo na tarčno beljakovino in s tem služijo za bolj kritično validacijo orodij molekulskega sidranja.

Postopek sidranja smo izvedli tako, da smo najprej za vsak par ligand-beljakovina identificirali votlino ali luknjo v beljakovini, kjer bo potekalo iskanje optimalnega položaja liganda. Nato smo za vsak par izvedli sidranje, pri čemer smo optimizacijski algoritem nastavili z zastavico `-n` na vrednost 1, kar pomeni, da smo optimizacijo izvedli enkrat. V primeru vrednosti 10 bi orodje izbralo najboljše položaje med desetimi ponovitvami. Dodatno smo uporabili zastavico `-H`, da smo vključili vse atome vodika, saj bi se sicer pri branju molekul iz datoteke izpustili atomi vodika, ki so vezani na ogljik.

Točnost napovedanih oziroma sidranih položajev smo ocenili s primerjavo z referenčnimi položaji, pri čemer smo izračunali koren povprečne kvadrirane deviacije (RMSD) med ujemajočimi se težkimi atomi (brez vodikov). Nižja vrednost RMSD pomeni boljše ujemanje napovedanega položaja z referenčnim. Za pravilno napovedan položaj se po navadi šteje tisti, pri katerem je RMSD manjši od 2,0 Å, saj je ta meja znotraj eksperimentalne natančnosti pridobljenih referenčnih položajev. RMSD smo izračunali s pomočjo knjižnice RDKit†.

Sidranje s CmDockom smo izvedli na 101 kompleksu zbirke DUD-E, saj je bil kompleks BRAF izključen zaradi težav pri branju s knjižnico RDKit. Sidranje smo ponovili 30-krat za vsako cenilko (PLP, CHEMPLP in VDW), da smo izračunali srednje vrednosti in tako zmanjšali vpliv šuma. Cenilka VDW, ki temelji na simulaciji Van der Waalsovih interakcij in je privzeta empirična cenilka CmDocka, je uporabljena za primerjavo poleg PLP in CHEMPLP. Rezultati RMSD za vse tri cenilke so prikazani na sliki 7.



Slika 7:: Graf povprečnih vrednosti kompleksov RMSD, ki je prikazan s škatlami z brki [12]. Siva črta znotraj vsake škatle prikazuje mediano, škatla se razteza od prvega do tretjega kvartila, na robovih pa so brki. Vsi primeri, ki ne spadajo znotraj brkov ali škatle, so osamelci, označeni s krogom.

Naša hipoteza je bila, da bo implementacija simulacije vodikovih vezi v cenilki CHEMPLP izboljšala njeno natančnost v primerjavi s PLP, a smo še vedno pričakovali, da bo cenilka VDW dosegla najboljše rezultate zaradi svoje kompleksnejše strukture in optimizacije za CmDock. Rezultati so pokazali, da med PLP in CHEMPLP ni bilo bistvenih razlik, saj sta obe dosegli vrednosti RMSD med 8 in 10 Å za večino kompleksov. Pravilnost simulacije vodikovih vezi smo potrdili s primerjavo s cenilko PLANTS, zato težava ne tiči v napačni implementaciji. Cenilka VDW se je najbolje izkazala, saj je večina vrednosti RMSD padla med 5 in 8 Å, kar potrjuje našo prvotno hipotezo, v nasprotju s cenilko CHEMPLP. Že v predhodnih testiranjih [7] se je cenilka VDW izkazala bolje kot PLP.

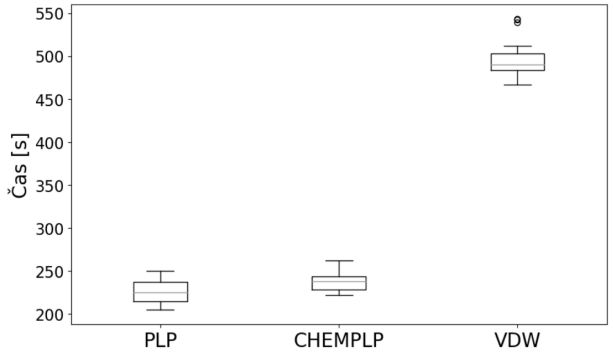
Dodajanje podpore za simulacijo kovinskih vezi v CHEMPLP ne bi vplivalo na rezultate, saj nobeden izmed receptorjev v zbirki DUD-E ne vsebuje kovinskih atomov. V takšnih primerih so vsi kovinski členi v enačbi 2 ničelni.

5.3 Hitrost izvajanja sidranja

Med meritvami smo izračunali čas izvajanja enega sidranja nad množico DUD-E vsake cenilke. Ker smo sidranje izvedli 30-krat za vsako cenilko, smo pridobili 30 časovnih vrednosti za vsako, ki so prikazane na sliki 8. Rezultati kažejo, da je cenilka CHEMPLP le nekoliko počasnejša od PLP, kar je mogoče pripisati simulaciji vodikovih vezi, ki jih PLP ne podpira. Cenilka VDW pa zahteva približno dvakrat toliko časa za eno sidranje v primerjavi s preostalima, kar je posledica njene kompleksnejše narave.

Poleg časovne zahtevnosti smo merili tudi število generacij genetskega algoritma, potrebnih za ustavitev.

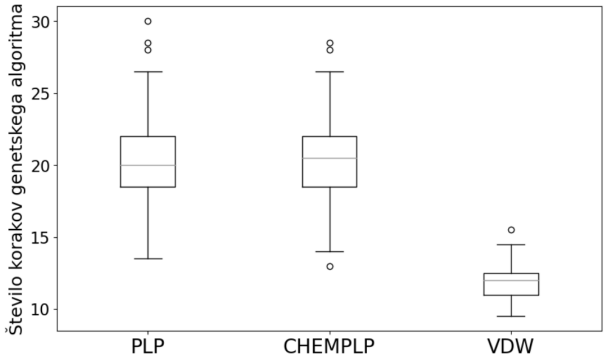
*https://dude.docking.org
†https://rdkit.org/.



Slika 8:: Časi izvajanja sidranja nad množico DUD-E.

CmDock ne uporablja fiksnega števila generacij kot ustavitveni pogoj. Namesto tega se algoritem ustavi, ko se ocena najboljšega osebka na podlagi cenilke ne spremeni skozi šest zaporednih generacij.

Natančnejša cenilka običajno zmanjša število potrebnih korakov genetskega algoritma, saj ga učinkovitejše usmerja proti globalnemu ali lokalnemu optimumu. Za vsak kompleks smo po 30 ponovitvah sidranja izračunali mediane števila generacij. Rezultati za vseh 101 kompleksov so prikazani na sliki 9. Kot pričakovano, je bilo za cenilko VDW potrebnih manj korakov do ustavitve v primerjavi s preostalima, medtem ko med PLP in CHEMPLP ni bilo večjih razlik.

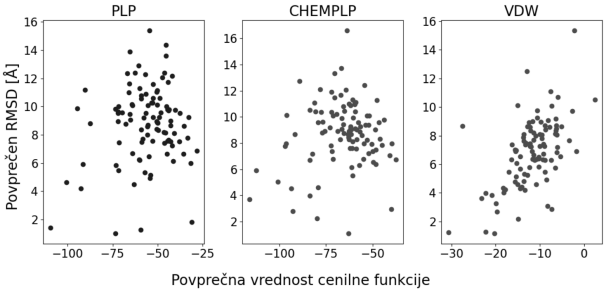


Slika 9:: Število generacij genetskega algoritma do ustavitve pri uporabi različnih cenilk.

5.4 RMSD in cenilne funkcije

Čeprav iz vrednosti cenilnih funkcij ne moremo neposredno sklepati o RMSD, med njima obstaja posredna povezava, s pomočjo katere lahko sklepamo tudi o vezavni energiji. Nižje vrednosti cenilnih funkcij običajno nakazujejo boljše prileganje liganda, kar se pogosto, vendar ne vedno, ujema z nižjimi vrednostmi RMSD. Na sliki 10 je prikazan graf raztrosa RMSD v odvisnosti od vrednosti cenilnih funkcij za posamezno cenilko. Pri PLP in CHEMPLP ni očitne linearne odvisnosti, medtem ko je pri VDW že videti bolj opazen trend.

Večji ligandi običajno povzročijo višje absolutne vrednosti cenilnih funkcij, saj več atomov prispeva k interakcijam med ligandom in receptorjem. Posledično prisotnost več parov atomov vpliva na večje vrednosti členov, kot so f_{plp}, in vodikovih vezi. Zato je smiselno, da cenilke normaliziramo glede na število težkih atomov v ligandu. S tem namreč postanejo rezultati primerljivi za ligande različnih velikosti. Na sliki 11 je prikazan graf raztrosa RMSD glede na normalizirane vrednosti cenilnih funkcij. Po normalizaciji postane linearna korelacija med RMSD in cenilkama PLP ter CHEMPLP bolj opazna.

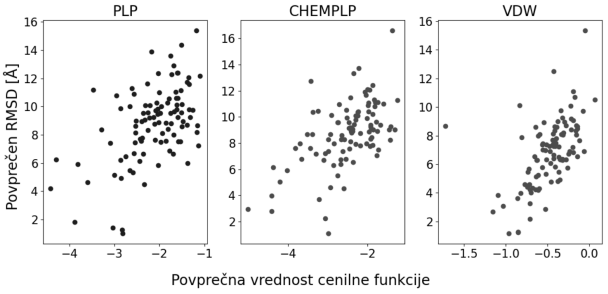


Slika 10:: Razsevni diagram RMSD v odvisnosti od vrednosti cenilne funkcije.

Za oceno linearne korelacije med RMSD in vrednostjo cenilk smo uporabili Pearsonov koeficient korelacije [11], ki ga izračunamo po enačbi (6). Koeficient se giblje med -1 in 1, pri čemer večje absolutne vrednosti kažejo na močnejšo linearno korelacijo med obema spremenljivkama, vrednosti bližje 0 pa kažejo na neobstoječo linearno korelacijo.

$$\begin{aligned}\rho_{\text{XY}} &= \frac{cov(X,Y)}{\sigma_{\text{X}} \cdot \sigma_{\text{Y}}} \\ &= \frac{\sum_{x_{\text{i}} \in X, y_{\text{i}} \in Y} (x_{\text{i}} - \overline{x})(y_{\text{i}} - \overline{y})}{\sqrt{\sum_{x_i \in X} (x_{\text{i}} - \overline{x})^2} \cdot \sqrt{\sum_{y_{\text{i}} \in Y} (y_{\text{i}} - \overline{y})^2}}\end{aligned} \tag{6}$$

Rezultati, prikazani v tabeli 2, potrjujejo, da sta cenilki PLP in CHEMPLP bolj linearno povezani z RMSD po normalizaciji. Koeficienta sta narasla s približno 0,15 na 0,55, medtem ko je cenilka VDW že brez normalizacije dosegala podobno raven linearne odvisnosti.



Slika 11:: Razsevni diagram RMSD v odvisnosti od normalizirane vrednosti cenilne funkcije.

	PLP	CHEMPLP	VDW
brez normalizacije	0,16	0,13	0,58
z normalizacijo	0,54	0,57	0,56

Tabela 2:: Vrednosti Pearsonovega koeficienta korelacije med RMSD in vrednostmi cenilne funkcije.

6 Zaključek

Uspešno smo razvili in implementirali novo cenilno funkcijo v orodje CmDock, izhajajoč iz kode orodja PLANTS. Na kratko smo proučili obstoječo cenilko PLP in njene ključne sestavne dele, pri čemer smo analizirali, kako različni deli cenilke vplivajo na položaje liganda. Cilj naloge je bil izboljšati cenilko PLP z dodajanjem simulacije vodikovih vezi, kar smo dosegli z realizacijo in integracijo treh vrst vodikovih vezi v CmDock.

Kljub pričakovanjem dodana simulacija vodikovih vezi ni izboljšala natančnosti cenilke CHEMPLP v primerjavi s PLP, medtem ko je cenilka VDW ponovno pokazala najboljše rezultate, tako glede RMSD kot tudi števila potrebnih korakov genetskega algoritma. Kljub temu sta PLP in CHEMPLP bistveno hitrejši cenilki, zaradi česar sta privlačni izbiri v primerih visokozmogljivega molekulskega sidranja, kjer je pomembna učinkovitost obdelave velikega števila ligandov na isti tarčni beljakovini [13].

Analiza korelacije med RMSD in vrednostmi cenilk je pokazala, da je cenilka VDW edina, ki kaže opazno linearno korelacijo brez normalizacije. Po normalizaciji glede na število težkih atomov v ligandu pa se je zmerna linearna korelacija pojavila tudi pri cenilkah PLP in CHEMPLP.

V prihodnje bi bilo cenilko CHEMPLP smiselno razširiti z modelom za interakcije med ligandom in kovinskimi atomi, da bi jo lahko uporabili tudi pri beljakovinah s kovinskimi atomi. Prav tako bi lahko s strojnim učenjem optimizirali uteži posameznih členov cenilke CHEMPLP, da bi našli optimalne parametre za orodje CmDock. Še ena ideja za izboljšavo bi bila hibridna uporaba cenilk, in sicer začetni del optimizacije bi opravili s cenilko CHEMPLP, ki je hitrejša in bi se hitro usmerila proti globalnemu ali lokalnemu optimumu. Za zadnji del pa bi uporabili kompleksnejšo cenilko VDW, ki bi našla natančnejšo točko optimuma. Na takšen način bi pohitrili izvajanje cenilke, a hkrati dobili podobne rezultate kot z boljšo cenilko VDW.

Te nadgradnje CmDocka bi prispevale k boljšim rezultatom molekulskega sidranja in povečale vrednost odprtokodne programske opreme CmDock za raziskave na področju odkrivanja novih zdravilnih učinkovin.

Literatura

[1] L.G. Ferreira, R.N. Dos Santos, G. Oliva in A.D. Andricopulo, "Molecular docking and structure-based drug design strategies", *Molecules*, let. 20, št. 7, str. 13384–13421, 2015, doi: 10.3390/molecules200713384.

[2] M. Xuan-Yu, Z. Hong-Xing, M. Mihaly in C. Meng, "Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery", *Current Computer-Aided Drug Design*, let. 7, št. 2, str. 146–157, 2011, doi: 10.2174/157340911795677602.

[3] "CmDock repozitorij kode na GitLab", (15. junij 2024), spletni naslov: https://gitlab.com/Jukic/cmdock/.

[4] "rDock Reference Guide", (10. junij 2024), spletni naslov: https://rdock.sourceforge.net/wp-content/uploads/2015/08/rDock_User_Guide.pdf.

[5] O. Korb, T. Stützle in T.E. Exner, "Empirical scoring functions for advanced protein- ligand docking with PLANTS", *Journal of chemical information and modeling*, let. 49, št. 1, str. 84–96, 2009, doi: 10.1021/ci800298z.

[6] O. Korb, T. Stützle in T.E. Exner, "Accelerating molecular docking calculations using graphics processing units", *Journal of chemical information and modeling*, let. 51, št. 4, str. 865–876, 2011, doi: 10.1021/ci100459b.

[7] M. Kovač, "Podpora za grafične pospeševalnike v orodju za molekulsko sidranje", *Univerza v Ljubljani, Fakulteta za računalništvo in informatiko*, 2023.

[8] T. Erent, "Vzporedni genetski algoritem v OpenCL za simulacijo molekulske dinamike", *Univerza v Ljubljani, Fakulteta za računalništvo in informatiko*, 2022.

[9] K. Szalewicz, "Hydrogen Bond", *Encyclopedia of Physical Science and Technology (Third Edition)*, str. 505–538, 2003, doi: 10.1016/B0-12-227410-5/00322-7.

[10] M.M. Mysinger, M. Carchia, J.J. Irwin in B.K. Shoichet, "Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking", *Journal of Medicinal Chemistry*, let. 55, št. 14, str. 6582–6594, 2012, doi: 10.1021/jm300687e.

[11] J.J. Berman, "Chapter 4 - Understanding Your Data", *Data Simplification*, str. 135–187, 2016, doi: 10.1016/B978-0-12-803781-2.00004-7.

[12] "A complete guide to box plots", (8. september 2024), spletni naslov: https://www.atlassian.com/data/charts/box-plot-complete-guide.

[13] A.V. Sadybekov in V. Katritch, "Computational approaches streamlining drug discovery", *Nature*, let. 616, str. 673–685, 2023, 10.1038/s41586-023-05905-z.

Primož Zidanšek je leta 2024 diplomiral na Fakulteti za računalništvo in informatiko Univerze v Ljubljani.

Črtomir Podlipnik je leta 2003 doktoriral na Univerzi v Ljubljani s področja teoretične kemije. Od leta 2011 je docent na Fakulteti za kemijo in kemijsko tehnologijo Univerze v Ljubljani. Raziskovalno deluje na področjih modeliranja molekul in kemoinformatike.

Davor Sluga je leta 2017 doktoriral na Univerzi v Ljubljani s področja računalništva. Na Fakulteti za računalništvo in informatiko Univerze v Ljubljani je zaposlen kot asistent. Raziskovalno se ukvarja z vzporednim programiranjem, strojnim učenjem in visokozmogljivim računalništvom.

Nejc Ilc je leta 2016 doktoriral na Univerzi v Ljubljani s področja računalništva. Od leta 2021 dalje je docent na Fakulteti za računalništvo in informatiko Univerze v Ljubljani. Na raziskovalnem področju se zanima za strojno učenje, vzporedne računalniške sisteme in bioinformatiko.