



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV INFORMAČNÍCH SYSTÉMŮ

DEPARTMENT OF INFORMATION SYSTEMS

VYHLEDÁVÁNÍ PŘÍBUZNÝCH ENZYMŮ

SEARCH OF RELATED ENZYMES

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

AUTOR PRÁCE

AUTHOR

VEDOUCÍ PRÁCE

SUPERVISOR

SIMEON BORKO

Ing. JIŘÍ HON

BRNO 2020

Zadání bakalářské práce



22863

Student: **Borko Simeon**
Program: Informační technologie
Název: **Vyhledávání příbuzných enzymů**
Search of Related Enzymes

Kategorie: Bioinformatika

Zadání:

1. Seznamte se se základními principy proteinového inženýrství a způsoby získávání enzymů s požadovanými vlastnostmi.
2. Seznamte se s existujícími přístupy a nástroji pro identifikaci příbuzných enzymů.
3. Navrhněte nový nástroj s webovým uživatelským rozhraním pro vyhledávání příbuzných enzymů. Při vyhledávání zohledněte pozice a typy esenciálních residuí. Výsledky doplňte o vhodné anotace z dostupných databází, predikci rozpustnosti a vizualizaci sekvenčního prostoru.
4. Po dohodě s vedoucím práce proveďte implementaci navrženého nástroje a ověřte jeho funkčnost na enzymatické rodině haloalkan dehalogenáz.
5. Zhodnoťte dosažené výsledky a diskutujte možnosti dalšího pokračování projektu.

Literatura:

- VANACEK, Pavel, Eva SEBESTOVA, Petra BABKOVA, et al. Exploration of Enzyme Diversity by Integrating Bioinformatics with Expression Analysis and Biochemical Characterization. ACS Catalysis. 2018, 8(3), 2402-2412

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

Vedoucí práce: **Hon Jiří, Ing.**

Vedoucí ústavu: Kolář Dušan, doc. Dr. Ing.

Datum zadání: 1. listopadu 2019

Datum odevzdání: 14. května 2020

Datum schválení: 22. října 2019

Abstrakt

Milióny nových proteínov, ktoré sa objavujú každý rok, nie je možné charakterizovať klasickými biochemickými metódami pre ich časovú náročnosť a cenu. Medzi nepreskúmanými proteínami sa môžu nachádzať enzýmy uplatniteľné v priemysle i v univerzitnom prostredí, najmä na ekologickú výrobu chemických zlúčenín. Výsledkom práce je webová aplikácia, ktorá na základe vstupných proteínov vyhľadá v databáze podobné proteíny, ktoré prefiltruje pomocou esenciálnych rezíduí vstupných proteínov a označí ich ako domnelé biokatalyzátory. K získaným proteínom sa pridajú anotácie, aby sa mohol používateľ informovane rozhodnúť, ktoré proteíny podrobí experimentálnemu overeniu v laboratóriu. Vytvorený nástroj uľahčuje viackrokovú analýzu a poskytuje návrhy proteínov na experimentálne overenie ich enzymatickej funkcie. Webové rozhranie je voľne dostupné na <https://loschmidt.chemi.muni.cz/enzymeminer/>. Nástroj bol publikovaný v medzinárodnom časopise Nucleic Acids Research.

Abstract

Millions of new proteins discovered each year cannot be characterized by classical biochemical methods due to their demands of time and cost. Among the unexplored proteins, there may be enzymes useful in both industry and academy, mostly for ecological production of chemical compounds. The result of the thesis is a web application which, based on the input proteins, searches the database for similar proteins. The proteins are filtered using essential residues of the input proteins and marked as putative biocatalysts. Finally, the proteins are annotated so that the user can make an informed decision about which proteins to select for experimental laboratory verification. The developed tool facilitates multi-step analysis and recommends proteins for experimental verification of their enzymatic function. The web interface is freely available at <https://loschmidt.chemi.muni.cz/enzymeminer/>. The tool was published in the international journal Nucleic Acids Research.

Kľúčové slová

dolovanie enzýmov, nové biokatalyzátory, vyhľadávanie príbuzných proteínov, predikcia funkcie proteínu, overenie katalytickej funkcie, katalytické reziduá, esenciálne reziduá, bioinformatika, haloalkán dehalogenázy

Keywords

enzyme mining, novel biocatalysts, related protein identification, protein function prediction, catalytic function verification, catalytic residues, essential residues, bioinformatics, haloalkane dehalogenase

Citácia

BORKO, Simeon. *Vyhledávání příbuzných enzymů*. Brno, 2020. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Ing. Jiří Hon

Vyhledávání příbuzných enzymů

Prehlásenie

Prehlasujem, že som túto bakalársku prácu vypracoval samostatne pod vedením pána Ing. Jiřího Hona. Uviedol som všetky literárne pramene, publikácie a ďalšie zdroje, z ktorých som čerpal.

.....
Simeon Borko
27. mája 2020

PodĎakovanie

Ďakujem vedúcemu práce Ing. Jiřímu Honovi za mnohé konzultácie, vedenie a podporu. Taktiež ďakujem Loschmidtovým laboratóriám za odborné vedenie a Mgr. Janovi Štouračovi za technickú podporu. Táto práca vznikla s podporou projektom „e-Infrastruktura CZ“ financovaného v rámci programu Projects of Large Research, Development and Innovations Infrastructures.

Obsah

1	Úvod	2
2	Proteíny a zarovnanie proteínových sekvencií	3
2.1	Podobnosť proteínových sekvencií	3
2.2	Zarovnanie dvoch sekvencií	4
2.3	Zarovnanie viacerých sekvencií	5
2.4	Sieť sekvenčnej podobnosti	7
3	Enzýmy a overenie enzymatickej funkcie	9
3.1	Enzymatická funkcia	9
3.2	Overenie enzymatickej funkcie podobných proteínov	9
4	Proteínové databázy	11
5	Návrh nástroja EnzymeMiner	12
5.1	Návrh výpočtového jadra	12
5.2	Návrh užívateľského rozhrania	14
5.3	Architektúra nástroja	18
6	Implementácia nástroja EnzymeMiner	21
6.1	Implementácia výpočtového jadra	21
6.2	Implementácia úlohového rozhrania	28
6.3	Implementácia webovej aplikácie	28
7	Overenie nástroja na haloalkán dehalogenázach	35
8	Záver	41
	Literatúra	43
A	Pokročilé nastavenia úlohy	46
B	Ukážka vstupných súborov	48
C	Sprievodca elektronickou prílohou	50
D	Článok v časopise Nucleic Acids Research	51

Kapitola 1

Úvod

Proteíny sú prírodné látky tvorené z molekúl zvaných aminokyseliny a sú základom každého živého organizmu. Jednou zo skupín proteínov sú enzýmy, ktoré umožňujú alebo urýchľujú priebeh chemických reakcií – nazývajú sa preto *katalyzátory*. Enzýmy v živých organizmoch riadia väčšinu biochemických procesov. Napríklad, enzým α -amyláza (ptyalín) produkovaný slinnými žľazami štiepi škrob na jednoduchšie sacharidy. Optimálne pH, pri ktorom dosahuje tento enzým najväčšiu účinnosť, je 7,0, aj keď rôzne organizmy môžu vytvárať tento enzým s mierne posunutým optimálnym pH [22]. Enzým α -amyláza má priemyselné využitie pri výrobe sladu a kukuričného sirupu a tiež v čistiaciach a pracích prostriedkoch na rozloženie škrobu na sacharidy rozpustné vo vode. Schopnosť enzýmov katalyzovať chemické reakcie závisí od teploty, pH a tlaku prostredia. Tieto vlastnosti sa môžu líšiť medzi príbuznými enzýmami poskytujúcimi rovnakú katalytickú funkciu. Vzhľadom k rozmanitosti enzýmových rodín je na priemyselné i vedecké účely vhodné hľadať a používať tie najefektívnejšie a najstabilnejšie enzýmy v potrebnom prostredí. Cieľom tejto práce bolo vytvoriť nástroj na užívateľsky prívetivé vyhľadanie príbuzných enzýmov s rovnakou katalytickou funkciou.

Kapitola 2 predstavuje podobnosť proteínov a spôsob ich porovnávanía. Kapitola 3 vysvetľuje, ako fungujú enzýmy a ako je možné výpočtovo overiť enzymatickú funkciu podobných proteínov. Kapitola 4 uvádza proteínové databázy. V kapitole 5 je návrh výpočtového jadra, užívateľského rozhrania a architektúry nástroja. Implementácia tohto návrhu je popísaná v kapitole 6. Overenie nástroja na enzýmovej rodine haloalkán dehalogenáz je v kapitole 7.

Cieľom teoretickej časti nie je podať presné detaily o biochemickej povahe proteínov, ale ponúknuť čitateľovi základné znalosti a princípy, ktoré boli využité pri návrhu a implementácii nástroja. Vytvorený nástroj s názvom EnzymeMiner vznikol v spolupráci s Loschmidtovými laboratóriami, ktoré poskytli odborné vedenie a technickú podporu. Táto práca naväzuje na diplomovú prácu Ing. Jiřího Hona *Vyhledávání příbuzných proteínů s modifikovanou funkcí* [8].



Obr. 1.1: Logo nástroja EnzymeMiner. Logo poskytli Loschmidtove laboratórie.

Kapitola 2

Proteíny a zarovnanie proteínových sekvencií

Proteíny, ktoré sú základom každého živého organizmu, sú prírodné látky tvorené z molekúl zvaných aminokyseliny. Existuje 20 aminokyselín, ktoré sa v bioinformatike reprezentujú ako písmená abecedy. Proteíny sa zapisujú ako reťazec písmen, ktorý sa označuje *sekvencia* proteínu. Lineárne poradie aminokyselín v tomto reťazci sa nazýva *primárnou* štruktúrou proteínu. Vyššie úrovne – *sekundárna* a *terciárna* štruktúra – popisujú priestorové usporiadanie aminokyselín proteínu. V oddiele 2.1 je vysvetlené, že proteíny je možné porovnávať, oddiel 2.2 predstavuje zarovnanie (porovnanie) dvoch proteínových sekvencií a oddiel 2.3 predstavuje zarovnanie viacerých sekvencií. Oddiel 2.4 ponúka spôsob, ako pomocou grafu znázorniť podobnostné vzťahy medzi proteínovými sekvenciami.

2.1 Podobnosť proteínových sekvencií

Každý proteín je členom proteínovej rodiny, ktorá združuje vývojovo príbuzné proteíny s podobnými vlastnosťami. Napríklad, ľudský inzulín a inzulín primáta makaka jávskeho patria do spoločnej rodiny a plnia v príslušných organizmoch rovnakú funkciu – znižujú koncentráciu glukózy (cukru) v krvi. Oba tieto proteíny sú tvorené zo 110 aminokyselín, ale nie sú úplne totožné – líšia sa v dvoch aminokyselinách na pozíciách 23 a 61. *Identita* je pomer počtu zhodných aminokyselín k dĺžke zarovnania, a teda identita týchto dvoch inzulínových sekvencií je 98 %. Ich sekvencie sú dostupné v databáze UniProtKB [23] s identifikátormi P01308 (INS_HUMAN) a P30406 (INS_MACFA).

Vývojovo príbuzné proteíny sa označujú ako *homológne* proteíny a majú spoločného predka. Príbuzné proteíny majú často rovnakú funkciu, a to v prípade, že o ňu neprišli vplyvom vývojových zmien. Sekvenčná podobnosť príbuzných proteínov klesá s množstvom vývojových zmien a rozširovaním vývojových vetiev. Tento jav znižovania podobnosti príbuzných proteínov sa nazýva *divergencia*.

Homológia a divergencia poskytujú akýsi teoretický rámec na uvažovanie o podobnosti proteínov. Avšak to, či sú dva proteíny homológne, nie je možné exaktne určiť na základe samotnej sekvenčnej podobnosti. Na vyvodenie exaktného záveru, že dva proteíny sú homológne, je potrebné presne objaviť ich spoločného predka a všetky medzičlánky [10]. Avšak v praxi je štatisticky významná sekvenčná podobnosť medzi proteínami považovaná za známku divergentného vývoja zo spoločného predka alebo, inými slovami, táto podobnosť je považovaná za dôkaz homológie [10].

SVETADIEL	SVETADIEL
SVETADIEL	LIETAD-EL
SVETADIEL	SVETADIEL
LVETADIEL	LIETAD--L
SVETADIEL	SVETADIEL-
LIETADIEL	LIETAD--LO

Obr. 2.1: Zostrojenie zarovnania slov *svetadiel* a *lietadlo*. Sivá farba označuje v danom kroku ešte nespracované pozície, zvýraznený znak bol z danom kroku zmenený, zelené znaky sú v oboch slovách rovnaké a spojovníky označujú medzery. Počet zmien medzi týmito slovami je päť, čo je ich jednoduchá editačná vzdialenosť.

V mikrobiológii dochádza takisto k javu, kedy sa predkovia z rôznych proteínových rodín vyvinú do proteínov s výraznou sekvenčnou podobnosťou. Tento jav sa nazýva *konvergencia*. V takomto prípade sú podľa definície homológie výsledné vyvinuté proteíny homológne iba k svojim predkom, a nie k sebe navzájom. Konvergenciou zapríčinené podobné proteíny môžu mať (i) medzi sebou rovnakú funkciu, (ii) zdieľať funkciu so svojím predkom alebo (iii) nemať žiadnu funkciu. A nakoniec, aj proteíny s veľmi rozdielnymi sekvenciami môžu mať rovnakú funkciu.

2.2 Zarovnanie dvoch sekvencií

Na určenie miery podobnosti proteínových sekvencií je možné porovnávať ich pomocou textových algoritmov. Uplatnením princípov všeobecnej editačnej vzdialenosti dokážu porovnávať proteínové sekvencie rôzne nástroje.

Editačná vzdialenosť

Na vyjadrenie podobnosti dvoch textových reťazcov sa v informatike používajú chybové modely, ktoré majú svoje využitie pri prenose dát nespoľahlivým kanálom, pri vyhľadávaní v texte s preklepmi alebo pravopisnými chybami a tiež v bioinformatike pri vyhľadávaní DNA sekvencií alebo proteínových sekvencií s možnými mutáciami. Jedným z najviac študovaných chybovým modelom je *editačná vzdialenosť*, ktorá povoľuje tri operácie: (i) vloženie znaku, (ii) odstránenie znaku a (iii) nahradenie znaku za iný znak. Ak majú rôzne operácie rôzne ceny, je reč o *všeobecnej editačnej vzdialenosti*. Inak, ak majú všetky operácie cenu 1, je reč o *jednoduchéj editačnej vzdialenosti*. Jednoduchá editačná vzdialenosť je teda minimálny počet vložení, odstránení a nahradení tak, aby boli oba reťazce rovnaké [13].

Porovnávané reťazce je možné zobrazit pod sebou a vložení *medzier* na správnych pozíciách vznikne *zarovnanie* textových reťazcov. Príklad zostrojenia zarovnania je na obrázku 2.1. Každá úprava potrebná na zmenu reťazca A na reťazec B sa aplikuje nasledujúcim spôsobom:

- Ak je úprava vloženie znaku do reťazca A, vloží sa medzera do reťazca B.
- Ak je úprava odstránenie znaku z reťazca A, vloží sa medzera do reťazca A.
- Ak je úprava zámena znaku, žiadna medzera sa nevloží.

Rozdielna cena editačných operácií

Nie pri všetkých použitíach editačnej vzdialenosti je rovnocenné ohodnocovanie úprav vhodné. V takýchto prípadoch sa používa všeobecná editačná vzdialenosť. Napríklad, lingvistická aplikácia na skúmanie slovotvorných procesov môže požadovať iné ohodnotenie zmeny dĺžky samohlásky než ohodnotenie zámenny samohlásky za spoluhlásku.

Podobne pri ohodnocovaní podobnosti proteínových sekvencií reprezentovaných vo forme textového reťazca je nahradenie rôznych znakov ohodnotených rôzne. Ceny jednotlivých nahradení musia reflektovať biologickú pravdepodobnosť zmien aminokyselín v proteíne a mieru dôsledkov týchto zmien na funkciu proteínu. Na ohodnotenie zarovnania sa používajú substitučné matice, ktoré určujú ohodnotenie všetkých možných dvojíc aminokyselín. Výsledné ohodnotenie zarovnania je súčtom ohodnotení dvojíc aminokyselín v zarovnaní. Jednou zo substitučných matíc je BLOSUM 62 [7], ktorá je zobrazená na obrázku 2.2. V kontexte proteínových sekvencií znamená vyššie ohodnotenie zarovnania to, že sekvencie sú si viac podobné; na rozdiel od jednoduchej editačnej vzdialenosti, ktorá vyjadruje počet zmien. Na obrázku 2.3 je vidieť zarovnanie mliečnych proteínov a na obrázku 2.4 je ukážka ohodnotenia zarovnania pomocou matice BLOSUM 62.

Nástroje na zarovnávanie a vyhľadávanie podobných sekvencií

V tejto práci sú použité tri zarovnávacie nástroje: PSI-BLAST [1], USEARCH [3] a nástroj MMseqs2 [21]. PSI-BLAST sa používa na vyhľadávanie podobných sekvencií v proteínových databázach, napríklad v databáze NCBI nr [17]. Zadaná sekvencia, ktorá sa používa na vyhľadávanie podobných sekvencií, je v tejto práci označovaná ako vstupná sekvencia.

Štatistická významnosť ohodnotenia zarovnania vstupnej sekvencie a nájdenej sekvencie sa uvádza ako hodnota E-value, čo je očakávaný počet náhodne nájdených sekvencií s rovnakým alebo vyšším ohodnotením zarovnania so vstupnou sekvenciou [12]. Zjednodušene je hodnota E-value počet očakávaných výsledkov podobnej kvality, ktoré je možné nájsť náhodou. Čím je hodnota E-value menšia, tým je výsledok lepší. Hodnota E-value môže byť prvým filtrom na vyhľadávanie v databáze – sú žiadané výsledky s hodnotou E-value rovnakou alebo lepšou ako zadanou. Ak sa vyhľadávaniu zadá hodnota E-value $1e-50$, budú výsledky veľmi podobné vstupným sekvenciám, ale bude ich málo. Hodnotu 0.01 je možné stále považovať za vhodnú na nájdenie sekvencií príbuzných proteínov. Hodnota 10 je implicitná hodnota pre BLAST a zahrnie výsledky, ktoré nemôžu byť považované za významné, ale môžu poskytnúť predstavu o možných vzťahoch [18].

2.3 Zarovnanie viacerých sekvencií

Zarovnanie viacerých sekvencií (v angličtine *multiple sequence alignment*, skratka MSA) je zovšeobecnením zarovnania dvoch sekvencií. Zatiaľ čo zarovnanie dvoch sekvencií je významné pri vyhľadávaní v databáze, zarovnanie viacerých sekvencií je zásadné pre analýzu proteínových rodín. MSA spresňuje zarovnanie a pre proteínovú rodinu poukazuje na spoločné a odlišné javy.

Na vytvorenie MSA sú všetky sekvencie najprv navzájom porovnané a zhľukovaním vytvoria vodiaci strom. Sekvencie sa potom postupne zarovnávajú v smere zdola nahor, počnúc koncovými zhľukmi v strome až po vnútorné uzly, až kým sa nedosiahne koreň. Akonáhle sú dve sekvencie zarovnané, ich zarovnanie je zafixované a je ďalej považované za jednu sekvenciu [10]. Vzhľadom k tomu, že zarovnávanie viacerých sekvencií je v podstate

Vtom jdouce, trefíme mezi jakési,
kteříž plnou síň cifer majíce, přebírali se v nich.
A: Někteří berouc z hromady, rozsazovali je;
B: jiní zase přehršlím shrňujíc, na hromádky kladli;
C: jiní opět z těch hromádek díl ubírali a obzvlášť sypali;
D: jiní opět ty díly v jedno snášeli;
E: a jiní zase to dělili a roznášeli,
až sem se tomu jejich dílu podivil.

A: Někteří berouc z
B: jiní zase přehršlím shrňujíc
C: jiní opět z tě-----ch
D: jiní opět ty
E: a jiní zase to

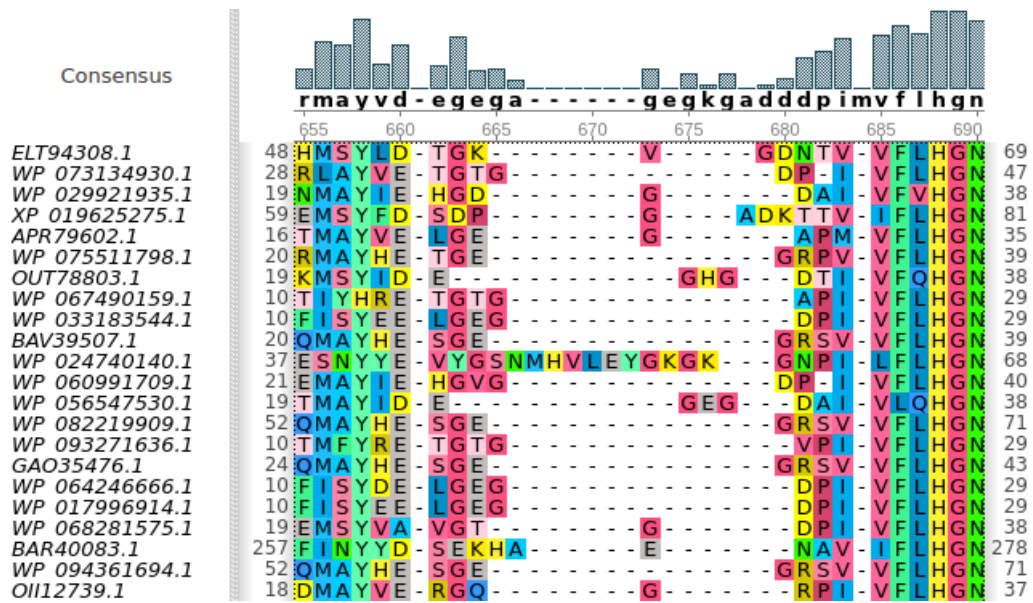
A: hromad-----y, rozsazova-li je
B: na hromád-k-----y kladli
C: hromádek d--íl ubíra-li a obzvlášť sypa--li
D: d--íly v jedno s-nášeli
E: dělili a roz-----nášeli

Obr. 2.5: Ukážka zarovnania viacerých reťazcov. Každý zarovnávaný riadok (A – E) zastupuje jednu proteínovú sekvenciu. Rovnaké znaky sú zvýraznené a zarovnané pod sebou. Pri zarovnaní iba dvoch reťazcov B a D by vzájomná poloha slov *opět* a *zase* bola nejednoznačná. Reťazec C toto zarovnanie spresňuje a dáva slovo *opět* pred slovo *zase*. Text pochádza z časti *Mezi aritmetiky* z knihy *Labyrinth světa a ráj srdce* od Jána Amosa Komenského [9].

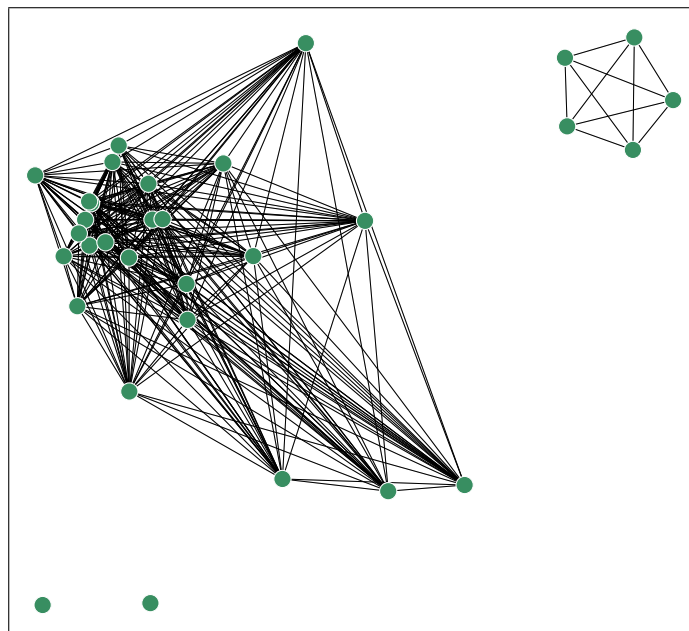
textovým algoritmom, je možné tento princíp ukázať na reťazcoch bez biologického kontextu – ako je tomu na obrázku 2.5. Na vytvorenie zarovnania viacerých proteínových sekvencií je možné použiť nástroj Clustal Omega [20]. Ukážka vytvoreného zarovnania je na obrázku 2.6.

2.4 Sieť sekvenčnej podobnosti

Vzájomnú podobnosť skupiny proteínových sekvencií – sekvenčný priestor – je možné vizualizovať pomocou *siete sekvenčnej podobnosti* (v angličtine *sequence similarity network*, skratka SSN). Uzly v grafe SSN zastupujú sekvencie a dĺžky hrán sú úmerné k sekvenčnej podobnosti spojovaných uzlov. Podobné sekvencie sú blízko seba, zatiaľ čo vzdialenejšie sekvencie sú spojené dlhšou hranou alebo po prekročení určitého prahu nie sú spojené vôbec. Na zlepšenie prehľadnosti môže byť graf SSN pre početnú skupinu sekvencií zjednodušený tým, že uzly budú zastupovať nielen jednu sekvenciu, ale zoskupenie veľmi podobných sekvencií, čím sa zredukuje počet uzlov v grafe. SSN poskytuje jasný vizuálny prístup na identifikovanie zhlukov veľmi podobných sekvencií a na rýchle odhalenie vzdialených sekvencií. Ukážka grafu SSN je na obrázku 2.7.



Obr. 2.6: Zarovnanie viacerých sekvencií zobrazené nástrojom Unipro UGENE [15].



Obr. 2.7: Graf siete sekvenčnej podobnosti. Uzly reprezentujú proteínové sekvencie a dĺžky hrán sú úmerné k sekvenčnej podobnosti spojovaných uzlov.

Kapitola 3

Enzýmy a overenie enzymatickej funkcie

Táto kapitola predstavuje enzýmy a popisuje spôsob, na základe ktorého sa v tejto práci overuje, či príbuzný proteín zdieľa so známym proteínom enzymatickú funkciu.

3.1 Enzymatická funkcia

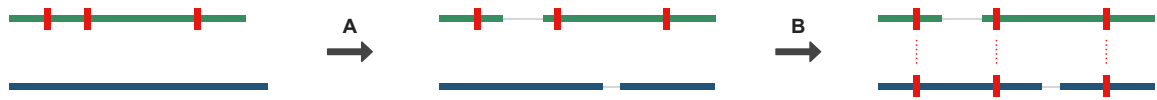
Proteíny s enzymatickou funkciou (enzýmy) katalyzujú chemické reakcie v živých organizmoch. Katalyzovaná chemická reakcia je transformácia *substrátu* na *produkt* za prítomnosti *katalyzátora*. Katalyzátor umožňuje alebo urýchľuje priebeh reakcie, pričom vystupuje z reakcie nezmenený. Vzhľadom k tomu, že enzýmy nie sú jedinou skupinou chemických látok slúžiacich ako katalyzátory, sa na ich označenie používa tiež pojem *biokatalyzátory*. Enzým je možné predstaviť si ako dielňu na výrobu produktu zo zdrojovej suroviny. Zdrojovú surovinu (substrát) je potrebné najprv doviesť do dielne (enzýmu), kde prebehne výroba produktu, ktorý následne dielňu opustí.

Za funkciu každého enzýmu sú zodpovedné konkrétne aminokyseliny nazývané *esenciálne rezíduá*. Esenciálne rezíduá vytvárajú s premieňanou látkou dočasné väzby, čím na ňu spoločne pôsobia. Napríklad, pre enzým s označením Dh1A sú to aminokyseliny na pozíciách 124, 125, 175, 260 a 289.

3.2 Overenie enzymatickej funkcie podobných proteínov

Príbuzné proteíny – proteíny s podobnými sekvenciami – môžu a nemusia mať rovnakú funkciu. Naopak i vzdialené proteíny môžu mať rovnakú funkciu. Neexistuje žiadny prah podobnosti sekvencií, pomocou ktorého je možné bezpečne určiť, či majú dva proteíny rovnakú funkciu [8].

Na zistenie, či dva proteíny majú rovnakú enzymatickú funkciu, je možné overiť prítomnosť esenciálnych rezíduí v skúmanej proteínovej sekvencii. *Templát* je enzým so známymi pozíciami esenciálnych rezíduí. Každé esenciálne rezíduum má pri katalýze svoju špecifickú rolu. Túto rolu môžu byť schopné plniť viaceré aminokyseliny. V zarovnaní skúmanej sekvencie s templátom sa overí, či na pozíciách určenými esenciálnymi rezíduami v templáte sú v skúmanej sekvencii vhodné aminokyseliny (3.1).



Obr. 3.1: Overenie prítomnosti esenciálnych rezíduí. Zelený obdĺžnik označuje templát a modrý obdĺžnik skúmanú sekvenciu. Červené značky reprezentujú esenciálne rezíduá. V kroku A sa templát zarovná so skúmanou sekvenciou. Vzniknuté medzery sú označené vodorovnými čiarami uprostred sekvencií. V kroku B sa prenesú pozície esenciálnych rezíduí z templátu na skúmanú sekvenciu. Ak sú na týchto pozíciách v skúmanej sekvencii vhodné aminokyseliny, je skúmaná sekvencia označená ako domnelý enzým.

V nástroji, ktorý je výstupom tejto práce, sa táto metóda používa so zarovnaniami dvoch sekvencií, rovnako ako aj so zarovnaniami viacerých sekvencií, kde je jeden templát a niekoľko skúmaných sekvencií.

Kapitola 4

Proteínové databázy

Množstvo objavených informácií o proteínoch každoročne rastie a rôzne inštitúcie tieto údaje zhrňujú do databáz. Databázy obsahujú napríklad: (i) proteínové sekvencie, (ii) pozície esenciálnych rezíduí, (iii) funkčné domény proteínov, (iv) trojrozmernú štruktúru proteínov, (v) zaradenie do proteínových rodín, (vi) popis alebo predikciu funkcie proteínov, (vii) odkazy na podobné proteíny a (viii) prepojenia na iné databázy. Proteínové databázy sú verejne dostupné a je možné k nim pristupovať cez webové rozhranie alebo stiahnutím celej databázy na disk.

Databáza *NCBI nr* (non-redundant) [17] združuje proteíny z rôznych databáz a výskumov a usiluje sa o to, aby v nej bola každá proteínová sekvencia iba raz, aby táto databáza neobsahovala identické proteíny viackrát. NCBI nr obsahuje 265 miliónov proteínových sekvencií (vydanie zo 4. 3. 2020). Nástroj BLAST umožňuje túto databázu stiahnuť a vyhľadávať v nej podobné sekvencie. Veľkosť databázy na disku je 206 GiB.

Databáza *UniProtKB* (UniProt Knowledgebase) [23] obsahuje popisy funkcií proteínov, pozície esenciálnych rezíduí enzýmov, zaradenie do proteínovej rodiny, funkčné domény a ďalšie informácie. Databáza UniProtKB je rozdelená na dve časti: (i) *Swiss-Prot*, kde sú anotácie o proteínoch skontrolované odborníkmi, a (ii) *TrEMBL*, kde sú anotácie získané iba výpočtovou charakterizáciou. Vo vydaní 2020_02 obsahuje Swiss-Prot 562 tisíc záznamov, zatiaľ čo TrEMBL obsahuje 180 miliónov záznamov. Tento nepomer poukazuje na dôležitosť výpočtovej analýzy proteínov. Obe časti databázy UniProtKB sú dostupné cez prehľadné webové rozhranie a tiež ako súbory vo formáte XML, ktoré je možné ľahko spracovať a nájsť v nich potrebné informácie.

Databáza *UniParc* (UniProt Archive) je rovnako ako databáza UniProtKB spravovaná konzorciom UniProt. UniParc je archívom všetkých verejne dostupných proteínových sekvencií na svete. Proteíny môžu pochádzať z rôznych zdrojových databáz. UniParc ukladá každú unikátnu sekvenciu iba raz a označuje ju identifikátorom UPI. Identifikátor UPI sa nikdy nezmaže, nezmení a ani nepriradí inej sekvencii. UniParc obsahuje iba samotné sekvencie a odkazy na záznamy zdrojových databáz. Pri týchto odkazoch sa uchováva dátum prvého a posledného zaznamenania danej sekvencie v zdrojovej databáze. Vo vydaní 2020_02 obsahuje UniParc 322 miliónov sekvencií.

Databáza *Pfam* [4] sa zameriava na proteínové rodiny. Proteínové rodiny sú zastúpené zarovnaniami viacerých sekvencií (MSA) a skrytými Markovovými modelmi (HMM). Proteíny sú zložené z jedného alebo viacerých funkčných regiónov, ktoré sa nazývajú domény. Identifikovaním domén v proteínových sekvenciách je možné nahliadnuť na funkciu proteínu.

Kapitola 5

Návrh nástroja EnzymeMiner

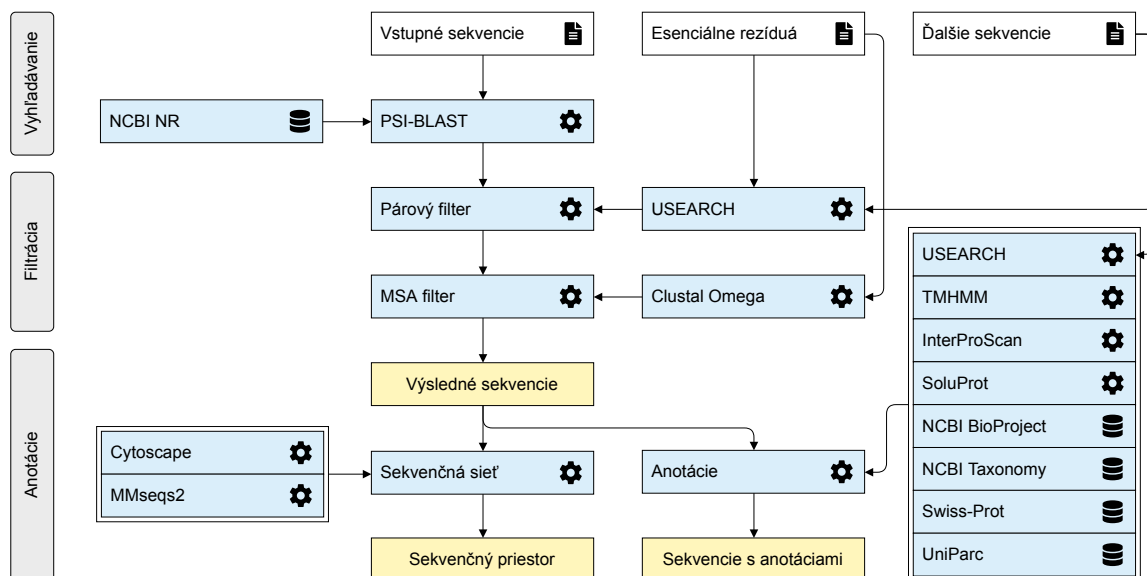
Nástroj na vyhľadávanie príbuzných enzýmov je potrebné dôsledne navrhnuť. V oddiele 5.1 sa nachádza návrh výpočtového jadra, ktoré zabezpečí hlavnú výpočtovú časť nástroja. Oddiel 5.2 predstavuje návrh užívateľského rozhrania, pomocou ktorého môže užívateľ (i) zadať nástroju úlohy na výpočet a (ii) vyzdvihnúť si výsledky zadaných úloh. Posledný oddiel 5.3 popisuje navrhnutú architektúru pre nástroj EnzymeMiner.

5.1 Návrh výpočtového jadra

Návrh výpočtu vychádza zo skúseností Loschmidtových laboratórií s hľadaním enzýmovej rodiny haloalkán dehalogenáz [24]. *Pracovný postup* nástroja má tri kroky: (i) vyhľadávanie homológov, (ii) filtrácia založená na esenciálnych rezíduách, (iii) anotácia výsledných sekvencií (5.1). Na vykonanie týchto úloh požaduje výpočtové jadro dva rôzne typy informácií: (i) vstupné sekvencie, (ii) templáty esenciálnych rezíduí. Vstupné sekvencie slúžia ako sekvencie na vyhľadávanie homológnych (príbuzných) sekvencií. Templáty esenciálnych rezíduí, definované ako dvojice proteínovej sekvencie a sady esenciálnych rezíduí v tejto sekvencii, umožnia nástroju vybrať spomedzi homológov sekvencie s najväčšou pravdepodobnosťou enzymatickej funkcie. Každé esenciálne rezíduum je definované názvom, pozíciou v proteínovej sekvencii a množinou povolených aminokyselín pre túto pozíciu.

V prvom kroku vyhľadávania homológov budú vstupné sekvencie použité ako vstup pre nástroj PSI-BLAST [1] na vyhľadávanie v proteínovej databáze NCBI nr [17]. Ak bude zadaných viacero sekvencií, vyhľadávanie sa spustí pre každú sekvenciu zvlášť. Umelé proteínové sekvencie, čiže sekvencie s označením artificial, synthetic construct, vector, vaccinia virus, plasmid, halotag alebo replicon, sa odstránia, pretože nepochádzajú z prírody a nemá zmysel zaoberať sa nimi v tomto nástroji, ktorý medzi proteínami predpokladá prirodzený vývoj. EnzymeMiner zoradí výsledky nástroja PSI-BLAST podľa hodnoty E-value a do ďalších krokov posunie maximálne 10 000 najlepších výsledkov. Dôvod tohto obmedzenia je zrejmý – príliš veľké množstvo sekvencií by spomalilo výpočet a vyžadovalo by príliš veľa diskového priestoru. Konkrétne použitie nástroja však môže požadovať vyšší limit počtu najlepších výsledkov, aby sa mohli odhaliť aj vzdialenejšie výsledky, a preto bude možné tento parameter nastaviť podľa potreby.

V druhom kroku filtrácie založenej na esenciálnych rezíduách sa výsledky vyhľadávania homológov prefiltrujú použitím templátov esenciálnych rezíduí. Po prvé, výsledné sekvencie z predošlého kroku sa rozdelia do templátových zhlukov. Každý zhluk bude obsahovať všetky sekvencie, ktoré sa s daným templátom zhodujú v esenciálnych rezíduách. Esenciálne rezíduá



Obr. 5.1: Pracovný postup nástroja EnzymeMiner. Základné kroky postupu: (i) vyhľadavanie homológov, (ii) filtrácia založená na esenciálnych rezíduách, (iii) anotácia výsledných sekvencií.

sa skontrolujú pomocou párového zarovnania s templátom nástrojom USEARCH [3]. Ak sa sekvencia zhoduje s viacerými templátmi esenciálnych rezíduí, bude priradená k templátu s najvyššou sekvenčnou identitou. Po druhé, po rozdelení sekvencií do zhlukov sa pre každý zhluk skonštruuje zarovnanie viacerých sekvencií (MSA) nástrojom Clustal Omega [20]. MSA dodatočne overí esenciálne rezíduá vo výsledkoch, a to skontrolovaním príslušného stĺpca v MSA. Sekvencie, ktoré sa nebudú zhodovať v esenciálnych rezíduách s templátom, budú zo zhluku odstránené. Po tretie, pre každý zhluk sa MSA skonštruuje znova a naposledy sa skontrolujú esenciálne rezíduá. Táto druhá iterácia filtrovania pomocou MSA spresní zarovnanie tým, že už v ňom nebudú divergentné sekvencie, ktoré mali vplyv na zarovnanie v prvej iterácii.

V poslednom anotačnom kroku sa výsledné proteínové sekvencie oannotujú použitím rôznych databáz a prediktorov: (i) transmembránové regióny sa predikujú nástrojom TMHMM [11], (ii) Pfam domény sa predikujú nástrojom InterProScan [16], (iii) anotácie o zdrojových organizmoch výsledných proteínov sa získajú z databázy NCBI Taxonomy [5] a z databázy NCBI BioProject [2], (iv) rozpustnosť proteínu sa predikuje nástrojom SoluProt (zatiaľ nepublikovaný nástroj Loschmidtových laboratórií), (v) sekvenčná identita so vstupnými sekvenciami, s ostatnými výslednými sekvenciami alebo s ďalšími voliteľnými vstupnými sekvenciami sa vypočíta nástrojom USEARCH [3], (vi) dátum prvého zaznamenania (*first seen*) sa zistí z databázy UniParc a (vii) v prípade, že má sekvencia revidovaný záznam v databáze Swiss-Prot, tak sa k výslednej sekvencii priradí označenie tohto záznamu.

Sekvenčný priestor výsledných sekvencií sa vizualizuje použitím sietí podobnosti reprezentatívnych sekvencií (SSN, oddiel 2.4). Siete sa vygenerujú na rôznych úrovniach detailu. Varianty s menšími detailmi budú mať uzly znázorňujúce viacero sekvencií a uzly variantov s väčšími detailmi znázornia menej sekvencií. Na zobrazenie siete na webe je potrebné, aby sieť nebola príliš veľká. Preto sa siete vygenerujú na rozličných prahoch zhľukovania použitím nástrojov MMseqs2 [21] a Cytoscape [19]. SSN poskytne jasný vizuálny prístup

Job Input

Swiss-Prot sequences
Custom sequences

EC number:

Load example

Select sequences from table

	Accession	ER	Length	Sequence plot
<input checked="" type="checkbox"/>	Q9ZER0	3	307	
<input checked="" type="checkbox"/>	P59336	5	290	
<input type="checkbox"/>	P9WMR8	3	280	
<input checked="" type="checkbox"/>	A5U5S9	3	240	

Essential residues: Nucleophile (106), Proton donor (130), Proton acceptor (230)

Select sequences from similarity network

Filter sequences by Pfam domains...

Select all
Deselect all

Show selected only

Select representative sequences of clusters

10%

20%

30%

40%

50%

60%

70%

80%

90%

100%

Page 1 of 10: Prev 1 2 3 4 5 Next

Obr. 5.2: Návrh grafického užívateľského rozhrania na zadanie vstupu použitím revidovaných sekvencií z databázy Swiss-Prot. Vstupné sekvencie môžu byť vybrané použitím: (i) tabuľky sekvencií, (ii) siete sekvenčnej podobnosti, (iii) prahu sekvenčnej identity.

na identifikovanie zhľukov veľmi podobných sekvencií a na rýchle odhalenie vzdialených sekvencií.

5.2 Návrh užívateľského rozhrania

Nové úlohy môžu byť zadané z webového užívateľského rozhrania. Nástroj EnzymeMiner poskytne dva konceptuálne odlišné spôsoby, ako definovať vstup pre *pracovný postup*: (i) použitím revidovaných sekvencií z databázy UniProtKB/Swiss-Prot [23] a (ii) použitím vlastných sekvencií. Možnosť vybrať revidované sekvencie je vhodná pre užívateľov, ktorí nemajú hĺbkovú znalosť zadávanej enzymatickej rodiny. Možnosť zadať vlastné sekvencie poskytuje plnú kontrolu nad vstupom pre nástroj EnzymeMiner a je vhodná pre užívateľov, ktorí dobre poznajú zadávanú enzymatickú rodinu a chcú nástroju zadať pokročilé začiatkové informácie na získanie lepších výsledkov. Posledná možnosť je kombinácia oboch prístupov, kde sú najprv vybrané revidované sekvencie a následne ich užívateľ upraví. Užívateľ môže tiež využiť možnosť načítať si predpripravený vzorový vstup.

Revidované sekvencie z databázy Swiss-Prot

V záložke *Swiss-Prot sequences* (5.2) môžu byť sekvencie získané na základe Enzyme Commission (EC) čísla z databázy Swiss-Prot. Po zadaní tohto čísla sa zobrazí tabuľka všetkých sekvencií danej EC skupiny a príslušná sieť sekvenčnej podobnosti. Tabuľka má štyri stĺpce: (i) názov sekvencie (accession) odkazujúci do databázy UniProt, (ii) počet

esenciálnych rezíduí, (iii) dĺžka sekvencie a (iv) sekvenčný graf. Sekvenčný graf ukazuje dve dôležité vlastnosti sekvencie – pozície esenciálnych rezíduí a identifikované Pfam domény. Esenciálne rezíduá sú konkrétne jednotlivé aminokyseliny a Pfam domény sú súvislé funkčné regióny tvorené viacerými aminokyselinami. Na zobrazenie iba riadkov so sekvenciami, ktorú majú určitú Pfam doménu, môže užívateľ využiť filtrovacie pole so zoznamom domén v EC skupine. Prepínač *Show selected only* zobrazí v tabuľke iba užívateľom vybrané sekvencie bez ohľadu na vybrané domény vo filtrovacom poli. Graf siete sekvenčnej podobnosti znázorňuje sekvenčný priestor všetkých sekvencií v zadanej EC skupine. Uzly zastupujú Swiss-Prot sekvencie. Dĺžky hran sú úmerné k sekvenčnej identite spojovaných uzlov. Podobné sekvencie sú blízko seba, zatiaľ čo vzdialenejšie sekvencie nie sú spojené vôbec. Aktuálne vybrané uzly sú farebne zvýraznené.

Vybrať Swiss-Prot sekvenciu ako vstupnú sekvenciu pre nástroj EnzymeMiner je možné tromi spôsobmi: (i) výber riadku v tabuľke sekvencií, (ii) výber uzlu v sieti sekvenčnej podobnosti alebo (iii) výber zástupcov zhlukov určením prahu sekvenčnej identity. Tlačidlá prahu sekvenčnej identity vyberú zástupcov zhlukov vytvorených zhlukovaním s daným percentuálnym prahom sekvenčnej identity. Pomocou tejto funkcie môže užívateľ automaticky vybrať malé množstvo sekvencií, ktoré pokrývajú celý sekvenčný priestor zadanej EC skupiny. Všetky vybrané Swiss-Prot sekvencie sú použité ako vstup na vyhľadanie homológov a sú tiež použité ako templáty esenciálnych rezíduí pre filtračný krok navrhnutého *pracovného postupu*. Na úpravu vybraných sekvencií a templátov esenciálnych rezíduí môže užívateľ prepnúť užívateľské rozhranie do záložky *Custom sequences* a výber si ručne prispôbiť.

Vlastné sekvencie

V záložke *Custom sequences* (5.3) môže používateľ zadať vlastné vstupné sekvencie a templáty esenciálnych rezíduí. Na tejto stránke môže používateľ nahrať celé zadanie úlohy zo súboru. Sekvencie je možné zadať vo formáte FASTA vyplnením vstupných polí alebo načítaním zo súboru so sekvenciami. Sekvencie vo formáte FASTA obsahujú: (i) hlavičku s názvom začínajúcu znakom *väčší než* (>) a (ii) textový reťazec sekvencie. Templáty esenciálnych rezíduí sa vyplňajú úpravou tabuľky.

Sekvencie zadané do pola *Query sequences* budú použité ako vstup na vyhľadávanie homológov (príbuzných sekvencií). Nepovinné sekvencie zadané do pola *Other known sequences* budú použité na vypočítanie identity s výslednými sekvenciami.

Templáty esenciálnych rezíduí sú zásadnými zdrojmi informácií potrebných vo filtračnom kroku *pracovného postupu* nástroja. Všetky výsledné sekvencie nástroja sa musia zhodovať v esenciálnych rezíduách s aspoň jedným templátom. Upravovateľná tabuľka templátov má pre každý templát jeden riadok a pre každé esenciálne rezíduum jeden stĺpec. Názvy sekvencie (accessions) musia tvoriť podmnožinu názvov sekvencií zadaných vo vstupných poliach sekvencií. Nie každá vstupná sekvencia musí byť templátom, ale každý templát musí mať sekvenciu v poli *Query sequences* alebo v poli *Other known sequences*. Okrem prvého stĺpca (*Accession*) sú názvy stĺpcov užívateľom definované názvy esenciálnych rezíduí označujúce ich roly pri katalýze. Pod názvom esenciálneho rezídua sú čiarkou oddelené jednoznakové značky povolených aminokyselín, ktoré sú schopné plniť rolu esenciálneho rezídua. Tieto aminokyseliny užívateľ vyberie pomocou modálneho okna zobrazeného po kliknutí na tlačidlo *Set residues*. Riadky i stĺpce je možné pridávať a odstraňovať podľa potreby. Vnútrnú časť tabuľky predstavujú číselné pozície rezídua určeného stĺpcom v sekvencii určenej riadkom. V prípade, že v sekvencii nebude na zadanej pozícii jedna z povolených aminokyselín, bude

Job Input

Swiss-Prot sequences
Custom sequences

Upload input file
Load example

Query sequences Upload FASTA

```
>Q6Q3H0
MINAIRTDPQRFNSNLDQYPFSPNYLDDLPGYPLRAHYLD
PDFFGFGKSDKPVDEEDYTFFHRNFLALIERLDRNIT
PAFSAFVTQPADGFTAWKYDLVTPSDLRLDQFMKRWAPTL
TEAISFWQNDWNGQTFMAIGMKDKLLGPDVMPMKALING
```

Other known sequences

Sequences in FASTA format...

Allowed residues for Nucleophile

<input type="checkbox"/> A	<input type="checkbox"/> C	<input type="checkbox"/> H	<input type="checkbox"/> M	<input type="checkbox"/> W
<input type="checkbox"/> R	<input type="checkbox"/> Q	<input type="checkbox"/> I	<input type="checkbox"/> F	<input type="checkbox"/> Y
<input type="checkbox"/> N	<input type="checkbox"/> E	<input type="checkbox"/> L	<input type="checkbox"/> P	<input type="checkbox"/> V
<input type="checkbox"/> D	<input type="checkbox"/> G	<input type="checkbox"/> K	<input type="checkbox"/> S	<input type="checkbox"/> T

CANCEL OK

Essential residue templates
✔ Add sequence (row)
➤ Add residue (column)

▼ Accession	▼ Halide 1 ✘	▼ Enter name ✘	▼ Nucleophile ✘	▼ Proton donor ✘	▼
	W, N	W, N	<i>Set residues</i> ●	D, E	
Q6Q3H0	125	175	124	260	✘
Q73Y99	<i>Enter position</i>	<i>Enter position</i>	123	250	✘
<i>Enter accession</i>	<i>Enter position</i>	<i>Enter position</i>	114	138	✘

Obr. 5.3: Návrh grafického užívateľského rozhrania na zadanie vlastných vstupných sekvencií a templátov esenciálnych rezíduí.

Job Results

Job output information

Job ID: example

Job Title: Example

Status: Done

Download results

Result table (XLSX)

Result table (TSV)

Raw results

Target selection table

Select all Deselect all Undo Redo

Minimum solubility Query identity range

Selected	Full Dataset	Extra Domain	Organism	Temperature	Salinity	Biotic Rel.	Disease	Transmembrane
Accession	Closest query	Kingdom	Solubility	Sequence length	Domain annotation	Organism	Essential residues	
<input checked="" type="checkbox"/>	WP_071575177.1	D4Z2G1_S1	B	0.9399	270	Abhydrolase_1	Corynebacterium diphtheriae	NDWEH
<input type="checkbox"/>	3SK0_A	D4Z2G1_S1	B	0.9393	311	Abhydrolase_1	Rhodococcus rhodochrous	NDWEH
<input checked="" type="checkbox"/>	AGS48406.1	D4Z2G1_S1	B	0.9357	290	Abhydrolase_1	Rhodobacteraceae	NDWEH
<input type="checkbox"/>	WP_003902310.1	D4Z2G1_S1	B	0.925	292	Abhydrolase_1	Renilla reniformis	NDWEH

Obr. 5.4: Návrh grafického užívateľského rozhrania na zobrazenie výsledkov úlohy.

užívateľ vyzvaný na úpravu sekvencie, množiny povolených aminokyselín alebo zadanej pozície.

Prechod z revidovaných sekvencií na vlastné sekvencie

Pri prechode zo záložky *Swiss-Prot sequences* do záložky *Custom sequences*, ak bolo zadané EC číslo a vybraná aspoň jedna sekvencia, sa vybrané sekvencie automaticky prenesú do poľa *Query sequences*, pole *Other known sequences* zostane prázdne a tabuľka templátov esenciálnych rezíduí sa vyplní podľa vybraných sekvencií. Pre každú sekvenciu sa vytvorí jeden riadok tabuľky. Všetky esenciálne rezíduá vybraných sekvencií sa združia podľa názvu rezídua a vytvoria stĺpce tabuľky. Povolené aminokyseliny sa automaticky nastavia ako zjednotenie aminokyselín nájdených na daných pozíciách v sekvenciách s týmto esenciálnym rezíduom.

Zhrnutie zadania úlohy

Po zadaní úlohy pomocou revidovaných alebo vlastných sekvencií sa užívateľovi zobrazí zhrnutie zadania úlohy, kde si môže zadanie stiahnuť a uložiť na opätovné použitie v budúcnosti. Po kontrole zadania úlohy užívateľ spustí úlohu, čím sa dostane na výsledkovú stránku nástroja.

Výsledková stránka

Výsledková stránka (5.4) bude rozvrhnutá do štyroch sekcií: (i) informácie o úlohe, (ii) tlačidlá na stiahnutie, (iii) výsledková tabuľka a (iv) sieť sekvenčnej podobnosti.

Pri informáciách o úlohe je zobrazený identifikátor úlohy, názov úlohy a aktuálny stav úlohy. Pomocou tlačidiel na stiahnutie je možné stiahnuť výsledkovú tabuľku vo formáte XLSX alebo vo formáte tabuľky oddelený text a tiež archív so všetkými výstupnými súborami z *pracovného postupu* výpočtového jadra.

Výsledková tabuľka (*Target selection table*) je najdôležitejším komponentom výsledkov nástroja EnzymeMiner. Tabuľka zobrazuje všetky domnelé enzýmové sekvencie identifikované nástrojom EnzymeMiner a pomáha vybrať sekvencie na experimentálnu charakterizáciu. Každý z deviatich hárkov tabuľky zobrazuje výsledky z rôznych pohľadov: (i) Hárak *Selected* zobrazuje všetky sekvencie vybrané z jednotlivých hárkov. Hárak *Selected* obsahuje oproti ostatným hárkom jeden stĺpec navyše, ktorý zaznamenáva dôvod vybraní. Implicitne je tento dôvod nastavený na názov hárku, v ktorom bola sekvencia vybraná, ale môže byť voľne zmenený. (ii) Hárak *Full Dataset* zobrazuje všetky identifikované sekvencie. (iii) Hárak *Extra Domain* zobrazuje sekvencie s dodatočnými Pfam doménami, ktoré boli nájdené v sekvencii, ale nie sú zadané v poli *Primary domains*. (iv) Hárak *Organism* zobrazuje sekvencie so známym zdrojovým organizmom. (v) Hárak *Temperature* zobrazuje sekvencie z organizmov s anotáciou optimálnej teploty v databáze NCBI BioProject. (vi) Hárak *Salinity* zobrazuje sekvencie z organizmov s anotáciou salinity v databáze NCBI BioProject. (vii) Hárak *Biotic Relationship* zobrazuje sekvencie z organizmov s anotáciou biotického vzťahu v databáze NCBI BioProject. (viii) Hárak *Disease* zobrazuje sekvencie z organizmov s anotáciou ochorenia v databáze NCBI BioProject. (ix) Hárak *Transmembrane* zobrazuje sekvencie s transmembránovými regiónmi predikovanými nástrojom TMHMM. Na zobrazenie výsledných sekvencií je možné použiť dvojaký filter: (i) nastavenie minimálnej solubility proteínu a (ii) nastavenie rozsahu sekvenčnej identity ku vstupnej sekvencii.

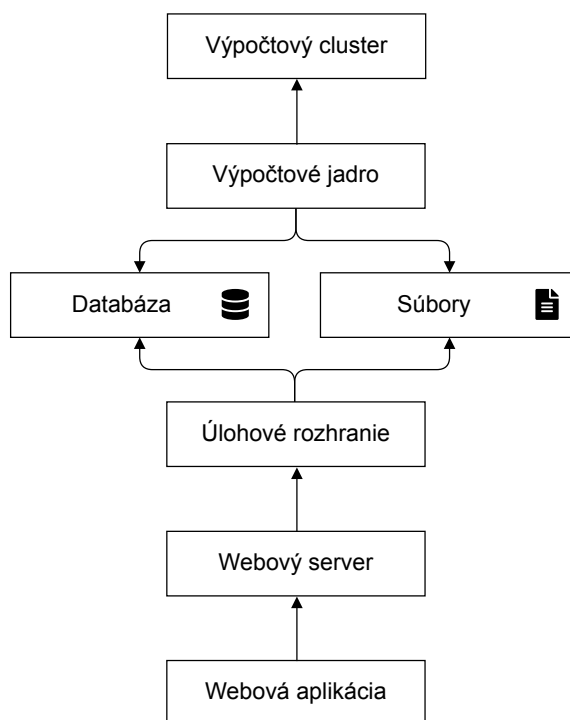
Sieť sekvenčnej podobnosti znázorňuje sekvenčný priestor identifikovaných sekvencií. Zhľuky podobných sekvencií, rovnako ako vzdialené sekvencie, môžu byť jednoducho identifikované. Vzhľadom k tomu, že výsledných sekvencií môžu byť tisíce, každý uzol môže reprezentovať viacero blízkych sekvencií, čo prispieva k prehľadnosti a znižuje požiadavky na výkon pri zobrazovaní.

5.3 Architektúra nástroja

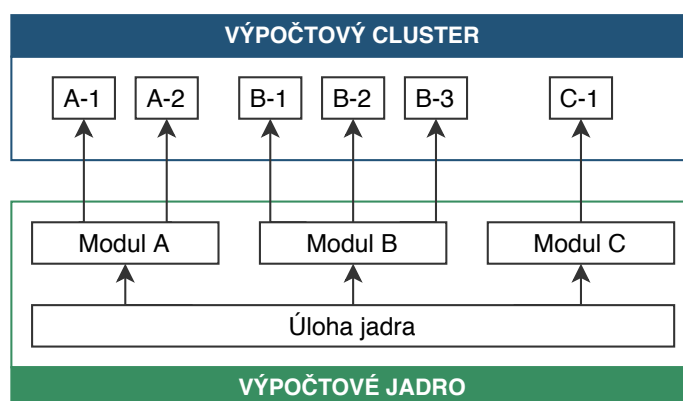
Mnoho krokov v rámci *pracovného postupu* je výpočtovo náročných, a preto vyžaduje nástroj architektúru šitú na mieru. Nástroj EnzymeMiner má viacvrstvovú architektúru, kde každá vrstva má špecifický účel a zodpovednosť (5.5).

Výpočtové jadro realizuje hlavný *pracovný postup* nástroja (5.1). Jadro prijíma úlohy z databázy a získava príslušné súbory z disku. Pre každú úlohu jadro vytvorí a naplánuje niekoľko výpočtových modulov, ktoré implementujú jednotlivé kroky *pracovného postupu*, pričom nezávislé moduly sú spustené súbežne. Na výpočet neobyčajne náročných procedúr využívajú moduly jadra *Výpočtový cluster*, ktorý použitím svojich zdrojov spracuje zadanú úlohu. To znamená, že je potrebné rozlišovať tri rôzne výpočtové jednotky (5.6): (i) užívateľom zadaná úloha pre výpočtové jadro, (ii) moduly jadra spravované výpočtovým jadrom a (iii) úlohy spravované výpočtovým clustrom.

Úlohové rozhranie zapuzdruje prácu s databázou a prácu so súborami do jednej vrstvy. Toto rozhranie zabezpečuje konzistenciu medzi databázou a súborami a umožňuje (i) pridať



Obr. 5.5: Viacvrstvá architektúra navrhnutá pre nástroj EnzymeMiner.



Obr. 5.6: Výpočtové jednotky nástroja EnzymeMiner. Tri typy výpočtových jednotiek: (i) úlohy jadra, (ii) moduly jadra a (iii) úlohy výpočtového clustra.

nové úlohy do databázy a uložiť vstupné súbory, (ii) stiahnuť stav úlohy a výsledky úlohy a (iii) získať štatistiku užívateľského využívania nástroja.

Webový server nástroja je štandardný webový server obsluhujúci užívateľské požiadavky klientov. Na túto obsluhu využíva tiež služby *Úlohového rozhrania*. S *Webovým serverom* úzko spojená *Webová aplikácia* vytvára grafické užívateľské rozhranie navrhnuté v oddiele 5.2.

Kapitola 6

Implementácia nástroja EnzymeMiner

Nástroj EnzymeMiner má netriviálnu architektúru viacerých vrstiev, kde každá vrstva má svoj účel a zodpovednosť (oddiel 5.3). Výpočtové jadro, databáza, úlohové rozhranie a webový server sú spustené a prepojené systémom Docker.

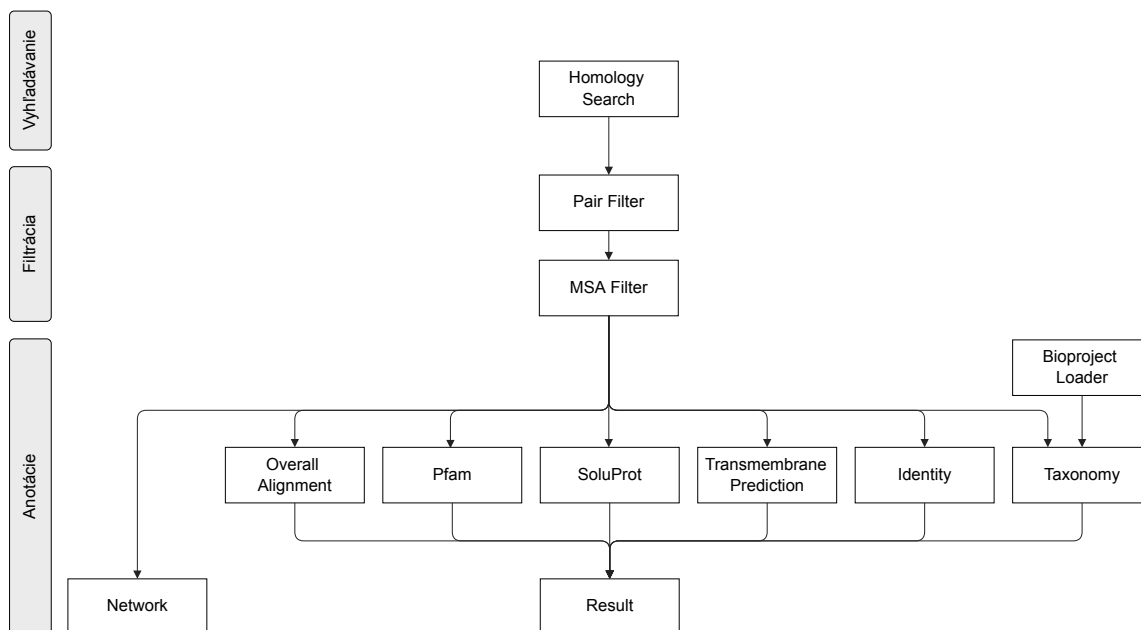
Výpočtový cluster používa na plánovanie úloh Portable Batch System (PBS) a je spravovaný Loschmidtovými laboratóriami. *Výpočtové jadro* (oddiel 6.1), implementované v jazyku Java, zabezpečuje hlavný výpočet nástroja. *Databáza MySQL* uchováva informácie o úlohách pre výpočtové jadro. *Úlohové rozhranie* (oddiel 6.2) je implementované v jazyku Java ako HTTP aplikačné rozhranie pomocou frameworku Spring Boot. *Webový server* spustený cez Node.js, ako jediný verejne prístupný server nástroja z Internetu, podáva súbory potrebné na zobrazenie webovej aplikácie a slúži ako brána k úlohovému rozhraniu. *Webová aplikácia* (oddiel 6.3) je implementovaná ako jednostránková aplikácia v systéme React s využitím správcu stavu Redux. Zdrojové súbory implementácie a skripty pre výpočtový cluster sú priložené elektronicky a ich prehľad je uvedený v prílohe C.

6.1 Implementácia výpočtového jadra

Výpočtové jadro realizuje hlavný *pracovný postup* nástroja (oddiel 5.1), komunikuje s databázou a súborovým systémom a využíva služby výpočtového clustra. Jadro je implementované v jazyku Java s použitím frameworku Loschmidtových laboratórií. Zoznam parametrov, ktoré je možné nastaviť pre výpočet úlohy, je uvedený v prílohe A.

Framework Loschmidtových laboratórií

Implementácia výpočtového jadra používa framework Loschmidtových laboratórií (ďalej ako framework), v spolupráci s ktorými tento projekt vznikol. Tento framework štruktúruje výpočetný problém (úlohu) do modulov a zabezpečuje základnú spoločnú funkcionálnu pre bioinformatické nástroje, a to napríklad: (i) plánovanie úloh, (ii) riadenie závislosti medzi modulmi, (iii) komunikácia s databázou, (iv) komunikácia s výpočtovým clusterom, (v) vytvorenie diskového priestoru pre úlohu, (vi) symbolické názvy pre súbory zdieľané medzi modulmi a (vii) e-mailové notifikácie o zmene stavu úlohy. Programátor s použitím frameworku potrebuje už iba (i) navrhnúť a implementovať moduly, (ii) určiť vstupné a výstupné súbory modulov a (iii) určiť závislosti medzi modulmi.



Obr. 6.1: Graf závislostí modulov jadra. Modul sa nemôže začať vykonávať, kým nie sú úspešne dokončené všetky moduly zobrazené nad daným modulom v grafe závislostí.

Vo výsledku jadro funguje nasledovným spôsobom: (i) jadro nájde v databáze novú úlohu, (ii) prečíta si špecifikáciu úlohy (vstupné súbory a parametre), (iii) vytvorí pracovnú adresárovú štruktúru na disku pre túto novú úlohu, (iv) skopíruje vstupné súbory do vytvorenej adresarovej štruktúry, (v) naplánuje spustenie výpočtových modulov a (vi) po dokončení všetkých modulov uloží do databázy informáciu, že úloha je dokončená, alebo, v prípade zlyhania nejakého modulu, že výpočet úlohy zlyhal.

Moduly a zdieľané súbory

Výpočet jadra je rozdelený do dvanástich modulov (obrázok 6.1), ktoré implementujú tri kroky *pracovného postupu*: (i) vyhľadávanie homológov, (ii) filtrácia a (iii) anotácie výsledkov. Každý z modulov má na disku svoj priestor, ale všetky moduly môžu takisto pristupovať k zdieľanej zložke, odkiaľ získajú vstupné súbory a kam uložia svoje výstupné súbory. Zoznam všetkých zdieľaných súborov je v tabuľke 6.1. Jeden symbolický názov súboru môže zastupovať skupinu súborov, a to tým, že skutočná cesta k súboru je parametrizovaná. Napríklad, skutočná cesta pre `pair_filter_cluster` je `pair_filter/cluster_{$number}.fa`; pri získavaní súboru je nutné zadať hodnotu parametra `number`.

Modul Homology Search

Modul *Homology Search* vyhľadáva v databáze NCBI nr [17] podobné proteínové sekvencie ako vstupné sekvencie, a to pomocou nástroja PSI-BLAST [1].

Predspracovanie vstupných súborov zahŕňa odstránenie anotácií z hlavičiek vstupných sekvencií, pridanie prípony k identifikátorom vstupných sekvencií a prevod sekvencií na veľké písmená. Každá sekvencia zo súboru `queries` sa uloží do samostatného súboru. Pre každý samostatný súbor s jednou sekvenciou sa spustí nástroj PSI-BLAST, ktorého výstupný súbor obsahuje zoznam identifikátorov nájdených sekvencií. Po dokončení všetkých samostatných

Symbolický názov	Popis
Vstupné súbory	
<code>user_queries</code>	vstupné sekvencie na vyhľadávanie homológov (FASTA)
<code>user_known_sequences</code>	všetky vstupné (známe) sekvencie (FASTA)
<code>user_catalytic_aas</code>	tabuľka templátov esenciálnych rezíduí
Predspracované vstupné súbory	
<code>queries</code>	predspracované vstupné sekvencie na vyhľadávanie homológov (FASTA)
<code>known_sequences</code>	predspracované všetky vstupné sekvencie (FASTA)
<code>input_annotations</code>	anotácie z hlavičiek vstupných sekvencií
<code>catalytic_aas</code>	tabuľka templátov esenciálnych rezíduí
Vyhľadávanie homológnych sekvencií	
<code>seqs_blast</code>	nájdene podobné sekvencie (FASTA)
<code>seqs_blast_pdb_accessions</code>	identifikátory PDB štruktúr nájdených sekvencií
<code>seqs_blast_alternative_accessions</code>	alternatívne identifikátory nájdených sekvencií
<code>seqs_blast_annotations</code>	anotácie z hlavičiek nájdených sekvencií
Párové filtrovanie	
<code>templates</code>	sekvencie templátov (FASTA)
<code>pair_filter_cluster_count</code>	počet zhlukov z párového filtrovania
<code>pair_filter_cluster *</code>	zhluky z párového filtrovania
MSA filtrovanie	
<code>msa_filter_a_cluster_count</code>	počet zhlukov z prvej iterácie MSA filtrovania
<code>msa_filter_a_cluster *</code>	zhluky z prvej iterácie MSA filtrovania
<code>msa_filter_a_positions *</code>	pozície esenciálnych rezíduí templátov v zarovnaní ich zhlukov z prvej iterácie MSA filtrovania
<code>msa_filter_b_cluster_count</code>	počet zhlukov z druhej iterácie MSA filtrovania
<code>msa_filter_b_cluster *</code>	zhluky z druhej iterácie MSA filtrovania
<code>msa_filter_b_positions *</code>	pozície esenciálnych rezíduí templátov v zarovnaní ich zhlukov z druhej iterácie MSA filtrovania
<code>accessions</code>	identifikátory výsledných vyfiltrovaných sekvencií
<code>alternative_accessions</code>	alternatívne identifikátory výsledných sekvencií
<code>identified_targets_templates</code>	mapovanie výsledných sekvencií na ich templáty
<code>identified_targets</code>	výsledné sekvencie (FASTA)
<code>identified_targets_gz</code>	výsledné sekvencie (FASTA, GZIP)
<code>identified_targets_residue_acids</code>	aminokyseliny na pozíciách esenciálnych rezíduí vo výsledných sekvenciách

Tabuľka 6.1: Súbory zdieľané medzi modulmi výpočtového jadra. Názvy označené hviezdíčkou sú parametrizované.

Symbolický názov	Popis
Celkové zarovnanie	
<code>overall_alignment_table</code>	tabuľka pozícií esenciálnych rezíduí celkového zarovnania
<code>overall_alignment_success</code>	úspech celkového zarovnania (yes alebo no)
<code>identified_targets_residue_acids_extra</code>	esenciálne rezíduá nájdené vďaka celkovému zarovnaniam
<code>identified_targets_msa</code>	celkové zarovnanie všetkých výsledných sekvencií
Anotácie	
<code>pfam</code>	informácie o Pfam doménach výsledných sekvencií
<code>soluprot</code>	rozpuštnosť výsledných sekvencií predikovaná nástrojom SoluProt
<code>transmembrane_prediction</code>	predikované transmembránové regióny výsledných sekvencií
<code>closest_known</code>	identitou najbližšie vstupné (známe) sekvencie k výsledným sekvenciám
<code>closest_all</code>	identitou najbližšie výsledné sekvencie k výsledným sekvenciám
<code>closest_query</code>	identitou najbližšie vstupné sekvencie, ktoré boli použité ako vstup pre vyhľadávanie, k výsledným sekvenciám
<code>tax_mapping</code>	mapovanie sekvencií na organizmy
<code>taxonomy</code>	názvy organizmov výsledných sekvencií
<code>extremophilicity</code>	informácie o organizmoch z databázy Bioproject
Sieť sekvenčnej podobnosti	
<code>identified_targets_lengths</code>	počet aminokyselín výsledných sekvencií
<code>network_options</code>	dostupné varianty siete
<code>network_nodes_edges</code> *	sieť sekvenčnej podobnosti (JSON)
<code>network_cluster</code> *	reprezentanti a členovia zhukov na vytvorenie siete
<code>network_session</code> *	sieť sekvenčnej podobnosti vo formáte pre Cytoscape [19]
Výsledková tabuľka	
<code>result_table</code>	výsledková tabuľka (JSON)

Tabuľka 6.1: Súbory zdieľané medzi modulmi výpočtového jadra. Názvy označené hviezdíčkou sú parametrizované (pokračovanie tabuľky).

	Halide 1 [NW]	Nucleophile [D]	Halide 2 [W]
D4Z2G1_S1	38	108	109
P22643_S2	125	124	175

Tabuľka 6.2: Pozície esenciálnych rezíduí v sekvenciách templátov esenciálnych rezíduí. Hlavička obsahuje názvy rezíduí a písmená v hranatých zátvorkách označujú povolené aminokyseliny, ktoré môžu plniť rolu esenciálneho rezídua. V prvom stĺpci sú identifikátory sekvencií templátov. Čísla označujú pozície aminokyselín v sekvenciách templátov.

vyhľadávaní sa zoznamy identifikátorov nájdených sekvencií zlúčia a nástrojom FASTAcmd sa získajú sekvencie vo formáte FASTA. Ak hlavička sekvencie obsahuje informáciu, že proteín je umelo vytvorený, alebo ak sekvencia obsahuje neznámu aminokyselinu, tak sa táto sekvencia odstráni. Hlavičky sekvencií môžu obsahovať viacero identifikátorov sekvencie a ďalšie poznámky (anotácie). Primárny identifikátor sa určí týmto spôsobom:

- ak je sekvencia zhodná s nejakou vstupnou sekvenciou, tak primárny identifikátor je identifikátor vstupnej sekvencie,
- inak, ak hlavička sekvencie obsahuje identifikátor PDB štruktúry, tak primárny identifikátor je tento identifikátor PDB štruktúry,
- inak je primárny identifikátor ten identifikátor, ktorý je uvedený v hlavičke ako prvý.

Ak hlavička sekvencie obsahuje identifikátor PDB štruktúry, je tento identifikátor uložený do súboru `seqs_blast_pdb_accessions`. Všetky alternatívne identifikátory sekvencie sa uložia do súboru `seqs_blast_alternative_accessions`. Anotácie z hlavičiek sekvencií sa uložia do súboru `seqs_blast_annotations`. Nakoniec, nájdeným sekvenciám sa nahradí hlavička ich primárnym identifikátorom sa uložia do súboru `seqs_blast`.

Modul Pair Filter

Modul *Pair Filter* odstráni z nájdených sekvencií tie sekvencie, ktoré podľa párového zarovnania nezdieľajú esenciálne rezíduá so žiadnym templátom esenciálnych rezíduí.

Zo súboru `known_sequences` sa vyselektujú sekvencie templátov a uložia sa do súboru `templates`. Templáty sú určené tabuľkou templátov esenciálnych rezíduí (6.2) v súbore `catalytic_aas`. Nástroj USEARCH [3] vytvorí párové zarovnanie každej nájdenej sekvencie (`seqs_blast`) s každým templátom. V každom zarovnaní sa zisťuje, či na pozíciách určenými esenciálnymi rezíduami templátu sú v skúmanej sekvencii povolené aminokyseliny. Na to je potrebné previesť *čisté* pozície rezíduí z tabuľky templátov na *zarovnané* pozície rezíduí – pozície rovnakých aminokyselín, ktoré sú od čistých pozícií väčšie o počet medzier vložených zarovnaním.

Pre každý templát je inicializovaný zhluk a daný templát je do tohto zhluku vložený. Ak sa skúmaná sekvencia zhoduje v esenciálnych rezíduách aspoň s jedným templátom, je zachovaná a vložená do zhluku tohto templátu, inak je odstránená. Ak sa skúmaná sekvencia zhoduje s viacerými templátmi, je vložená do zhluku templátu, s ktorým má najväčšiu identitu. Vytvorené zhluky sa uložia do súborov `pair_filter_cluster`.

Modul MSA Filter

Modul *MSA Filter* vytvorí z templátových zhlukov zarovnanie viacerých sekvencií (v angličtine *multiple sequence alignment*, skratka MSA) a odstráni zo zhlukov tie sekvencie, ktoré sa podľa zarovnania zhľuku nezhodujú s templátom v esenciálnych rezíduách. Toto filtrovanie sa vykonáva v dvoch iteráciách.

Templátové zhľuky vytvorené v modul *Pair Filter* obsahujú jeden templát a sekvencie, ktoré sa podľa párového filtrovania zhodujú s týmto templátom v esenciálnych rezíduách. Pri zarovnaní celého zhľuku je však možné, že vzájomná poloha medzi templátom a sekvenciou zhľuku bude iná než pri párovom zarovnaní (oddiel 2.3).

Na zarovnanie zhlukov sa použije nástroj Clustal Omega [20] a zistia sa zarovnané pozície esenciálnych rezíduí v templátoch, ktoré sa uložia do súborov `msa_filter_a_positions`. Po odstránení s templátom sa nezhodujúcich sekvencií zo zhlukov sa zhľuky uložia ako nezarovnané do súborov `msa_filter_a_cluster`. Z týchto prefiltrovaných zhľuky opäť vytvoria zarovnania, zarovnané pozície esenciálnych rezíduí v templátoch sa uložia do súborov `msa_filter_b_positions` a po odstránení nezhodujúcich sa sekvencií sa tentokrát zarovnané zhľuky – s vloženými medzerami – uložia do súborov `msa_filter_b_cluster`.

Všetky sekvencie v týchto podruhékrát prefiltrovaných zhľukov sú označené ako výsledné proteínové sekvencie a uložia sa do súboru `identified_targets`. Identifikátory sekvencií sa uložia do súboru `accessions` a alternatívne identifikátory sa uložia do súboru `alternative_accessions`. Priradenie výsledných sekvencií k templátom sa uchová v súbore `identified_targets_templates`. Do súboru `identified_targets_residue_acids` sa uložia konkrétne aminokyseliny, ktoré boli nájdené vo výsledných sekvenciách na pozíciách esenciálnych rezíduí.

Modul Overall Alignment

Modul *Overall Alignment* vytvára celkové spoločné zarovnanie všetkých templátových zhlukov. Celkové zarovnanie sa môže, ale nemusí podariť. Predpoklad úspešného zarovnania je rovnaké poradie esenciálnych rezíduí s rovnakým názvom vo všetkých templátoch.

Poradie zlučovania zhlukov je dané vzdialenosťou medzi zhľukmi. Blízke zhľuky – tie s podobnejšími sekvenciami – sú zlúčené ako prvé. Vzdialenosť zhlukov je určená ako priemerná identita párového zarovnania medzi sekvenciami zhlukov. Na zistenie vzdialenosti sa nezarovnávajú všetky sekvencie, ale iba 100 náhodných sekvencií z každého zhľuku.

Tabuľka vzdialeností uchováva vzdialenosti všetkých dvojíc zhlukov. Zhľuky sú na začiatku získané zo súborov `msa_filter_b_cluster`. Dvojica zhlukov s najmenšou vzdialenosťou sa zlúči do jedného nového zhľuku. Zarovnania týchto dvoch zhlukov sa spolu zarovnajú (zlúčia) pomocou nástroja Clustal Omega [20]. Z tabuľky vzdialeností sa odstránia zaniknuté zhľuky, vloží sa nový zhľuk a vypočítajú sa preň vzdialenosti k ostatným zhľukom. Tento proces zlučovania sa opakuje, kým nezostane nakoniec iba jeden výsledný zhľuk so všetkými sekvenciami.

Celkové zarovnanie je úspešné vtedy, ak pre každé esenciálne rezíduum platí, že ak sa nachádza vo viacerých templátoch, tak v celkovom zarovnaní je vo všetkých templátových sekvenciách na rovnakej pozícii. Tým, že je esenciálne rezíduum všetkých templátov na rovnakej pozícii, je toto rezíduum na rovnakých pozíciách vo všetkých výsledných sekvenciách. Pomocou úspešného celkového zarovnania je možné odhaliť vo výsledných sekvenciách dodatočné esenciálne rezíduá, ktoré nemá ich templát, ale má ich iný templát.

Do súboru `overall_alignment_table` sa uloží tabuľka pozícií esenciálnych rezíduí v celkovom zarovnaní. Úspech celkového zarovnania (yes alebo no) sa zaznačí do sú-

boru `overall_alignment_success`. Aminokyseliny na pozíciách esenciálnych rezíduí výsledných sekvencií, ktoré boli nájdené vďaka celkovému zarovnaniu, sa uložia do súboru `identified_targets_residue_acids_extra`. Samotné celkové zarovnanie sa uloží do súboru `identified_targets_msa`.

Anotačné moduly

Modul *Pfam* predikuje pomocou nástroja InterProScan [16] Pfam domény vo výsledných sekvenciách.

Modul *SoluProt* predikuje pomocou vlastného nástroja Loschmidtových laboratórií rozpustnosť výsledných proteínov.

Modul *Transmembrane Prediction* predikuje pomocou nástroja TMHMM [11] transmembránové regióny sekvencií.

Modul *Identity* hľadá ku každej výslednej sekvencii pomocou nástroja USEARCH [3] identitou najbližšie sekvencie z troch skupín: (i) všetky vstupné sekvencie, (ii) všetky vstupné sekvencie, ktoré boli použité ako vstup pre vyhľadávanie príbuzných sekvencií, a (iii) všetky výsledné sekvencie.

Modul *Bioproject Loader* prevádza databázu NCBI BioProject [2] z formátu XML do formátu TSV (tabmi oddelené hodnoty).

Modul *Taxonomy* zistí, z akých organizmov pochádzajú výsledné proteínové sekvencie, a nájde informácie o týchto zdrojových organizmoch. Prepojenie medzi identifikátormi sekvencií a organizmami je z databázy NCBI Taxonomy [5] a informácie o organizmoch pochádzajú z databázy NCBI BioProject.

Modul Network

Modul *Network* vytvára sieť sekvénčnej podobnosti pre výsledné sekvencie (SSN, oddiel 2.4). Najprv sa určia vzdialenosti medzi každou dvojicou sekvencií a potom sa pre rôzne varianty siete určí, ktoré sekvencie v nich budú zobrazené.

Všetky dvojice výsledných sekvencií sa zarovnávajú nástrojom *MMseqs2* [21]. Na základe získaných hodnôt podobnosti sekvencií (*bitscore*) a dĺžok sekvencií (len_1 , len_2) sa vypočíta skóre na dĺžku hrany (*score*). Aby neboli vzdialené sekvencie (uzly) v grafe spojené dlhými hranami, sú hrany so skóre väčším než 40 odstránené. Vzorec na výpočet skóre na dĺžku hrany je získaný zo zdrojových súborov nástroja EFI-EST [6, 14]:

$$score = -\log_{10}(len_1 * len_2) + bitscore * \log_{10} 2$$

Vzhľadom k tomu, že výsledných sekvencií môžu byť tisíce, nie je vždy potrebné zobrazovať každú sekvenciu ako samostatný uzol. Varianty siete združujú skupinu podobných sekvencií do jedného uzlu a sú identifikované percentuálnymi sekvénčnými identitami. V module Network je implementovaných deväť variantov s identitami v rozsahu 10 % až 90 %. Pre každý variant nástroj *MMseqs2* rozradí všetky výsledné sekvencie do zhlukov tak, že všetky sekvencie v každom zhluke zdieľajú s reprezentantom zhľuku identitu danú variantom. Varianty s počtom zhlukov (uzlov) väčším než 800 sa vyradia. Pre každý zo zvyšných variantov sa vytvorí popis grafu vo formáte XGMML založenom na XML. Do grafov sa vložia uzly pre reprezentantov zhlukov a hrany spájajúce tieto uzly s vypočítaným skóre na dĺžku hrany. Nástroj Cytoscape [19] z popisov grafov vytvorí rozmiestnenie grafu v dvojrozmernom priestore na základe dĺžok hrán – uzlom sa určia súradnice x a y . Výsledné rozmiestnené siete z výstupu nástroja Cytoscape sa prevedú z formátu XML do formátu JSON.

Informácie o počte uzlov a hrán dostupných sietí sa uložia do súboru `network_options`. Rozmiestnené siete sekvenčnej podobnosti sa uložia do súborov `network_nodes_edges`. Rerezentanti a členovia zhlukov sekvencií sa uložia do súborov `network_cluster`. Výstupné súbory nástroja Cytoscape vo formáte ZIP sa uložia do súborov `network_session`.

Modul Result

Modul *Result* zhrňuje výsledné sekvencie a ich anotácie do výsledkovej tabuľky, ktorá sa uloží do súboru `result_table`. Jednotlivé polia tabuľky pre výsledné sekvencie sú zobrazené v tabuľke 6.3.

6.2 Implementácia úlohového rozhrania

Úlohové rozhranie poskytuje *webovému serveru* prístup k databáze a súborovému systému. Úlohové rozhranie má HTTP koncové body (URL adresy) na zadanie úloh pre výpočtové jadro, na prevzatie výsledkov úloh a na získanie štatistiky užívateľského používania nástroja. Okrem toho, úlohové rozhranie slúži ako server na podávanie vstupných súborov pre výpočtové jadro. Úlohové rozhranie je implementované v jazyku Java s použitím frameworku Spring Boot. Všetky koncové body sú zobrazené v tabuľke 6.4.

6.3 Implementácia webovej aplikácie

Webová aplikácia je implementovaná ako jednostránková aplikácia v systéme React v jazyku TypeScript, čo je rozšírený JavaScript so statickým typovaním. Na správu globálneho stavu aplikácie sa používa knižnica Redux a na prácu s URL adresou knižnica React Router.

React využíva deklaratívny prístup ku generovaniu a zmene obsahu webovej stránky. Programátor definuje štruktúru stavu aplikácie a usporiadanie komponentov, pričom ich zobrazenie je závislé na stave. React sa stará o to, aby obsah stránky (DOM) bol automaticky zmenený po zmene stavu aplikácie. Zmenou stavu je možné dynamicky meniť obsah stránky bez komunikácie so serverom, čo však má za následok to, že URL adresa v prehliadači je vždy rovnaká. Špecifikácia HTML5 definuje rozhranie na prácu s históriou prehliadača *History API*, čo využíva knižnica React Router na zmenu URL adresy.

Knižnica Redux podporuje znovupoužiteľnosť *zobrazovacích komponentov* a oddeľuje ich od funkcionality. Zobrazovacie komponenty sú parametrizované – zobrazujú obsah podľa hodnôt parametrov bez znalosti pôvodu týchto hodnôt. Parametre môžu byť dáta, ale i referencie na funkcie. Tým pádom môže mať zobrazovacia komponenta pre tlačidlo ako parameter funkciu, ktorá sa má zavolať pri stlačení tlačidla, bez toho, aby bolo telo tejto funkcie v komponente definované.

Programátor si má určiť stromovú štruktúru globálneho stavu a udalosti meniace tento stav. *Kontajnery* určujú hodnoty parametrov pre zobrazovacie komponenty. Kontajnery získavajú dáta zo stavu aplikácie a poskytujú zobrazovacím komponentom funkcie na vyvolanie udalostí meniacich stav. *Reducer* obsluhuje udalosti a ako jediný mení globálny stav aplikácie. Reducer môže poveriť dielčie reducery obsluhou udalostí pre podstromy stavu.

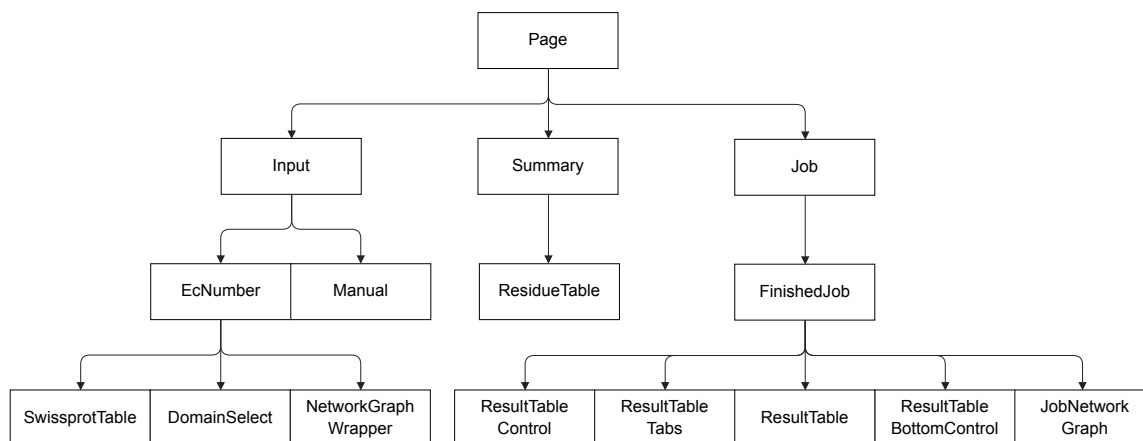
Webová aplikácia implementuje užívateľské rozhranie navrhnuté v oddiele 5.2. Hlavné komponenty sú zobrazené na obrázku 6.2. Zadanie úlohy použitím revidovaných sekvencií umožňuje komponent *EcNumber* (obr. 6.3), pre vlastné sekvencie je komponent *Manual* (obr. 6.4). Zhrnutie zadania úlohy je implementované v komponente *Summary* a výsledkovú stránku pre úspešne dokončenú úlohu zobrazuje komponent *FinishedJob*, kde je výsledková

Názov poľa	Popis
accession	identifikátor sekvencie
annotation	anotácie z hlavičky sekvencie
gi	identifikátor sekvencie GI
domains	Pfam domény
soluprot	rozpustnosť podľa nástroja SoluProt
transmembrane	transmembránový región
closestQuery	identifikátor najbližšej vstupnej vyhľadávacej sekvencie
identityClosestQuery	percentuálna identita s najbližšou vstupnou vyhľadávacou sekvenciou
closestKnown	identifikátor najbližšej vstupnej sekvencie
identityClosestKnown	percentuálna identita s najbližšou vstupnou sekvenciou
closestAll	identifikátor najbližšej výslednej sekvencie
identityClosestAll	percentuálna identita s najbližšou výslednou sekvenciou
organism	názov zdrojového organizmu
kingdom	riša zdrojového organizmu (B pre baktérie, E pre eukaryoty)
salinity	anotácia zdrojového organizmu <i>salinity</i>
optimumTemp	anotácia zdrojového organizmu <i>optimum temperature</i>
tempRange	anotácia zdrojového organizmu <i>temperature range</i>
bioticRel	anotácia zdrojového organizmu <i>biotic relationship</i>
disease	anotácia zdrojového organizmu <i>disease</i>
erTemplate	identifikátor templátu sekvencie
residueMap	priradenie aminokyselín názvom esenciálnych rezíduí
catalyticResidues	aminokyseliny na pozíciách esenciálnych rezíduí spojené do jedného reťazca (napríklad NDWEH)
upi	UPI – identifikátor záznamu v databáze UniParc [23]
firstSeen	dátum prvého zaznamenania sekvencie podľa databázy UniParc
swissprot	identifikátor v databáze revidovaných sekvencií Swiss-Prot [23]
structure	identifikátor PDB štruktúry
sequence	sekvencia – reťazec všetkých aminokyselín
sequenceLength	dĺžka sekvencie – počet aminokyselín

Tabuľka 6.3: Polia výsledkovej tabuľky vytvorenej v module *Result*.

URL adresa	Popis koncového bodu
Trieda EcNumberController	
/ecNumberOptions	našepkávač dostupných hodnôt pri zadávaní EC čísla s použitím knižnice Apache Lucene
/ecNumber	sekvencie z databázy Swiss-Prot pre zadané EC číslo
Trieda NewJobController	
/addNewJob	vloženie novej úlohy pre výpočtové jadro
Trieda FilesController	
/files/{jobId}/{filename}	podávanie vstupných súborov výpočtovému jadru
Trieda CreatedJobController	
/jobInfo	informácie o stave úlohy
/getResultTable	výsledková tabuľka vo formáte JSON na zobrazenie vo webovej aplikácii
/getExtraDomainSheetRows	čísla riadkov na zobrazenie v hárku <i>Extra Domain</i> po zmene primárnych domén
/getResultTableXlsx	výsledková tabuľka vo formáte XLSX
/getResultTableTsv	výsledková tabuľka vo formáte TSV
/getNetworkOptions	dostupné varianty siete sekvenčnej podobnosti
/getNetwork	sieť na zobrazenie vo webovej aplikácii
/getNetworkSession	sieť na zobrazenie nástrojom Cytoscape
/getResultZip	všetky výstupné súbory výpočtového rozhrania v archíve ZIP
/getInputFile	súbor na opätovné zadanie úlohy
/countJobs	počet všetkých vložených úloh pre výpočtové jadro
Trieda PiwikController	
/visitorsCount	počet návštevníkov webovej aplikácie – funkcionality spravovaná Loschmidtovými laboratóriami

Tabuľka 6.4: Koncové body úlohového rozhrania. Koncový bod je HTTP zdroj adresovateľný URL adresou. Koncové body sú v tabuľke rozdelené podľa tried, v ktorých sú implementované.



Obr. 6.2: Strom hlavných komponentov.

tabuľka (obr. 6.5). Všetky tieto logické komponenty majú zobrazovací komponent s príponou *Component* a kontajner s príponou *Container*.

Na stránke pre zadanie úlohy použitím revidovaných sekvencií (obr. 6.3) sa využíva stav na uchovanie výberu sekvencií. Na základe stavu sa zobrazujú riadky tabuľky ako vybrané a uzly siete sa farebne zvyrazňujú. Výber sekvencií je uchovaný v stavovom objekte ako pole identifikátorov sekvencií a mení sa: (i) výberom riadku v tabuľke, (ii) kliknutím na uzol v sekvenčnej sieti, (iii) určením sekvenčnej identity zhlukov a (iv) tlačidlami *Select all* a *Deselect all*.

Na stránke pre zadanie úlohy použitím vlastných sekvencií (obr. 6.4) sa overuje, či sú v templátových sekvenciách na zadaných pozíciách esenciálnych rezíduí povolené aminokyseliny. Po každej zmene polí pre sekvencie sa skontroluje formát zadaných sekvencií a unikátnosť ich identifikátorov. Táto kontrola sa deje najviac raz za jednu sekundu. Využíva sa na to oneskorené vykonávanie, známe ako metóda *debounce*. Pri kontrole sekvencií sa sekvencie prevedú do formy asociatívneho poľa, kde je kľúčom identifikátor sekvencie a hodnotou reťazec aminokyselín. To sa využije pri kontrole pozícií v tabuľke templátov.

Výsledková tabuľka (obr. 6.5) je najzložitejšia časť webovej aplikácie. Tabuľka umožňuje užívateľovi filtrovať a zoradzovať výsledky rôznymi spôsobmi tak, aby si mohol čo najlepšie vytipovať proteínové sekvencie, ktorých enzymatické funkcie podrobí laboratórnemu overeniu. Vybrané sekvencie sa ukladajú do úložiska *localStorage*, čím zostanú vybrané aj po obnovení stránky. Stav výsledkovej tabuľky tvorí: (i) zoznam vybraných riadkov, (ii) výber hárku, (iii) obdobie prvého vloženia sekvencií do databázy, (iv) minimálna rozpustnosť, (v) rozsah sekvenčnej identity k vyhľadávacím vstupným sekvenciám, (vi) primárne domény a (vii) nastavenie prepínačov na vylúčenie sekvencií. Tým, že React zabezpečuje deterministické vykresľovanie jedine na základe stavu, je veľmi jednoduché stav verzovať a pomocou tlačidiel *Undo* a *Redo* umožniť užívateľovi vracať a obnovovať vykonané zmeny.

JOB INPUT

Swiss-Prot sequences ? Custom sequences ?

3.8.1.5 - Haloalkane dehalogenase. (33) Load example

Select sequences from table ? Select sequences from similarity network ?

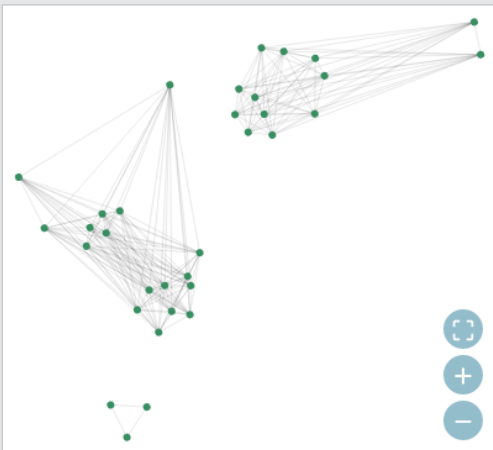
Accession	ER ?	Length	Sequence plot ?
<input type="checkbox"/> A0A1L5BTC1	3	296	
<input type="checkbox"/> A1KLS7	3	300	
<input type="checkbox"/> A5U5S9	3	300	
<input type="checkbox"/> B0SY51	3	302	
<input type="checkbox"/> B2HJU9	3	297	
<input type="checkbox"/> B4RF90	3	301	
<input type="checkbox"/> B8H3S9	3	302	
<input type="checkbox"/> C1AF48	3	300	
<input type="checkbox"/> D4Z2G1	5	296	
<input type="checkbox"/> P0A3G2	3	293	
<input type="checkbox"/> P0A3G3	3	293	
<input type="checkbox"/> P0A3G4	3	293	
<input type="checkbox"/> P22643	5	310	
<input type="checkbox"/> P59336	5	294	

Selected 0 of 33 1 to 14 of 33 |< < Page 1 of 3 > >|

Filter sequences by Pfam domains... | v

Select all Deselect all Show selected only

Advanced options



Select representative sequences of clusters ?

10 % 20 % 30 % 40 % 50 %
60 % 70 % 80 % 90 % 100 %

Obr. 6.3: Zadanie úlohy použitím revidovaných sekvencií z databázy Swiss-Prot. Po zadání EC čísla sa zobrazí tabuľka sekvencií a sekvenčná sieť (SSN). Tabuľka a sieť sú medzi sebou prepojené – výberom riadku v tabuľke sa zafarbí príslušný uzol v sieti a kliknutím na uzol sa vyberie jeho riadok v tabuľke.

JOB INPUT

Swiss-Prot sequences ? Custom sequences ?

Upload input file ? Load example

Query sequences ? Upload FASTA

Sequences in FASTA format...

Other known sequences (optional) ? Upload FASTA

Sequences in FASTA format...

Essential residue templates ? Add protein (row) Add residue (column)

Accession	Enter name
Enter accession	Enter position

Advanced options

Obr. 6.4: Zadanie úlohy použitím vlastných sekvencií. Tabuľka templátov esenciálnych rezíduí má pre každý templát jeden riadok a pre každé esenciálne rezíduum jeden stĺpec. Vnútro tabuľky tvoria pozície esenciálnych rezíduí v templátoch. Templátom môže byť ktorákoľvek zo zadaných sekvencií. Webová aplikácia overí prítomnosť povolených aminokyselín v templátoch na zadaných pozíciách esenciálnych rezíduí. Povolené aminokyseliny sa nastavujú v druhom riadku pomocou tlačidla *Set residues*.

TARGET SELECTION TABLE

Select all Deselect all Undo Redo

First seen ? Earliest date Latest date Histogram

Minimum solubility 0.00 Query identity range 0 100

Primary domains ? PF00561 (Abhydrolase_1) x

Selected Full Dataset Extra Domain Organism Temperature Salinity Biotic Relationship Disease Transmembrane 3D Structure Network

Accession	Annotation	First seen	Closest query	Identity closest query	Kingdom	Solubility	Sequence le
<input type="checkbox"/> 2PSD_A	Chain A, Crystal Struct...	2007-05-23	D4Z2G1_S1	41.5	E	0.9735	318
<input type="checkbox"/> 2PSF_A	Chain A, Crystal Struct...	2007-05-23	D4Z2G1_S1	41.5	E	0.9615	310
<input type="checkbox"/> 2PSJ_A	Chain A, Crystal Struct...	2007-06-06	D4Z2G1_S1	41.5	E	0.9614	319
<input type="checkbox"/> 2PSH_A	Chain A, Crystal Struct...	2007-06-06	D4Z2G1_S1	41.2	E	0.9586	319
<input type="checkbox"/> WP_071575177.1	haloalkane dehalogen...	2016-04-06	D4Z2G1_S1	70.8	B	0.9399	270
<input type="checkbox"/> 3SK0_A	Chain A, structure of ...	2012-06-22	D4Z2G1_S1	46.2	B	0.9393	311
<input type="checkbox"/> 4BRZ_A	Chain A, Haloalkane D...	2014-05-09	D4Z2G1_S1	61.7		0.9357	290
<input type="checkbox"/> AGS48406.1	haloalkane dehalogen...	2014-05-19	D4Z2G1_S1	61.3	B	0.925	292
<input type="checkbox"/> WP_003902310.1	haloalkane dehalogen...	2009-01-17	D4Z2G1_S1	69.3	B	0.9219	300
<input type="checkbox"/> 4C6H_A	Chain A, Haloalkane D...	2014-05-09	D4Z2G1_S1	61.5		0.9202	291
<input type="checkbox"/> 2QVB_A	Chain A, Crystal Struct...	2008-10-13	D4Z2G1_S1	69.5	B	0.9187	297
<input type="checkbox"/> 2O2H_A	MULTISPECIES: haloal...	1996-11-01	D4Z2G1_S1	69.3	B	0.917	300
<input type="checkbox"/> WP_057342388.1	haloalkane dehalogen...	2015-03-28	D4Z2G1_S1	68.9	B	0.9162	300
<input type="checkbox"/> CMB85140.1	haloalkane dehalogen...	2015-03-28	D4Z2G1_S1	67.9	B	0.9141	275
<input type="checkbox"/> 1CIJ_A	Chain A, PROTEIN (HA...	2003-03-27	P22643_S2	99.7	B	0.9136	310
<input type="checkbox"/> WP_055370782.1	haloalkane dehalogen...	2015-03-28	D4Z2G1_S1	69.3	B	0.9134	300

Number of rows: 3903

Exclude: transmembrane sequences extra domain sequences Swiss-Prot sequences ?

Download table (XLSX)

Obr. 6.5: Výsledková tabuľka úspešne dokončenej úlohy. Hárky slúžia ako pohľady na výsledky. Každý hárok zobrazuje inú podmnožinu výsledkov.

Kapitola 7

Overenie nástroja na haloalkán dehalogenázach

Funkčnosť vytvoreného nástroja bola overená na enzymatickej rodine haloalkán dehalogenáz. Nástroj identifikoval 7 433 sekvencií patriacich do tejto rodiny. Pri praktickom použití si užívateľ z týchto výsledných sekvencií na základe anotácií vyberie sadu sekvencií, ktoré podrobí experimentálnemu overeniu funkcie v laboratóriu. Výpočet trval 36 CPU hodín a všetky výsledné súbory majú veľkosť 3 GB. Súbory zdieľané medzi modulmi jadra sú dostupné v elektronickej prílohe (C) v adresári *Overenie nástroja*.

Rodina haloalkán dehalogenáz

Haloalkán dehalogenázy (EC číslo 3.8.1.5, skratka HLD) sú enzýmy, ktoré katalyzujú zmenu haloalkánu (halogénalkánu) a vody na alkohol a halid (halogén). Ako halid sa označujú chemické prvky fluór, chlór, bróm, jód a astát. Dehalogenázy umožňujú rozklad niektorých toxických látok znečisťujúcich životné prostredie na produkty, ktoré toxické už nie sú. Zjednodušený vzorec chemickej reakcie katalyzovanej HLD:



Vstupné súbory

Medzi haloalkán dehalogenázami boli identifikované tri podrodiny: HLD-I, HLD-II a HLD-III. Vstupný súbor vyhľadávacích sekvencií `user_queries` preto obsahuje po jednom zástupcovi z každej podrodiny, aby boli vo výsledkoch zahrnuté príbuzné sekvencie všetkých známych druhov dehalogenáz. Sú to proteíny s označením DhlA (HLD-I), LinB (HLD-II) a DrbA (HLD-III). Okrem týchto sekvencií boli ako templáty použité ďalšie tri sekvencie, po jednej z každej podrodiny. Sú to proteíny s označením DmbB, DhaA a DmbC a sú v súbore `user_known_sequences`. Oba tieto vstupné súbory sú uvedené v prílohe B. Pozície esenciálnych rezíduí sú v tabuľke 7.1.

Hľadanie podobných sekvencií

Pri vyhľadávaní podobných proteínov bolo nájdených 9 784 sekvencií. Tabuľka 7.2 zobrazuje zastúpenie jednotlivých podrodín. Sekvencie nájdené pre viaceré vyhľadávacie sekvencie sú priradené k tej najbližšej podľa hodnoty E-value.

Templát	nucleophile	acid1	acid2	base	halide1	halide2	halide3
	D	DE	DE	H	HNQWY	HNQWY	HNQWY
DhlA	124	–	260	289	–	125	175
DmbB	123	–	250	279	–	124	164
LinB	108	132	–	272	38	109	–
DhaA	106	130	–	272	41	107	–
DrbA	139	–	272	300	71	140	–
DmbC	109	–	238	267	43	110	–

Tabuľka 7.1: Pozície esenciálnych rezíduí v sekvenciách templátov esenciálnych rezíduí. Hlavička obsahuje názvy rezíduí a písmená pod nimi označujú povolené aminokyseliny, ktoré môžu plniť rolu esenciálneho rezídua. V prvom stĺpci sú identifikátory sekvencií templátov. Čísla označujú pozície aminokyselín v sekvenciách templátov.

Sekvencia	Podrodina	Sekvencií	Podiel
DhlA	HLD-I	3 292	33,6 %
LinB	HLD-II	2 416	24,7 %
DrbA	HLD-III	4 076	41,7 %
Σ		9 784	

Tabuľka 7.2: Zastúpenie podrodín vo výsledkoch hľadania podobných sekvencií.

Templát	Sekvencií	Podrodina	Sekvencií		Úbytok	
			Počet	Podiel	Počet	Podiel
DhlA	457	HLD-I	2 089	28,1 %	1 203	36,5 %
DmbB	1 632					
LinB	1 159	HLD-II	2 175	29,3 %	241	10,0 %
DhaA	1 016					
DrbA	2 193	HLD-III	3 169	42,6 %	907	22,3 %
DmbC	976					
Σ			7 433		2 351	24,0 %

Tabuľka 7.3: Výsledky filtrácie. Z každej podrodiny boli použité dva templáty. V druhom stĺpci je uvedený počet výsledných sekvencií v templátových zhluchoch a v štvrtom stĺpci je ich súčet pre podrodinu. Posledné dva stĺpce zobrazujú úbytok oproti počtu nájdených sekvencií (tabuľka 7.2).

Filtrácia

Pri párovom filtrovaní sa nájdené sekvencie priradia k templátom. Tie sekvencie, ktoré sa nezhodujú so žiadnym templátom, sú vylúčené. Z 9 784 nájdených sekvencií sa priradilo k šiestim templátom 7 542 sekvencií a 2 242 sekvencií bolo párovým filtrom vylúčených. Zarovnaním zhlučov sa v MSA filtri odstránilo v prvej iterácii 107 sekvencií a v druhej iterácii ďalšie dve sekvencie z podrodiny HLD-I. Výsledný počet domnelých nájdených dehalogenáz je 7 433. Celkovo bolo filtrovaním vylúčených 24 % z nájdených sekvencií. Rozdelenie výsledných sekvencií medzi templátmi je zobrazené v tabuľke 7.3.

Celkové zarovnanie

Pozície esenciálnych rezíduí v templátových zhluchoch sú v tabuľke 7.4. Za výrazne odlišné pozície zhuku pre templát DhaA môže sekvencia s označením OQO00754.1 s nezvyčajnou dĺžkou 3 372 aminokyselín. Celkové zarovnanie bolo úspešné – všetky rezíduá sú vo všetkých templátoch na rovnakých pozíciách (7.5). Pomocou celkového zarovnania bolo identifikovaných 636 dodatočných esenciálnych rezíduí v 622 sekvenciách. Dodatočné esenciálne rezíduá sekvencie sú tie, ktoré templát sekvencie nemá, ale má ich templát iného zhuku a boli identifikované v celkovom zarovnaní.

Analýza výsledných sekvencií

Sekvenčná podobnosť medzi 7 433 sekvenciami je použitím siete sekvenčnej podobnosti (SSN) vizualizovaná na obrázku 7.1, kde je tiež možné ľahko identifikovať vstupné sekvencie a podrodiny.

Každý rok sú do proteínových databáz vkladané nové sekvencie, pričom tento rast má exponenciálny charakter. To, koľko sekvencií bolo prvýkrát zaznamenaných v jednotlivých rokoch, je zobrazené na obrázku 7.2. Vzhľadom k stúpajúcej tendencii objavovaných proteínových sekvencií má vytvorený nástroj EnzymeMiner veľký potenciál, pretože opätovné spustenie úloh môže poukázať na nové pozoruhodné sekvencie.

Templát	nucleophile	acid1	acid2	base	halide1	halide2	halide3
DhlA	397	–	627	659	–	398	494
DmbB	661	–	964	1 002	–	662	772
LinB	495	522	–	735	405	496	–
DhaA	3 440	3 465	–	3 654	3 359	3 441	–
DrbA	519	–	769	805	359	520	–
DmbC	571	–	833	870	451	572	–

Tabuľka 7.4: Pozície esenciálnych rezíduí v samostatných templátových zhľukoch po MSA filtrovaní. Všetky sekvencie v jednotlivých zhľukoch majú esenciálne rezíduá na týchto pozíciách.

Templát	nucleophile	acid1	acid2	base	halide1	halide2	halide3
DhlA	3 810	–	4 301	4 370	–	3 811	3 980
DmbB	3 810	–	4 301	4 370	–	3 811	3 980
LinB	3 810	3 869	–	4 370	3 607	3 811	–
DhaA	3 810	3 869	–	4 370	3 607	3 811	–
DrbA	3 810	–	4 301	4 370	3 607	3 811	–
DmbC	3 810	–	4 301	4 370	3 607	3 811	–
∪	3 810	3 869	4 301	4 370	3 607	3 811	3 980

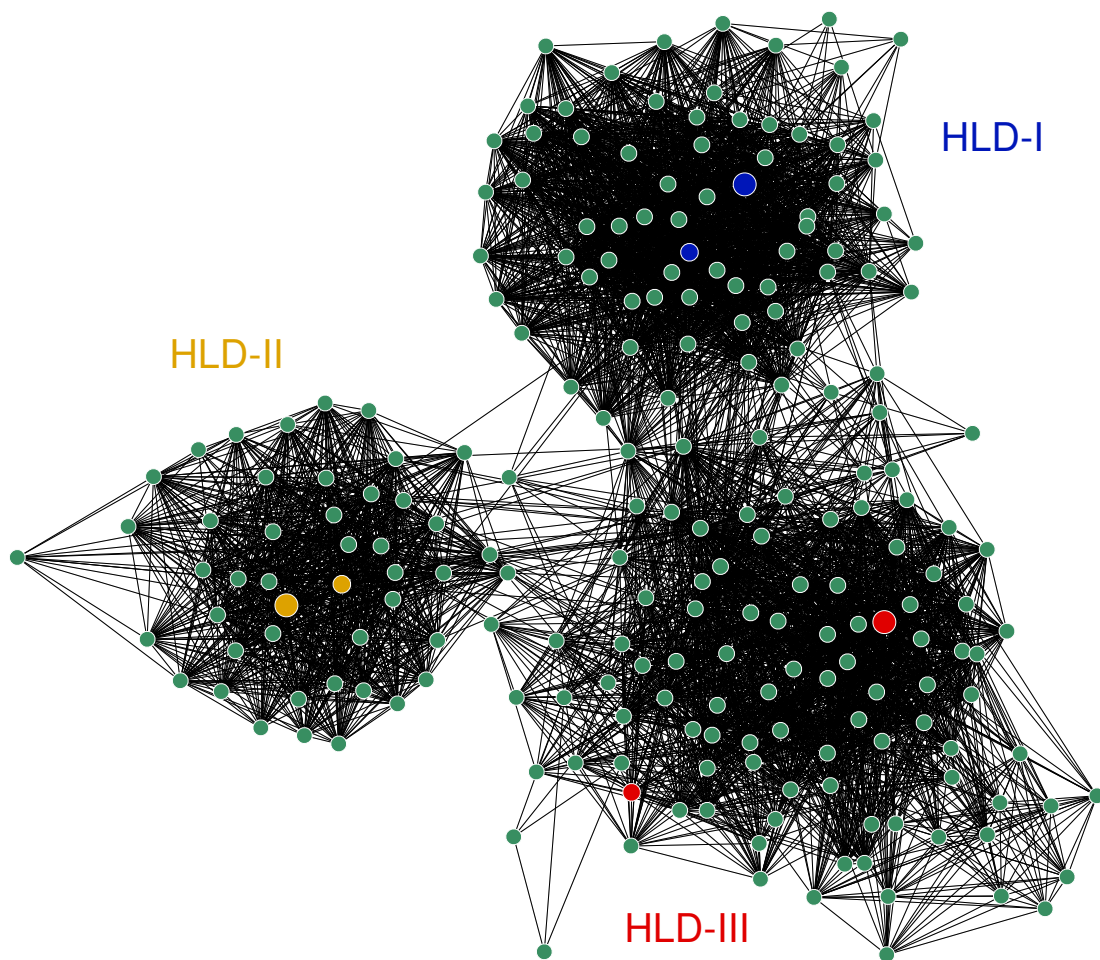
Tabuľka 7.5: Pozície esenciálnych rezíduí v celkovom zarovnaní.

Archaea	Baktérie	Eukaryoty	Σ
39	7 213	126	7 378

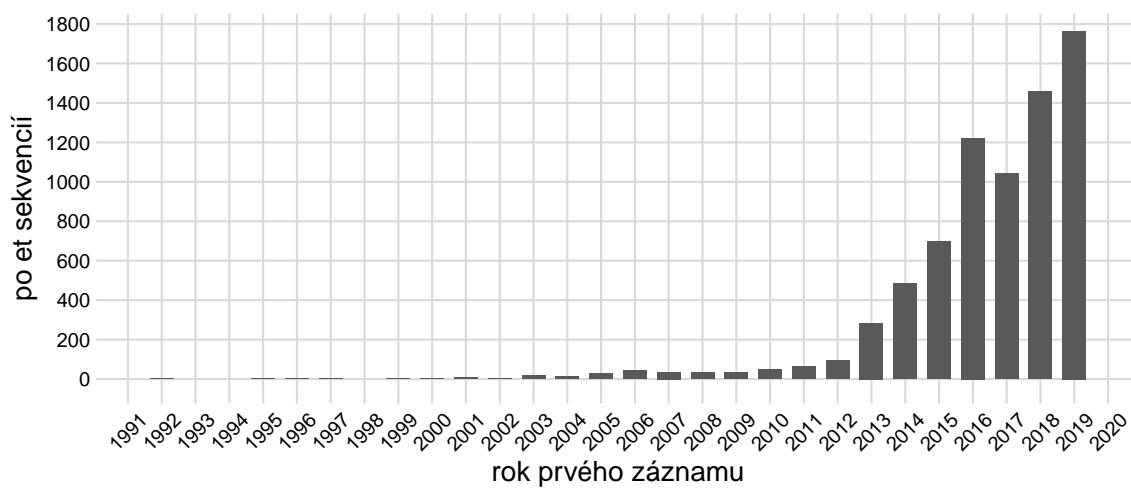
Tabuľka 7.6: Zastúpenie ríš zdrojových organizmov. Celkový súčet nie je rovný celkovému počtu výsledných sekvencií (7 433), pretože nie pre každú sekvenciu je možné nájsť záznam v databáze NCBI Taxonomy.

Termofilné	Psychrofilné	Mierne halofilné	Extrémne halofilné
9	55	113	21

Tabuľka 7.7: Počty zdrojových organizmov s extrémnymi vlastnosťami.



Obr. 7.1: Sieť sekvenčnej podobnosti výsledných proteínových sekvencií. Jednotlivé podrodiny HLD tvoria vizuálne identifikovateľné zhluky. Templáty sú farebne zvýraznené a vyhľadávacie sekvencie majú väčší uzol. Na tomto obrázku je zobrazený variant siete s identitou 50 %, ktorý obsahuje 216 uzlov prepojených 4 713 hranami.



Obr. 7.2: Graf počtu výsledných sekvencií prvýkrát zaznamenaných (*first seen*) v jednotlivých rokoch. Dátumy sú získané z databázy UniParc [23].

Kapitola 8

Záver

Proteínové databázy rastú rýchlym tempom, zatiaľ čo klasické biochemické metódy nie sú schopné charakterizovať toto množstvo proteínov v adekvátnom čase. Táto práca sa preto zameriava na vyhľadávanie príbuzných enzýmov, a to s ohľadom na ich esenciálne rezíduá. Vytvorený nástroj EnzymeMiner získava pre výsledné enzýmy anotácie, na základe ktorých sa používateľ rozhodne, ktoré proteíny podrobí experimentálnemu overeniu v laboratóriu, ktoré potvrdí alebo vyvráti katalytickú funkciu daného proteínu. Webové rozhranie nástroja je verejne dostupné na adrese <https://loschmidt.chemi.muni.cz/enzymeminer/>. Nástroj bol publikovaný v odbornom časopise Nucleic Acids Research (impakt faktor 11,147). Článok, ktorého som zdieľaným prvým autorom, je pripojený v prílohe D.

Existujúcim prístupom na získavanie príbuzných proteínov bez znalosti trojrozmernej štruktúry je vyhľadávanie proteínov v databázach, napríklad pomocou nástroja BLAST alebo cez webové rozhrania dostupných databáz. Získané proteíny je však často potrebné podrobiť ďalším analýzám. Pri enzýmoch je možné zohľadniť tiež pozície esenciálnych rezíduí. Práve to je v tejto práci hlavným princípom pri vyhľadávaní príbuzných enzýmov.

Hlavným prínosom nástroja EnzymeMiner vytvoreného v tejto práci oproti existujúcim prístupom je flexibilita vstupu a interaktívna prezentácia výsledkov bohatá na anotácie. Používateľ môže zadať vlastné sekvencie a vlastný popis esenciálnych rezíduí na vyhľadávanie špecifických enzymatických rodín alebo podrodín. Výstupom nástroja je výsledková tabuľka s možnosťou označovania, ktorá obsahuje sekvencie s anotáciami rozdelené do hárkov podľa rôznych kritérií. Táto tabuľka pomáha vybrať rôznorodú sadu sekvencií na laboratórnu charakterizáciu. Dva kľúčové uprednostňovacie kritériá sú: (i) predikovaná rozpustnosť a (ii) sekvenčná identita k vyhľadávacím sekvenciám doplnená interaktívnou sieťou sekvencnej podobnosti zobrazenou priamo vo webovom rozhraní, ktorá môže byť použitá na nájdenie rozmanitých sekvencií. Nástroj EnzymeMiner je univerzálny nástroj použiteľný pre akúkoľvek enzymatickú rodinu. Redukuje čas potrebný na zber dát, viackrokovú analýzu a uprednostnenie výsledných sekvencií z dní na hodiny. Všetky funkcie nástroja a potrebné nástroje sú integrované do serverového riešenia a na používanie nástroja nie sú potrebné žiadne externé nástroje. Výpočet úlohy môže trvať niekoľko hodín až dní v závislosti na zadanom vstupe a vyťažení výpočtových prostriedkov.

Nástroj EnzymeMiner je zložený z niekoľkých funkčných jednotiek spojených do robustnej architektúry (kapitola 5). *Výpočtové jadro* realizuje hlavný pracovný postup nástroja, prijíma úlohy z databázy, na náročné operácie využíva výpočtový cluster a ukladá výsledné súbory na disk. Výpočtové jadro sa skladá z troch krokov: (i) vyhľadávanie príbuzných proteínových sekvencií, (ii) filtrácia sekvencií pomocou enzýmových templátov esenciálnych rezíduí a (iii) anotácia výsledných enzýmov. Anotácie zahrňujú informácie o zdrojových

organizmoch, predikované domény, predikciu rozpustnosti a vizualizáciu sekvenčného priestoru pomocou siete sekvenčnej podobnosti (SSN). *Úlohové rozhranie* zapuzdruje prácu s databázou a prácu so súbormi do jednej vrstvy a umožňuje pridávať nové úlohy a sťahovať výsledky dokončených úloh. *Webová aplikácia* vytvára webové užívateľské rozhranie umožňujúce (i) zadať novú úlohu pomocou revidovaných sekvencií z databázy Swiss-Prot alebo pomocou vlastných sekvencií a (ii) zobrazíť výsledky v interaktívnej výsledkovej tabuľke prepojenej so sieťou sekvenčnej podobnosti výsledných enzýmov.

Funkčné jednotky boli implementované vhodnými technológiami s ohľadom na najnovšie trendy a udržateľnosť (kapitola 6). Výpočtové jadro, databáza, úlohové rozhranie a webový server sú spustené a prepojené systémom Docker. *Výpočtové jadro* zabezpečujúce hlavný výpočet nástroja je implementované v jazyku Java vo frameworku Loschmidtových laboratórií. *Úlohové rozhranie* je implementované v jazyku Java ako HTTP aplikačné rozhranie pomocou frameworku Spring Boot. *Webový server* je spustený cez Node.js, podáva súbory potrebné na zobrazenie webovej aplikácie a slúži ako brána k úlohovému rozhraniu. *Webová aplikácia* je implementovaná ako jednostránková aplikácia v systéme React s využitím správcu stavu Redux.

Funkčnosť vytvoreného nástroja bola overená na enzymatickej rodine haloalkán dehalogenáz (kapitola 7). Na spustenie úlohy sa použili tri vyhľadávacie sekvencie a šesť templátov, rovnomerne z každej z troch podrodín haloalkán dehalogenáz. Počet výsledných sekvencií bol 7 433, z čoho takmer 1 800 sekvencií bolo prvýkrát zaznamenaných v roku 2019. Výpočet trval 36 CPU hodín a vygeneroval dáta s veľkosťou 3 GB.

Na ďalší rozvoj nástroja EnzymeMiner sú v ponuke tri vylepšenia: (i) predikcia terciárnej (priestorovej) štruktúry výsledných sekvencií, (ii) automatické periodické vyhľadávanie – nástroj bude spúšťať analýzu periodicky a bude informovať užívateľa o nových sekvenciách, ktoré boli nájdené od posledného vyhľadávania, a (iii) inteligentný sprievodca nastavenia filtračných kritérií vo výsledkovej tabuľke.

Literatúra

- [1] ALTSCHUL, S. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*. Oxford University Press (OUP). september 1997, zv. 25, č. 17, s. 3389–3402. DOI: 10.1093/nar/25.17.3389. Dostupné z: <https://doi.org/10.1093/nar/25.17.3389>.
- [2] BARRETT, T., CLARK, K., GEVORGYAN, R., GORELENKOV, V., GRIBOV, E. et al. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Research*. Oxford University Press (OUP). december 2011, zv. 40, D1, s. D57–D63. DOI: 10.1093/nar/gkr1163. Dostupné z: <https://doi.org/10.1093/nar/gkr1163>.
- [3] EDGAR, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. Oxford University Press (OUP). august 2010, zv. 26, č. 19, s. 2460–2461. DOI: 10.1093/bioinformatics/btq461. Dostupné z: <https://doi.org/10.1093/bioinformatics/btq461>.
- [4] EL GEBALI, S., MISTRY, J., BATEMAN, A., EDDY, S. R., LUCIANI, A. et al. The Pfam protein families database in 2019. *Nucleic Acids Research*. Oxford University Press (OUP). október 2018, zv. 47, D1, s. D427–D432. DOI: 10.1093/nar/gky995. Dostupné z: <https://doi.org/10.1093/nar/gky995>.
- [5] FEDERHEN, S. The NCBI Taxonomy database. *Nucleic Acids Research*. Oxford University Press (OUP). december 2011, zv. 40, D1, s. D136–D143. DOI: 10.1093/nar/gkr1178. Dostupné z: <https://doi.org/10.1093/nar/gkr1178>.
- [6] GERLT, J. A., BOUVIER, J. T., DAVIDSON, D. B., IMKER, H. J., SADKHIN, B. et al. Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*. Elsevier BV. august 2015, zv. 1854, č. 8, s. 1019–1037. DOI: 10.1016/j.bbapap.2015.04.015. Dostupné z: <https://doi.org/10.1016/j.bbapap.2015.04.015>.
- [7] HENIKOFF, S. a HENIKOFF, J. G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*. Proceedings of the National Academy of Sciences. november 1992, zv. 89, č. 22, s. 10915–10919. DOI: 10.1073/pnas.89.22.10915. Dostupné z: <https://doi.org/10.1073/pnas.89.22.10915>.
- [8] HON, J. *Vyhledávání příbuzných proteinů s modifikovanou funkcí*. Brno, CZ, 2015. Diplomová práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedúci práce MARTÍNEK, T. Dostupné z: <https://www.fit.vut.cz/study/thesis/17337/>.

- [9] KOMENSKÝ, J. A. *Labyrint světa a ráj srdce*. František Jeřábek, 1809. 69–70 s. Kapitola XI: Motaniny filozofů, odstavec 8: Mezi aritmetiky. Dostupné z: <https://books.google.cz/books?id=GyAoAAAAyAAJ>.
- [10] KOONIN, E. V. a GALPERIN, M. Y. *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics* [online]. Boston: Kluwer Academic, 2003 [cit. 2020-04-27]. Dostupné z: <https://www.ncbi.nlm.nih.gov/books/NBK20255/>.
- [11] KROGH, A., LARSSON, B., HEIJNE, G. von a SONNHAMMER, E. L. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. Cohen. *Journal of Molecular Biology*. Elsevier BV. január 2001, zv. 305, č. 3, s. 567–580. DOI: 10.1006/jmbi.2000.4315. Dostupné z: <https://doi.org/10.1006/jmbi.2000.4315>.
- [12] MĂNDOIU, I., SUNDERRAMAN, R. a ZELIKOVSKY, A. *Bioinformatics Research and Applications: Fourth International Symposium, ISBRA 2008, Atlanta, GA, USA, May 6-9, 2008, Proceedings*. Springer, 2008. 51 s. LNCS sublibrary: Bioinformatics. ISBN 9783540794493. Dostupné z: https://books.google.cz/books?id=TKyr_V6eofgC.
- [13] NAVARRO, G. A guided tour to approximate string matching. *ACM Computing Surveys (CSUR)*. Association for Computing Machinery (ACM). marec 2001, zv. 33, č. 1, s. 31–88. DOI: 10.1145/375360.375365. Dostupné z: <https://doi.org/10.1145/375360.375365>.
- [14] OBERG, N. *Súbor xgmml_create_all.pl v repozitári EnzymeFunctionInitiative/EST* [online]. GitHub, Inc., 4. augusta 2018. Revidované 26. 6. 2019 [cit. 2020-05-23]. Vzorec na výpočet skóre pre tvorbu siete sekvenčnej podobnosti v nástroji EnzymeMiner je získaný zo zdrojových súborov nástroja EFI-EST [6]. Dostupné z: https://github.com/EnzymeFunctionInitiative/EST/blob/7eb600fd54ca6e890ac9d0af37dc865b47eb3468/xgmml_create_all.pl#L288.
- [15] OKONECHNIKOV, K., GOLOSOVA, O. a FURSOV, M. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics*. Oxford University Press (OUP). február 2012, zv. 28, č. 8, s. 1166–1167. DOI: 10.1093/bioinformatics/bts091. Dostupné z: <https://doi.org/10.1093/bioinformatics/bts091>.
- [16] QUEVILLON, E., SILVENTOINEN, V., PILLAI, S., HARTE, N., MULDER, N. et al. InterProScan: protein domains identifier. *Nucleic Acids Research*. Oxford University Press (OUP). júl 2005, zv. 33, Web Server, s. W116–W120. DOI: 10.1093/nar/gki442. Dostupné z: <https://doi.org/10.1093/nar/gki442>.
- [17] SAYERS, E. W., AGARWALA, R., BOLTON, E. E., BRISTER, J. R., CANESE, K. et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*. Oxford University Press (OUP). november 2018, zv. 47, D1, s. D23–D28. DOI: 10.1093/nar/gky1069. Dostupné z: <https://doi.org/10.1093/nar/gky1069>.
- [18] SCHOLZ, M. *E-value & Bit-score* [online]. [cit. 2020-05-06]. Dostupné z: <http://web.archive.org/web/20200506183113/http://www.metagenomics.wiki/tools/blast/evalue>.

- [19] SHANNON, P. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*. Cold Spring Harbor Laboratory. november 2003, zv. 13, č. 11, s. 2498–2504. DOI: 10.1101/gr.1239303. Dostupné z: <https://doi.org/10.1101/gr.1239303>.
- [20] SIEVERS, F., WILM, A., DINEEN, D., GIBSON, T. J., KARPLUS, K. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*. EMBO. január 2011, zv. 7, č. 1, s. 539. DOI: 10.1038/msb.2011.75. Dostupné z: <https://doi.org/10.1038/msb.2011.75>.
- [21] STEINEGGER, M. a SÖDING, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*. Springer Science and Business Media LLC. október 2017, zv. 35, č. 11, s. 1026–1028. DOI: 10.1038/nbt.3988. Dostupné z: <https://doi.org/10.1038/nbt.3988>.
- [22] SUNDARRAM, A. a MURTHY, T. P. K. Alpha-Amylase Production and Applications: A Review. *Journal of Applied & Environmental Microbiology*. Science and Education Publishing. 2014, zv. 2, č. 4, s. 166–175. DOI: 10.12691/jaem-2-4-10. Dostupné z: <http://pubs.sciepub.com/jaem/2/4/10>.
- [23] THE UNIPROT CONSORTIUM. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*. November 2018, zv. 47, D1, s. D506–D515. DOI: 10.1093/nar/gky1049. ISSN 0305-1048. Dostupné z: <https://doi.org/10.1093/nar/gky1049>.
- [24] VANACEK, P., SEBESTOVA, E., BABKOVA, P., BIDMANOVA, S., DANIEL, L. et al. Exploration of Enzyme Diversity by Integrating Bioinformatics with Expression Analysis and Biochemical Characterization. *ACS Catalysis*. American Chemical Society (ACS). február 2018, zv. 8, č. 3, s. 2402–2412. DOI: 10.1021/acscatal.7b03523. Dostupné z: <https://doi.org/10.1021/acscatal.7b03523>.

Príloha A

Pokročilé nastavenia úlohy

V tejto prílohe sú uvedené parametre ovplyvňujúce výpočet úlohy rozdelené do troch skupín: vyhľadávanie, filtrácia a vizualizácia. Pri každom parametri je zobrazený (i) názov uvedený vo webovom rozhraní, (ii) kľúč (identifikátor) použitý v implementácii, (iii) implicitná hodnota a (iv) popis parametra. Implementácia výpočtového jadra je predstavená v oddiele [6.1](#).

Vyhľadávanie

Názov: E-value

Kľúč: evalue

Hodnota: 1e-20

Popis: Počet očakávaných výsledkov, ktoré sa nájdu pri náhodnom vyhľadávaní v databáze danej veľkosti. Hodnota E-value je ohodnotením podobnosti nájdenej sekvencie s vyhľadávacou sekvenciou. Toto nastavenie určuje maximálnu hodnotu E-value, ktorú môže mať nájdená sekvencia.

Názov: Inclusion E-value threshold

Kľúč: inclusion_ethresh

Hodnota: 1e-20

Popis: Hranica štatistickej významnosti na zahrnutie sekvencie do modelu používanom nástrojom PSI-BLAST pri vykonávaní ďalšej iterácie vyhľadávania.

Názov: Number of iterations

Kľúč: num_iterations

Hodnota: 2

Popis: Počet iterácií vyhľadávania nástrojom PSI-BLAST.

Názov: Maximum number of hits

Kľúč: max_best_hits

Hodnota: 10000

Popis: Ohraničenie počtu výsledkov vyhľadávania. Všetky sekvencie nájdené nástrojom PSI-BLAST sa zoradia a ďalšej analýze sú podrobené iba najlepšie výsledky.

Názov: Artificial tags
Kľúč: `artificial.tags`
Hodnota: artificial, synthetic construct, vector, vaccinia virus, plasmid, halotag, replicon
Popis: Všetky sekvencie s týmito značkami v popise sú vylúčené.

Filtrácia

Názov: Minimum identity
Kľúč: `min_identity`
Hodnota: 25
Popis: Prah minimálnej sekvenčnej identity. Všetky výsledné sekvencie musia mať prahovú sekvenčnú identitu s aspoň jednou vyhľadávacou sekvenciou. Identita je vypočítaná globálnym zarovnaním (Needleman-Wunsch).

Názov: Clustal iterations
Kľúč: `clustalo_iter`
Hodnota: 3
Popis: Počet iterácií pri konštrukcii zarovnaní viacerých sekvencií nástrojom Clustal Omega.

Vizualizácia

Názov: Score threshold
Kľúč: `score_cutoff`
Hodnota: 40
Popis: Minimálne skóre na vloženie hrany do grafu siete sekvenčnej podobnosti.

Príloha B

Ukážka vstupných súborov

Ukážka proteínových sekvencií použitých na overenie nástroja (kapitola 7). Všetky súbory zdieľané medzi modulmi jadra sú dostupné v elektronickej prílohe (C) v adresári *Overenie nástroja*.

Súbor `user_queries`

Sekvencie použité pri vyhľadávaní sekvencií z rodiny haloalkán dehalogenáz.

>Dh1A

```
MINAIRTPDQRFNSNLDQYPPSPNYLDDLPGYPLRAHYLDEGNSDAEDVFLCLHGEP  
TWSYLVRKMIPVFAESGARVIAPDFFGFKSDKPVDEEDYTFEFHRNFLALIERL  
DLRNITLVVQDWGGFLGLTLPADPSRFKRLIIMNACLMTDPVTQPAFSAFVT  
QPADGFTA  
WKYDLVTPSDLRLDQFMKRWAPTLTEAEASAYAAPPDTSYQAGVRKFKMVA  
QRDQACIDISTEASISFWQNDWNGQTFMAIGMKDKLLGPDVMYPMKALING  
CPEPLEIADAGHFVQEFGEQVAR  
EALKHFAETE
```

>LinB

```
MSLGAKPFGEKKFIEIKGRRMAYIDEGTGDPILFQHGNPTSSYLWRNIMPH  
CAGLGR  
LIACDLIGMGSDKLDPSGPERYAYAEHRDYLDALWEALDLGDRVVLVVDW  
GSALGFDWARRHRERVQGIAYMEAIAMPIEWADFPEQDRDLFQAFRSQAGE  
ELVLDQNVFVEQVLPGLILRPLSEAE  
MAAYREPFLAAGEARRPTLSWPRQIPIAGTPADVVAIARDYAGWLS  
ESP  
IPKLFINAEPGALTTGRMRDFCRTWPNQTEITVAGAHFIQEDSPDEIGAA  
IAAFVRRLRPA
```

>DrbA

```
MSCRLSSNRRGSSKLAAMTNLASDLFPHPSSELSIDGHTLRYIDTAASSD  
IPSSAVGSSDGEPTFLCVHGNPTWSFYRRRIIERYGKQQRVIAVDHIGGR  
SDKPS  
EDEFPYTMAAHRDNLIRLVDEL  
DLKNVILIAHDWGGAIGLSAMHARRDRLAGIGLLNTAAFP  
PPYMPQRIAA  
CRMPVLGTPAVRGLNLFARA  
AVTMAMSR  
TKMKPDVAAGLLAPYDNWKNRVAIDRFV  
RDIPLNDSHPTMKT  
LRQLES  
DLPDLASLPISLIWGMKDWC  
FRPECLRRFQSVWPDAE  
VTE  
LATGHHYVIEDSPEETLAAIDSL  
LARVKERIGAA
```

Súbor user_known_sequences

Súbor user_known_sequences obsahuje okrem vyhľadávacích sekvencií nasledovné sekvencie, ktoré sú použité ako templáty esenciálnych reziduí.

>DmbB

```
MDVLRTPDSRFEHLVGYPFAPHYVDVTAGDTQPLRMHYVDEGPGDGPPIVLLHGPTWSY
LYRTMIPPLSAAGHRVLAPDLIGFGRSDKPTRIEDYTYLRHVEWVTSWFENLDLHDVTLF
VQDWGSLIGLRIAAEHGDRIARLVVANGFLPAAQGRTPLPFYVWRAFARYSPVLPAGRLV
NFGTVHRVPAGVRAGYDAPFPDKTYQAGARAFPRLVPTSPDDPAVPANRAAWEALGRWDK
PFLAIFGYRDPILGQADGPLIKHIPGAAGQPHARIKASHFIQEDSGTELAERMLSWQQAT
```

>DhaA

```
MSEIGTGFPFDPHYVEVLGERMHYVDVGPVDGTPVFLHGNPTSSYLWRNIIPHVAPSHR
CIAPDLIGMGKSDKPDLDYFFDDHVRYLDAFIEALGLEEVVLVIHDWGSALGFHWAKRNP
ERVKGIACMEFIRPIPTWDEWPEFARETFQAFRTADVGRELIIDQNAFIEGALPKCVVRP
LTEVEMDHYREPFLKPVREPLWRFPNELPIAGEPANIVALVEAYMNWLHQSPVPKLLFW
GTPGVLIPPAEAARLAESLPNCKTVDIGPGLHYLQEDNPDLIGSEIARWLPAL
```

>DmbC

```
MSIDFTDPQLYPFESRWFSSRGRHYVDEGTGPPILLCHGNPTWSFLYRDIIVALRDR
FRCVAPDYLGFGLSERPSGFGYQIDEHARVIGEFVDHLGLDRYLSMGQDWGGPISMAVAV
ERADRVRGVVLGNTWFWPADTLAMKAFSRVMSPPVQYAILRRNFFVERLIPAGTEHRPS
SAVMAHYRAVQPNAAARRGVAEMPQKILAAARPLLARLAREVPATLGTKPTLLIWGMKDVA
FRPKTIIPRLSATFPDHVVLVELPNAKHFIQEDAPDRIAAAI IERFG
```

Príloha C

Spríevodca elektronickou prílohou

Elektronická príloha obsahuje nasledujúce adresáre:

- **Text BP** – Text tejto práce vo formáte PDF a zdrojové súbory pre systém L^AT_EX.
- **Overenie nástroja** – Súbory zdieľané medzi modulmi výpočtového jadra (tabuľka 6.1) pre haloalkán dehalogenázy (kapitola 7), prepojenie medzi symbolickými a skutočnými názvami súborov vo formáte XML a použitá konfigurácia úlohy vo formáte XML.
- **EnzymeMiner HPC scripts** – Skripty na spustenie úloh pre výpočtový cluster.
- **EnzymeMiner Core** – Zdrojové súbory výpočtového jadra (oddiel 6.1).
- **EnzymeMiner Commons** – Zdrojové súbory zdieľané medzi výpočtovým jadrom a úlohovým rozhraním.
- **EnzymeMiner Backend** – Zdrojové súbory úlohového rozhrania (oddiel 6.2).
- **EnzymeMiner Client** – Zdrojové súbory webovej aplikácie (oddiel 6.3).
- **EcNumber Loader** – Skripty v jazyku Python na spracovanie dát pre EC skupiny proteínov. Získané dáta sa používajú vo webovom rozhraní na výber revidovaných sekvencií. Skripty v tomto adresári extrahujú z databázy Swiss-Prot EC skupiny, sekvencie patriace do jednotlivých EC skupín a pozície esenciálnych rezíduí. Pfam domény sa vo všetkých sekvenciách určia nástrojom InterProScan. Pre každú EC skupinu sa vytvorí sieť sekvenčnej podobnosti (SSN) a identifikujú sa reprezentatívne sekvencie s rôznymi prahmi sekvenčnej podobnosti.
- **Swiss-Prot** – Prevod databázy Swiss-Prot z formátu XML do formátu SQLite 3.
- **UniParc** – Prevod databázy UniParc z formátu XML do formátu SQLite 3.

Príloha D

Článok v časopise Nucleic Acids Research

Odborný časopis Nucleic Acids Research (NAR) publikuje výsledky špičkového výskumu nukleových kyselín a proteínov. Impakt faktor časopisu NAR je 11,147. Článok s názvom „EnzymeMiner: automated mining of soluble enzymes with diverse structures, catalytic properties and stabilities“ bol vypracovaný pod vedením tímu Loschmidtových laboratórií a bol zverejnený 11. 5. 2020.

EnzymeMiner: automated mining of soluble enzymes with diverse structures, catalytic properties and stabilities

Jiri Hon^{1,2,3,†}, Simeon Borko^{1,2,†}, Jan Stourac^{1,3}, Zbynek Prokop^{1,3}, Jaroslav Zendulka², David Bednar^{1,3}, Tomas Martinek² and Jiri Damborsky^{1,3,*}

¹Loschmidt Laboratories, Department of Experimental Biology and Research Center for Toxic Compounds in the Environment RECETOX, Faculty of Science, Masaryk University, Brno, Czech Republic, ²IT4Innovations Centre of Excellence, Faculty of Information Technology, Brno University of Technology, Bozotechnova 2, Brno, Czech Republic and ³International Clinical Research Center, St. Anne's University Hospital Brno, Brno, Czech Republic

Received March 08, 2020; Revised April 13, 2020; Editorial Decision April 27, 2020; Accepted April 29, 2020

ABSTRACT

Millions of protein sequences are being discovered at an incredible pace, representing an inexhaustible source of biocatalysts. Despite genomic databases growing exponentially, classical biochemical characterization techniques are time-demanding, cost-ineffective and low-throughput. Therefore, computational methods are being developed to explore the unmapped sequence space efficiently. Selection of putative enzymes for biochemical characterization based on rational and robust analysis of all available sequences remains an unsolved problem. To address this challenge, we have developed EnzymeMiner—a web server for automated screening and annotation of diverse family members that enables selection of hits for wet-lab experiments. EnzymeMiner prioritizes sequences that are more likely to preserve the catalytic activity and are heterologously expressible in a soluble form in *Escherichia coli*. The solubility prediction employs the in-house SoluProt predictor developed using machine learning. EnzymeMiner reduces the time devoted to data gathering, multi-step analysis, sequence prioritization and selection from days to hours. The successful use case for the haloalkane dehalogenase family is described in a comprehensive tutorial available on the EnzymeMiner web page. EnzymeMiner is a universal tool applicable to any enzyme family that provides an interactive and easy-to-use web interface freely available at <https://loschmidt.chemi.muni.cz/enzymeminer/>.

INTRODUCTION

There are currently >259 million non-redundant protein sequences in the NCBI nr database (release 2020-02-10) (1). Despite their enormous promise for biological and biotechnological discovery, experimental characterization has been performed on only a small fraction of the available sequences. Currently, there are about 560 000 protein sequences reliably curated in the UniProtKB/Swiss-Prot database (release 2020.01) (2).

The low ratio of characterized to uncharacterized sequences reflects the sharp contrast in time-demanding/low-throughput biochemical techniques versus fast/high-throughput next-generation sequencing technology. Although more efficient biochemical techniques employing miniaturization and automation have been developed (3–5), the most widely used experimental methods do not provide sufficient capacity for biochemical characterization of proteins spanning the ever-increasing sequence space. Therefore, computational methods are currently the only way to explore the immense protein diversity available among the millions of uncharacterized sequence entries.

Two different computational strategies are generally used for exploration of the unknown sequence space. The first strategy takes a novel uncharacterized sequence as input and predicts functional annotations. The method involves annotating the unknown input sequences by predicting protein domains (6), Enzyme Commission (EC) number (7) or Gene Ontology terms that are a subject of the initiative named the Critical Assessment of Functional Annotation (8). These methods are often universal and applicable to any protein sequence. However, they often lack specificity as the automatic annotation rules or statistical models need to be substantially general. A significant advantage of these methods is their seamless integration into available

*To whom correspondence should be addressed. Tel: +420 5 4949 3467; Email: jiri@chemi.muni.cz

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

databases. Submission of a query sequence to a database is sufficient, with no need for running computation- and memory-intensive bioinformatics pipelines locally. A model example of this approach is the automatic annotation workflow of the UniProtKB/TrEMBL database (2).

The second strategy takes a well-known characterized sequence as an input and applies a computational workflow, typically based on a homology search, to identify novel uncharacterized entries in genomic databases that are related to the input query sequence (5,9). The homology search is often followed by a filtration step, which checks the essential sequence properties, e.g. domain structure or presence of catalytic residues. The main advantage of these methods is the higher specificity of the analysis. A disadvantage is that it may be complicated to apply the developed workflow to protein families other than those for which it was designed. Moreover, these workflows typically require running complex bioinformatics pipelines and are usually not available through a web interface.

The fundamental unsolved problem is how to deal with the overwhelming number of sequence entries identified by these methods and select a small number of relevant hits for in-depth experimental characterization. For example, a database search for members of the haloalkane dehalogenase model family using the UniProt web interface yields 3598 sequences (UniProtKB release 2020.01). It is impossible to rationally select several tens of targets for experimental testing without additional bioinformatics analyses to help prioritize such a large pool of sequences.

To address the challenge of exploring the unmapped enzyme sequence space and rational selection of attractive targets, we have developed the EnzymeMiner web server. EnzymeMiner identifies novel enzyme family members, comprehensively annotates the targets and facilitates efficient prioritization and selection of representative hits for experimental characterization. To the best of our knowledge, there is currently no other tool available that allows such a comprehensive analysis in a single easy-to-run integrated workflow on the web.

MATERIALS AND METHODS

EnzymeMiner implements a three-step workflow: (i) homology search, (ii) essential residue based filtering and (iii) hits annotation (Figure 1). To execute these tasks, the server requires two different types of input information: (i) query sequences and (ii) essential residue templates. The query sequences serve as seeds for the initial homology search. The essential residue templates, defined as pairs of a protein sequence and a set of essential residues in that sequence, allow the server to prioritize hits that are more likely to display the enzyme function. Therefore, the essential residues may be the catalytic and ligand- or cofactor-binding residues that are indispensable for proper catalytic function. Each essential residue is defined by its name, position and a set of allowed amino acids for that position.

In the first *homology search step*, a query sequence is used as a query for a PSI-BLAST (10) two-iteration search in the NCBI nr database (1). If more than one query sequence is provided, a search is conducted for each sequence separately. Besides a minimum *E*-value threshold 10^{-20} , the PSI-

BLAST hits must share a minimum of 25% global sequence identity with at least one of the query sequences. Artificial protein sequences, i.e. sequences described by the term artificial, synthetic construct, vector, vaccinia virus, plasmid, halotag or replicon, are removed. EnzymeMiner sorts the PSI-BLAST hits by *E*-value and passes a maximum of 10,000 best hits to the next steps in the workflow. The default parameters for the homology search step, as well as the other steps, can be modified using advanced options in the web server.

In the second *essential residue based filtering step*, the homology search hits are filtered using the essential residue templates. First, the hits are divided into template clusters. Each cluster contains all hits matching essential residues of a particular template. Essential residues are checked using global pairwise alignment with the template calculated by USEARCH (11). When multiple essential residue templates match, the hit is assigned to the template with the highest global sequence identity. Second, for each cluster, an initial multiple sequence alignment (MSA) is constructed using Clustal Omega (12). The MSA is used to revalidate the essential residues of identified hits by checking the corresponding column in the MSA. Sequences not matching essential residues of the template are removed from the cluster. Third, the MSA is constructed again for each template cluster and the essential residues are checked for the last time. The final set of identified sequences reported by EnzymeMiner contains all sequences left in the template clusters.

In the third *annotation step*, the identified sequences are annotated using several databases and predictors: (i) transmembrane regions are predicted by TMHMM (13), (ii) Pfam domains are predicted by InterProScan (14), (iii) source organism annotation is extracted from the NCBI Taxonomy (15) and the NCBI BioProject database (16), (iv) protein solubility is predicted by the in-house tool SoluProt for prediction of soluble protein expression in *Escherichia coli* and (v) sequence identities to queries, hits or other optional sequences are calculated by USEARCH (11). SoluProt is based on a random forest regression model that employs 36 sequence-based features (<https://loschmidt.chemi.muni.cz/soluprot/>). It has been shown to achieve an accuracy of 58%, specificity of 73% and sensitivity of 44% on a balanced independent test set of 3788 sequences (Hon et al., manuscript in preparation). Alternative solubility prediction tools are summarised in a recently published review (17). It is not advised to use the solubility score for other expression systems because it was trained solely on *E. coli* data. We expect further intensive development of protein solubility predictors in coming years and will ensure that the solubility score in the EnzymeMiner stays at the cutting-edge in terms of its accuracy and reproducibility.

The sequence space of the identified hits is visualized using representative sequence similarity networks (SSNs) generated at various clustering thresholds using MMseqs2 (18) and Cytoscape (19). SSNs provide a clean visual approach to identify clusters of highly similar sequences and rapidly spot sequence outliers. SSNs proved to facilitate identification of previously unexplored sequence and function space (20). The SSN generation method used in EnzymeMiner is inspired by the EFI-EST tool (21). The minimum align-

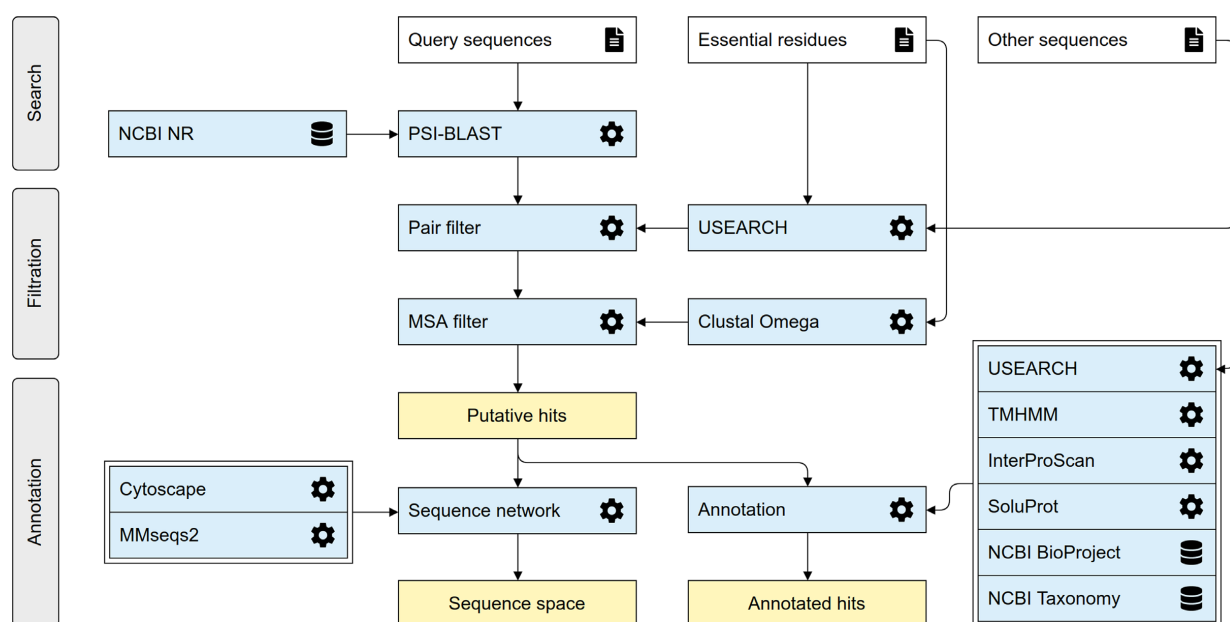


Figure 1. The EnzymeMiner workflow. The workflow consists of three distinct steps: (i) sequence homology search, (ii) filtration of functional sequences, and (iii) annotation of hits. These steps are executed consecutively and automatically. EnzymeMiner has only two required inputs: (i) query sequences, and (ii) essential residue templates. The *Other sequences* are optional inputs that allow EnzymeMiner to calculate the sequence identity between these sequences and all the hits. Input files are highlighted by a white background, tools and databases have a light blue background, outputs are highlighted by a yellow background.

ment score to include an edge between two representative sequences in an SSN is 40.

DESCRIPTION OF THE WEB SERVER

Job submission

New jobs can be submitted from the EnzymeMiner homepage. EnzymeMiner provides two conceptually different ways to define the input of the workflow: (i) using curated sequences from the UniProtKB/Swiss-Prot database and (ii) using custom sequences. We recommend the UniProtKB/Swiss-Prot option for users who do not have in-depth knowledge of the enzyme family. In contrast, the *Custom sequences* tab gives full control over the EnzymeMiner input—query sequences and essential residue templates are specified manually by the user. This is recommended for users who have good knowledge about the enzyme family and want to provide additional starting information to obtain refined results. The last option is a combination of both approaches, where Swiss-Prot sequences can be pre-selected first and then the input can be modified in the *Custom sequences* tab.

In the *Swiss-Prot sequences* tab (Figure 2A), sequences from the Swiss-Prot database can be queried by Enzyme Commission (EC) number. As a result, a table of all sequences annotated by the EC number and corresponding SSN is generated. The table has four columns: (i) sequence accessions hyperlinked to the UniProt database, (ii) number of essential residues, (iii) sequence length and (iv) sequence plot. The sequence plot summarizes two important features of the sequence – positions of essential residues and identi-

fied Pfam domains. The positions of essential residues are obtained from the Swiss-Prot database. The SSN visualizes the sequence space of all the sequences in the current EC group. Nodes represent Swiss-Prot sequences, whereas edge lengths are proportional to the pairwise sequence identities. Similar sequences are close to each other, whereas more distant sequences are not connected at all.

There are three strategies possible for selecting Swiss-Prot sequences as the EnzymeMiner query: (i) select a row from the sequence table, (ii) select a node in the SSN and (iii) select cluster representatives by defining a sequence identity threshold. The sequence identity threshold buttons select cluster representatives at the given percentage threshold. Using this feature, the user can automatically select a small set of sequences that cover the whole known sequence space of the current EC group. All selected Swiss-Prot sequences are used as a query in the homology search step and also as essential residue templates for the filtration step. To modify the selected sets of queries and essential residue templates, the user can switch to the *Custom sequences* tab and refine the selection manually.

EnzymeMiner results

The results page is organized into four sections: (i) *job information* box, (ii) *download results* box, (iii) *target selection table* and (iv) *sequence similarity network*.

In the *job information* box, the user can find the job ID, title, start time and status of the job. There is also a rerun button for rerunning the same analysis without the need for re-entering the same input. This feature is handy for periodically mining new sequences as the sequence databases

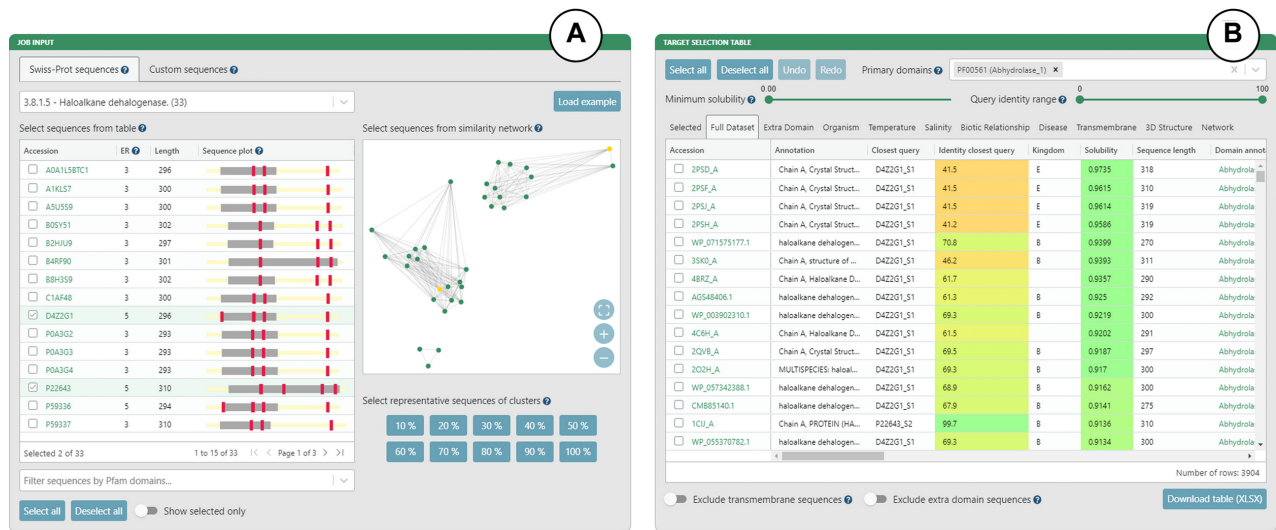


Figure 2. The EnzymeMiner graphical user interface showing example inputs and results for the haloalkane dehalogenase family (EC 3.8.1.5). (A) Inputs based on curated sequences from the UniProtKB/Swiss-Prot database. The input sequences can be selected using: (i) the sequence table, (ii) the SSN or (iii) the sequence identity threshold. (B) Target selection table. The table is organized into eleven sheets that summarize the results from different perspectives. The table can be filtered using solubility and identity sliders, and transmembrane and extra domain exclusion switches.

grow. For example, there are hundreds of new hits for the haloalkane dehalogenase family every year. In the *download results* box, the user can download the results table in XLSX format or tab-separated text format. A ZIP archive containing all output files from the EnzymeMiner workflow can also be downloaded.

The *target selection table* is the most important component of the EnzymeMiner results (Figure 2B). It presents all the putative enzyme sequences identified by EnzymeMiner and helps to select targets for experimental characterization. The table is organized into eleven sheets summarizing the results from different perspectives. (i) The *Selected* sheet shows all the sequences selected from individual sheets. It contains an extra column to track the argument used for the selection. By default, it is prefilled by the name of the sheet from which the sequence was selected, but it can be freely changed. (ii) The *Full Dataset* sheet shows all identified sequences. (iii) The *Extra domain* sheet shows sequences with extra Pfam domains found in the sequence but not listed in the *Primary domains* selection box. (iv) The *Organism* sheet shows sequences with known source organisms. (v) The *Temperature* sheet shows sequences from organisms having extreme optimum temperature annotation in the NCBI BioProject database, including sequences from thermophilic or cryophilic organisms. (vi) The *Salinity* sheet shows sequences from organisms having extreme salinity annotation in the NCBI BioProject database. (vii) The *Biotic Relationship* sheet shows sequences from organisms having biotic relationship annotation in the NCBI BioProject database. (viii) The *Disease* sheet shows sequences from organisms having disease annotation in the NCBI BioProject database. (ix) The *Transmembrane* sheet shows sequences with transmembrane regions predicted by the TMHMM tool. (x) The *3D Structure* sheet shows sequences with an available 3D structure in

the Protein Data Bank (22). (xi) The *Network* sheet shows sequences clustered into a selected sequence similarity network node.

There are four options for filtering the identified sequences displayed in the target selection table. The first option is the minimum solubility slider. Sequences with lower predicted solubility will be hidden. We recommend setting the solubility threshold to >0.5 to increase the probability of finding soluble protein expression in *E. coli*. We do not recommend to set the solubility threshold too high because of possible trade-off between enzyme solubility and activity (23). The second option is the identity range bar. Only sequences with identity to query sequences in the specified range will be visible. The third option is to exclude transmembrane proteins. We recommend removing these sequences as they are usually difficult to produce and tend to have lower predicted solubility. The fourth option is to exclude proteins with an extra domain. Extra domains are defined as domains found in the sequence but not listed in the *Primary domains* selection box. We recommend avoiding sequences with extra domains, but these sequences may also show interesting and unusual biological properties. The selection table can be sorted by clicking on a column header. Holding 'Shift' while clicking on the column headers allows sorting by multiple columns.

The SSN visualizes the sequence space of all identified sequences. Both clusters of similar sequences and sequence outliers can be easily identified. As there might be thousands of sequences, the sequences are clustered at the identity threshold and only an SSN of the representative sequences is shown for performance reasons. Sequences having greater sequence identity are consolidated into a single metanode. Edges indicate high sequence identity between representative sequences of the connected metanodes. Clicking on a metanode displays the *Network* sheet showing

which sequences are represented by a particular metanode. The SSN can be downloaded as a Cytoscape session file for further analysis and custom visualization. Networks clustered at different identities are available. The numbers of nodes and edges are indicated for each identity threshold. The SSN is interactively linked to the target selection table. All nodes representing selected sequences are automatically highlighted in the SSN.

Target selection

The target selection table and SSN facilitate the selection of a diverse set of soluble putative enzyme sequences for experimental validation. First, we recommend setting the maximum sequence identity to queries to 90%. This will remove all hits that are very similar to already known proteins. Second, we recommend selecting a few sequences from individual sheets to cover different phyla from the domains Archea, Bacteria and Eukarya. The most exciting enzymes might be from extremophilic organisms. Third, the SSN can be used to check that the selection covers all sequence clusters. Fourth, users can select sequences from all subfamilies of the enzyme family of interest. The members of different subfamilies can be easily recognized by the *Closest query* or *Closest known* column in the selection table (note: requires representative sequences of subfamilies as job input). Fifth, the available filtering options can be used to (i) prioritize sequences with the highest predicted solubility, (ii) prioritize sequences with known tertiary structures, (iii) eliminate proteins with predicted transmembrane regions and (iv) eliminate sequences with extra domains.

EXPERIMENTAL VALIDATION OF THE EnzymeMiner WORKFLOW

The EnzymeMiner workflow has been thoroughly experimentally validated using the model enzymes haloalkane dehalogenases (5). The sequence-based search identified 658 putative dehalogenases. The subsequent analysis prioritized and selected 20 candidate genes for exploration of their protein structural and functional diversity. The selected enzymes originated from genetically unrelated Bacteria, Eukarya and, for the first time, also Archaea and showed novel catalytic properties and stabilities. The workflow helped to identify novel haloalkane dehalogenases, including (i) the most catalytically efficient enzyme ($k_{\text{cat}}/K_{0.5} = 96.8 \text{ mM}^{-1} \text{ s}^{-1}$), (ii) the most thermostable enzyme showing a melting temperature of 71°C, (iii) three different cold-adapted enzymes active at near to 0°C, (iv) highly enantioselective enzymes, (v) enzymes with a wide range of optimal operational temperature from 20 to 70°C and an unusually broad pH range from 5.7–10 and (vi) biocatalysts degrading the warfare chemical yperite and various environmental pollutants. The sequence mining, annotation, and visualization steps from the workflow published by Vanacek and co-workers (5) were fully automated in the EnzymeMiner web server. The successful use case for the haloalkane dehalogenase family is described in an easy-to-follow tutorial available on the EnzymeMiner web page. Additional extensive validation of the fully automated version of EnzymeMiner,

experimentally testing the properties of another 45 genes of the haloalkane dehalogenases, is currently ongoing in our laboratory.

CONCLUSIONS AND OUTLOOK

The EnzymeMiner web server identifies putative members of enzyme families and facilitates their prioritization and well-informed manual selection for experimental characterization to reveal novel biocatalysts. Such a task is difficult using the web interfaces of the available protein databases, e.g. UniProtKB/TrEMBL and NCBI Protein, since additional analyses are often required. The major advantage of EnzymeMiner over existing protein sources is the flexibility of input and concise annotation-rich interactive presentation of results. The user can input custom queries and a custom description of essential residues to focus the search on specific protein families or subfamilies. The output of EnzymeMiner is an interactive selection table containing the annotated sequences divided into sheets based on various criteria. The table helps to select a diverse set of sequences for experimental characterization. Two key prioritization criteria are (i) the predicted solubility score, which can be used to prioritize the identified sequences and increase the chance of finding enzymes with soluble protein expression, and (ii) the sequence identity to query sequences complemented with an interactive SSN displayed directly on the web, which can be used to find diverse sequences. Additionally, source organism and domain annotations help to select sequences with diverse properties. EnzymeMiner is a universal tool applicable to any enzyme family. It reduces the time needed for data gathering, multi-step analysis and sequence prioritization from days to hours. All the EnzymeMiner features are implemented directly on the web server and no external tools are required. The web server was optimized for modern browsers including Chrome, Firefox and Safari. An EnzymeMiner job can take a few hours or days to compute, depending on the current load of the server. In the next EnzymeMiner version, we plan three major improvements. First, we will implement automated tertiary structure prediction based on homology modeling and threading for all identified sequences. The structural predictions will allow subsequent analysis of active site pockets/cavities and access tunnels. Structural features will significantly enrich the set of annotations and help to identify additional attractive targets for experimental characterization. Second, we will implement automated periodical mining. When enabled, EnzymeMiner will rerun the analysis periodically and inform the user about novel sequences found since the last search. Finally, we will implement a wizard for automated selection of hits based on input criteria provided by a user.

ACKNOWLEDGEMENTS

We thank the participants of the 1st Hands-on Computational Enzyme Design Course (Brno, Czech Republic) for giving valuable feedback on the EnzymeMiner user interface. Their comments inspired us to make the sequence similarity network visualization more interactive.

FUNDING

Czech Ministry of Education [857560, 02.1.01/0.0/0.0/18_046/0015975, CZ.02.1.01/0.0/0.0/16.026/0008451, CZ.02.1.01/0.0/0.0/16_019/0000868, LQ1602]; European Commission [720776, 814418]; AI Methods for Cybersecurity and Control Systems project of the Brno University of Technology [FIT-S-20-6293]; Computational resources were supplied by the project ‘e-Infrastruktura CZ’ [e-INFRA LM2018140] provided within the program Projects of Large Research, Development and Innovations Infrastructures, and by the ELIXIR-CZ project [LM2015047], part of the international ELIXIR infrastructure. Funding for open access charge: Czech Ministry of Education.

Conflict of interest statement. None declared.

REFERENCES

- Sayers, E.W., Agarwala, R., Bolton, E.E., Brister, J.R., Canese, K., Clark, K., Connor, R., Fiorini, N., Funk, K., Hefferon, T. *et al.* (2019) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **47**, D23–D28.
- UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
- Colin, P.-Y., Kintsjes, B., Gielen, F., Miton, C.M., Fischer, G., Mohamed, M.F., Hyvönen, M., Morgavi, D.P., Janssen, D.B. and Hollfelder, F. (2015) Ultrahigh-throughput discovery of promiscuous enzymes by picodroplet functional metagenomics. *Nat. Commun.*, **6**, 1–12.
- Beneyton, T., Thomas, S., Griffiths, A.D., Nicaud, J.-M., Dreville, A. and Rossignol, T. (2017) Droplet-based microfluidic high-throughput screening of heterologous enzymes secreted by the yeast *Yarrowia lipolytica*. *Microb. Cell Fact.*, **16**, 18.
- Vanacek, P., Sebestova, E., Babkova, P., Bidmanova, S., Daniel, L., Dvorak, P., Stepankova, V., Chaloupkova, R., Brezovsky, J., Prokop, Z. *et al.* (2018) Exploration of enzyme diversity by integrating bioinformatics with expression analysis and biochemical characterization. *ACS Catal.*, **8**, 2402–2412.
- El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A. *et al.* (2019) The Pfam protein families database in 2019. *Nucleic Acids Res.*, **47**, D427–D432.
- Li, Y., Wang, S., Umarov, R., Xie, B., Fan, M., Li, L. and Gao, X. (2018) DEEPre: sequence-based enzyme EC number prediction by deep learning. *Bioinformatics*, **34**, 760–769.
- Zhou, N., Jiang, Y., Bergquist, T.R., Lee, A.J., Kacsóh, B.Z., Crocker, A.W., Lewis, K.A., Georgiadiou, G., Nguyen, H.N., Hamid, M.N. *et al.* (2019) The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol.*, **20**, 244.
- Mak, W.S., Tran, S., Marcheschi, R., Bertolani, S., Thompson, J., Baker, D., Liao, J.C. and Siegel, J.B. (2015) Integrative genomic mining for enzyme function to enable engineering of a non-natural biosynthetic pathway. *Nat. Commun.*, **6**, 1–10.
- Altschul, S. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.
- Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.
- Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
- Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R. and Lopez, R. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res.*, **33**, W116–W120.
- Federhen, S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.
- Barrett, T., Clark, K., Gevorgyan, R., Gorelenkov, V., Gribov, E., Karsch-Mizrachi, I., Kimelman, M., Pruitt, K.D., Resenchuk, S., Tatusova, T. *et al.* (2012) BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res.*, **40**, D57–D63.
- Musil, M., Konegger, H., Hon, J., Bednar, D. and Damborsky, J. (2019) Computational design of Stable and Soluble Biocatalysts. *ACS Catal.*, **9**, 1033–1054.
- Steinegger, M. and Söding, J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, **35**, 1026–1028.
- Shannon, P. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
- Copp, J.N., Akiva, E., Babbitt, P.C. and Tokuriki, N. (2018) Revealing unexplored sequence-function space using sequence similarity networks. *Biochemistry*, **57**, 4651–4662.
- Gerlt, J.A., Bouvier, J.T., Davidson, D.B., Imker, H.J., Sadkhin, B., Slater, D.R. and Whalen, K.L. (2015) Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochim. Biophys. Acta (BBA) - Proteins Proteomics*, **1854**, 1019–1037.
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
- Klesmith, J.R., Bacik, J.-P., Wrenbeck, E.E., Michalczyk, R. and Whitehead, T.A. (2017) Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. *Proc. Natl Acad. Sci. U.S.A.*, **114**, 2265–2270.