Excel@FIT 2015
http://excel.fit.vutbr.cz

Vyhledávání příbuzných enzymů s modifikovanou funkcí v proteinových databázích

Jiří Hon*

Abstrakt
Hledání příbuzných enzymů v biologických databázích patří mezi obvyklé činnosti v oblasti proteinového inženýrství a pro tento účel existuje řada zavedených nástrojů. Chceme-li však stávajícími nástroji hledat příbuzné enzymy *s modifikovanou funkcí* – a zamýšlíme tím modifikaci pouze ve smyslu rychlosti reakce, stability enzymu apod., nikoliv přímo typu enzymatické funkce – pak spoléháme na slabou hypotézu, že sekvenčně podobné enzymy mají pravděpodobně stejný typ funkce. Důsledkem toho je buď příliš velké množství nežádoucích výsledků vyhledávání, nebo naopak jejich nedostatečná různorodost. Proto ve spolupráci s Loschmidtovými laboratořemi navrhuji rozšířit současné nástroje o nové filtrační kroky omezující výsledky pouze na relevantní enzymy, jejichž základem je zohlednění pozice a typu katalytických reziduí hledaného enzymu a doplnění anotací. Implementací a aplikací filtračních kroků se mi podařilo dosáhnout řádového snížení počtu výsledků při zachování jejich různorodosti a díky tomu jsem usnadnil a zlevnil výběr zajímavých sekvencí v množině nalezených putativních enzymů, které by mohly mít vhodné vlastnosti pro proteinové inženýrství a vyplatilo by se investovat do jejich experimentálního ověření v laboratoři.

Klíčová slova: Příbuzné proteiny — Ověření funkce proteinu — Zachování aktivního místa

Přiložené materiály: *N/A*

*xhonji01@stud.fit.vutbr.cz, *Fakulta informačních technologií, Vysoké učení technické v Brně*

1. Úvod

Proteinové inženýrství je obor s bohatými možnostmi uplatnění v praxi – od potravinářského průmyslu[1, 2, 3], přes ochranu životního prostředí[4, 5], lékařství[6, 7], až k produkci biopolymerů[8, 9] nebo vývoji nanobiotechnologií[10, 11]. Úspěch v tomto oboru je významně určen schopností, co nejdříve využít veškeré nově dostupné informace o proteinech. Situace je o to složitější, že nové informace přibývají rychlostí, která je nezvládnutelná pro člověka – například jen počet proteinových sekvencí v biologických databázích roste přibližně exponenciálně[12]. Proto každá nová strojová analýza biologických dat nebo i každé zlepšení existujících analýz může znamenat výrazný posun v oboru.

Vyhledávání příbuzných enzymů v biologických databázích je jednou z typických automatizovaných úloh v proteinovém inženýrství, nejčastěji je prováděno za účelem konstrukce vícenásobného zarovnání a analýzy konzervovanosti jednotlivých reziduí. Na příbuzné enzymy je ale možné nahlížet i z jiného úhlu – jako na nové sekvence, které by mohly být zahrnuty jako vstupy pro metody proteinového inženýrství a tím zvýšit šanci na úspěch při vývoji enzymu s vylepšenými vlastnostmi. V tomto případě se jedná o specializované vyhledávání příbuzných enzymů s konkrétní modifikovanou funkcí, přičemž modifikací v tomto případě není myšlen odlišný typ funkcionality proteinu – např. změna katalyzovaného substrátu u enzymů – ale její modifikace ve smyslu rychlosti reakce, stability proteinu, apod.

Nástroj, který by takovou analýzu řešil, by měl poskytovat relativně malý počet výsledků s co největší různorodostí a zároveň s co nejvyšší biologickou relevancí. Míru různorodosti lze ověřit oproti existujícím přístupům k vyhledávání příbuzných proteinů na základě sekvenční podobnosti. Pokud nastavíme parametry porovnávaných nástrojů tak, aby počet výsledků

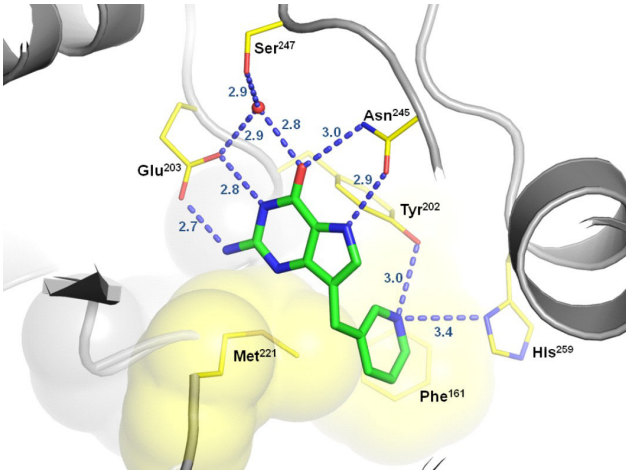
byl přibližně stejný, můžeme ověřit míru podobnosti nalezených sekvencí vzhledem ke vstupním enzymům, která implikuje různorodost. Enzymy jsou ale druhově i funkčně natolik rozmanité molekuly, že biologickou relevanci nalezených putativních enzymů je obtížné automatizovaně hodnotit. Proto je ponecháno na posouzení proteinových inženýrů případ od případu, které výsledky jsou z biologického hlediska zajímavé a které nikoliv. Nicméně je vhodné pro usnadnění posouzení výsledků a dobrou orientaci doplnit výsledky o užitečné anotace z dostupných databází a nástrojů, zejména o informace týkající se tvaru a umístění kapes nebo tunelů vedoucích k místu, na které se váže substrát a probíhá v něm reakce specifická pro daný typ enzymu – tzv. *aktivní místo* (viz obrázek 1).

2. Existující přístupy

V současné literatuře není dohledatelná zmínka o žádném již existujícím nástroji pro vyhledávání příbuzných proteinů *s modifikovanou funkcí*. V některých článcích sice lze objevit analýzy zahrnující hledání příbuzných proteinů a ověření jejich funkce [13], ale nejedná se o obecný postup nebo algoritmus, který by byl použitelný i pro jiné typy proteinů, respektive enzymů, než pro ty, na něž byla konkrétní studie zaměřena. Tento stav je možné vysvětlit tím, že se doteď pravděpodobně nikdo na problém nepodíval z nového úhlu pohledu. Když si totiž zadání problému rozdělíme na dvě části, za prvé na problém hledání příbuzných proteinů a za druhé na problém predikce funkce proteinu, pak v každé z těchto oblastí nalezneme dostatečné množství literatury a nástrojů. Zatím tedy pouze nikdo nevytvořil algoritmus, který by propojoval existující přístupy pro vyhledávání příbuzných proteinů s predikcí jejich funkce, aby ve výsledku vedly k vyhledávání příbuzných proteinů *s modifikovanou funkcí*.

Nejznámějším nástrojem pro hledání homologních sekvencí je BLAST[14], který aproximuje funkci algoritmu Smith-Waterman[15], ale díky použité heuristice je mnohonásobně rychlejší při zachování rozumné přesnosti. Mezi dalšími nástroji stojí za zmínku například PatternHunter[16] a nejnovější HHblits[17], jehož základem je konstrukce a porovnání profile-HMM[18]. Citlivost homologního vyhledávání lze ovlivnit nastavením prahové hodnoty E-value (Expect value), která vyjadřuje minimální požadovanou statistickou významnost jednotlivých výsledků. E-value lze interpretovat jako přibližný počet výsledků se stejnou mírou podobnosti vzhledem ke vstupním sekvencím, který můžeme očekávat, pokud bychom prohledávali náhodnou databázi o stejné velikosti.

Pro predikci funkce proteinu existuje několik me-



Obrázek 1. Prostorová ilustrace aktivního místa enzymu *Sm*PNP[23]. Aktivní místo je oblast na povrchu nebo uvnitř enzymu, na kterou se váže substrát a probíhá v něm katalyzovaná reakce. Žlutou barvou jsou označena tzv. katalytická rezidua, tj. aminokyseliny, bez jejichž přítomnosti a správné pozice nemůže reakce proběhnout. Zelená molekula značí navázaný substrát, v tomto konkrétním případě inhibitor aktivního místa.

tod. Některé z nich, založené na sekvenčním motivu, předpokládají, že je možné v neznámé sekvenci vyhledat známý sekvenční motiv – například z databáze PROSITE[19] – a přiřadit k nim anotace naznačující funkci, kterou obvykle plní proteiny s daným sekvenčním motivem. Další metody staví na hypotéze, že struktura proteinu je obecně více konzervovaná než jeho sekvence, proto strukturní podobnost dvou proteinů je dobrým indikátorem toho, že by mohly mít i stejnou funkci. Například nástroj ProFunc[20] využívá znalosti struktury proteinu s neznámou funkcí a snaží se nalézt strukturně podobné proteiny v databázi PDB [21]. Pokud struktura proteinu s neznámou funkcí není dostupná, což je obecně častější případ, mohou se některé nástroje, například I-TASSER[22], nejdříve pokusit strukturu proteinu predikovat a teprve poté vyhledávat v databázi známých struktur.

3. Vyhledávání enzymů s modifikovanou funkcí

Vzhledem k neexistenci nástroje, který by propojoval identifikaci homologních sekvencí s funkční predikcí, která je nutná pro vyhledávání enzymů s modifikovanou funkcí, navrhuji rozšířit jeden z nástrojů pro hledání příbuzných proteinů o filtrační kroky, které by zohlednily specifika aktivního místa výchozího enzymu a tím přispěly k výrazné redukci irelevantních výsledků. Výběr konkrétního nástroje, na jehož výsledky bude aplikována filtrace, není klíčový –

všechny totiž fungují zpravidla na principu sekvenční podobnosti a liší se spíše jen rychlostí nebo citlivostí vyhledávání. Základem samotné filtrace je ověření přítomnosti a správné pozice všech katalytických reziduí nutných pro průběh enzymatické reakce (pro ilustraci viz obrázek 1). V dalších krocích je vytvořena terciární struktura pomocí homologního modelování s tím, že jsou pozice katalytických reziduí v prostoru zafixovány. Pokud se podaří sestrojit takovýto homologní model s malou chybou, tzn. s vysokým skóre, je vysoce pravděpodobné, že nalezená homologní sekvence bude mít i stejnou enzymatickou funkci.

4. Návrh a implementace

Návrh výsledného nástroje je koncipován jako posloupnost několika navazujících bioinformatických analýz a filtračních kroků. V takovémto návrhu zpravidla výstup jedné analýzy slouží jako vstup analýzy další, proto bude i následující popis nástroje strukturován podle jednotlivých kroků – celkový náhled na průběh workflow ilustruje obrázek 2. Všechny filtrační kroky a automatizované spouštění potřebných nástrojů jsem implementoval v jazyce Java a vycházejí z myšlenek a zkušeností odborníků v Loschmidtových laboratořích. Přínosem z technického hlediska je především vhodná kombinace jednotlivých kroků a transformace vstupů a výstupů mezi nimi, která zpravidla není přímočará.

4.1 Povinné vstupy vyhledávání

Známé sekvence – sekvence enzymů s požadovanou funkcí, například aktuálně experimentálně potvrzené sekvence bez mutantů.

Doplňující sekvence – sekvence putativních enzymů, které nejsou laboratorně potvrzené, ale je pravděpodobné, že budou mít požadovanou funkci.

Sekvence pro vyhledávání – sekvence proteinů, jenž budou použity ve fázi prohledávání databáze všech známých proteinů. Optimálně aspoň jedna sekvence pro každou podrodinu. Pokud má jako vyhledávací sekvence sloužit např. vícedoménový protein a zajímá nás jen jedna doména, předpokládáme, že tato sekvence bude zkrácena takovým způsobem, aby obsahovala jen doménu, o kterou máme zájem.

Specifikace katalytických reziduí – pozice a typy reziduí pro *známé sekvence*.

Hlavní templát – struktura, na kterou budou přiloženy všechny homologní modely.

Vlastní sekvence – sekvence pro vlastní struktury, které nejsou v PDB databázi, ale bylo by vhodné je zařadit mezi potenciální templáty pro homologní modelování.

Specifikace katal. reziduí ve vlastních sekvencích

Vlastní struktury – PDB a mmCIF soubory.

4.2 Hledání příbuzných enzymů

V této fázi je provedeno základní vyhledávání příbuzných proteinových sekvencí s využitím nástroje PSI-BLAST pro hledání homologních sekvencí. Jako dotaz se použijí všechny sekvence pro vyhledávání a bude se vyhledávat v databázi NR, která obsahuje všechny sekvence proteinů z databází RefSeq[24], Swiss-Prot [25], PDB[21], PIR[26] a PRF spolu s překlady všech genů z databáze GeneBank do proteinových sekvencí. Výstupem je velké množství putativních sekvencí, které jsou v dalších krocích filtrovány a anotovány.

4.3 Zkrácení sekvencí a konstrukce matice vzdáleností

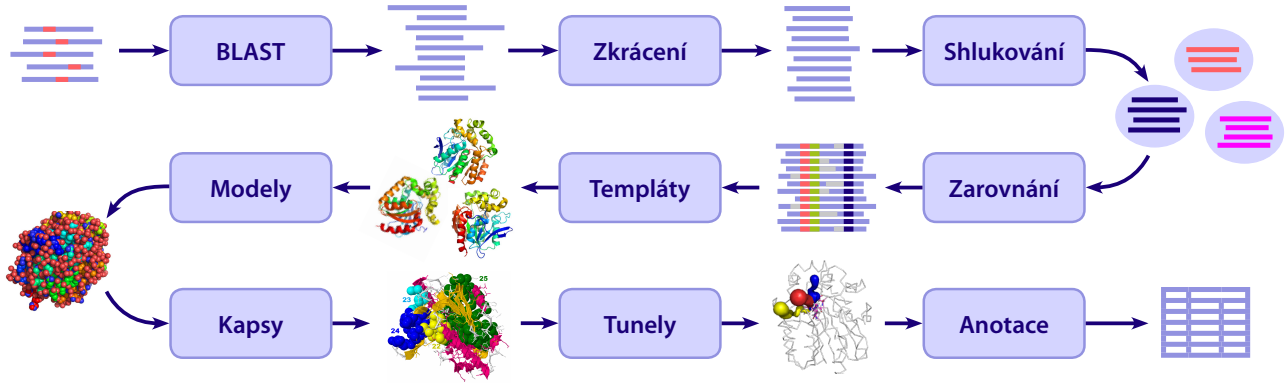
Všechny sekvence identifikované pomocí PSI-BLAST je nutné zkrátit tak, aby délkou zhruba odpovídaly sekvencím pro vyhledávání. Tímto budou z nalezených výsledků odstraněny domény, které se nevyskytují v sekvencích pro vyhledávání. Pokud by odstraněny nebyly, mohly by nastat problémy v dalších fázích analýzy – např. při shlukování. Na závěr této fáze se provede párové přiložení všech zkrácených sekvencí se všemi známými sekvencemi a provede se i vzájemné globální párové přiložení všech zkrácených sekvencí mezi sebou a připraví se matice párových vzdáleností ve formátu vhodném pro UPGMA shlukování.

4.4 Shlukování

Shlukování je nutným krokem před výpočtem vícenásobného zarovnání (MSA) nalezených sekvencí, jinak by nebylo dosaženo kvalitního výsledku. Pomocí programu MC-UPGMA[27] je provedeno hierarchické shlukování sekvencí dle matice párových vzdáleností. Získaný strom shluků je pak ořezán na různých úrovních a je vybráno takové rozdělení, které nejlépe pokrývá množinu identifikovaných sekvencí a ideálně v každém shluku je jedna ze známých sekvencí. Na závěr se vytvoří soubory obsahující sekvence patřící do daných shluků, což v další fázi umožní provést konstrukci profile-profile MSA.

4.5 Konstrukce vícenásobného zarovnání

Pomocí nástroje Clustal Omega[28] se pro každý shluk z předchozí fáze vytvoří MSA. Dále podle hierarchie, která je dána stromem shluků, se postupně provede profile-profile MSA a to tak, že se v každém kroku budou vybírat dva nejbližší shluky. Všechny sekvence, které v hlavičce obsahují některé z klíčových slov *synthetic, vector,* apod., jsou označeny jako umělé. V získaném MSA se podle specifikace katalytických



Obrázek 2. Ilustrace kroků filtrace a anotace. Podrobnější popis naleznete v příslušných kapitolách – BLAST (4.2), zkrácení (4.3), shlukování (4.4), zarovnání (4.5), templáty (4.6), modely, kapsy, tunely (4.7), anotace (4.8).

reziduí určí jejich pozice a pro každou sekvenci se zkontroluje, jestli obsahuje specifikovaná katalytická rezidua. Sekvence, které neobsahují všechna potřebná rezidua (např. z důvodu delece, či substituce) jsou označeny jako nekompletní respektive degenerované a smazány. Tento krok vede k tomu, že z velkého množství sekvencí zůstanou pouze takové, které pravděpodobně budou tvořit protein s mírně modifikovanou funkcí.

4.6 Vyhledávání templátů

Z identifikovaných sekvencí se vyberou takové, pro které existuje záznam v PDB databázi struktur. K těmto sekvencím se přidají všechny vlastní sekvence – tj. sekvence vlastních proteinů s určenou strukturou, které dosud nejsou dostupné v PDB databázi – a vytvoří se vlastní databáze pro BLAST. Ve výsledném vícenásobném zarovnání se stanoví konsenzuální začátek a konec sekvencí se známou strukturou. Např. začátek se stanoví na pozici MSA, kde má aspoň 50 % sekvencí nějaké reziduum, obdobně pro konec. Vícenásobné zarovnání se pak ořeže podle konsenzuálního začátku a konce. Ořezané sekvence budou použity pro konstrukci homologních modelů a tím pádem pro všechny následné strukturní analýzy. Pro každou sekvenci se uloží počet odstraněných reziduí (velikost domény) ze začátku a konce a určí se katalytická rezidua a jejich pozice. Každá ořezaná sekvence je použita jako dotaz pro vyhledávání pomocí BLAST ve vlastní databázi (viz výše). Nakonec se z výsledků vyhledávání určí vhodný templát pro všechny sekvence k následnému homologním modelování.

4.7 Homologní modelování, analýza kapes a tunelů

Všechny potenciální templáty se vhodným způsobem ohodnotí – do ohodnocení templátu se zahrne kvalita jeho párového přiložení s modelovanou sekvencí, správnost pozic a typů katalytických reziduí i rozlišení templátu. Nakonec se vybere templát s nejlepším ohodnocením a použije se ke konstrukci homologního modelu nástrojem Modeller. Každý homologní model se strukturně zarovná nástrojem jCE[29] s hlavní strukturou a na zarovnaném modelu se následně spustí CASTp[30] analýza. Zarovnání s hlavní strukturou je nutné především kvůli tomu, abychom mohli nastavit univerzální parametry pro CASTp analýzu všech homologních modelů. Provede se výpočet a shlukování tunelů ve všech zkonstruovaných a zarovnaných homologních modelech nástrojem CAVER[31]. Homologní modely jsou analyzovány jako jedna trajektorie, což navíc umožní vzájemné porovnání tunelů.

4.8 Databáze Taxonomy, Bioproject a Pfam

V závěrečné fázi se stáhnou data z NCBI databází Taxonomy a Bioproject a pro každou putativní sekvenci se vyextrahuje informace o hostitelském organismu, ze kterého daná sekvence pochází, informace o salinitě, optimální teplotě a rozmezí teplot. Pro každou putativní sekvenci se provede i sekvenční prohledání Pfam databáze a vyextrahuje se informace o identifikovaných doménách.

4.9 Výstupy

Každé putativní sekvenci se určí podrodina, do které patří, podle příslušnosti ke shlukům z fáze shlukování. Putativní sekvence se globálně zarovnají se všemi známými sekvencemi a určí se nejbližší známá sekvence. Pro každou sekvenci se provede predikce solubility (rozpustnosti v roztoku). Výsledkem je tabulka, která obsahuje všechna data získaná během analýzy.

5. Hledání dehalogenáz s modifikovanou funkcí

Z popisu povinných vstupů pro vyhledávání je zřejmé, že algoritmus vyžaduje poměrně velké množství expertních znalostí – ať už to jsou známé sekvence

enzymů, specifikace katalytických reziduí nebo strukturní informace. Proto jsem prezentovaný algoritmus ověřil ve spolupráci s Loschmidtovými laboratořemi na proteinové rodině haloalkan dehalogenáz, na kterou se tyto laboratoře specializují a věnují se jejímu inženýrství. Dehalogenázy jsou mikrobiální enzymy, které katalyzují hydrolytické štěpení vazby mezi halogenem a uhlíkem, jehož výsledkem je primární alkohol a halogenidový iont. Vyznačují se vysokou detoxikační schopností – dokáží například neutralizovat jedovatou a karcinogenní látku trichlorpropan, nebo bojový plyn yperit – proto jejich úspěšné proteinové inženýrství může vést k rozsáhlým aplikacím v oblasti ochrany životního prostředí.

Aplikace vyhledávání na známé dehalogenázy vedla na identifikaci celkem 1 057 putativních dehalogenáz s modifikovanou funkcí. Pokud bych vynechal filtrační kroky a použil jen hledání homologních sekvencí, dostal bych *o řád více* výsledků – 10 274. Filtrace tedy dokáže vyloučit významnou část domnělých homologních sekvencí, přičemž největší počet je odstraněn ve fázi shlukování, další pak při konstrukci vícenásobného zarovnání, kdy se odstraňují umělé a degenerované sekvence. Konkrétní počty odstraněných sekvencí jsou uvedeny v následující tabulce.

Tabulka 1. Vliv filtračních kroků na počet nalezených dehalogenáz.

BLAST	Shlukování	Umělé	Degenerované	$\sum$
10 274	−8 744	−39	−434	1 057

Nicméně je nutné ověřit, zda filtrační kroky spolu s falešnými výskyty neodstraňují i stejné množství užitečných sekvencí. Kdyby tomu tak bylo, stačilo by zúžit parametry homologního vyhledávání, abych obdržel podobný počet výsledků jako výpočetně a expertně náročnou filtrací širšího vyhledávání. Proto jsem provedl čistě homologní vyhledávání se stejnými vstupy, ale nastavil jsem o několik řádů nižší E-value (10^{-70} oproti původní hodnotě 10^{-20}) tak, aby počet nalezených sekvencí po triviálním odfiltrování umělých sekvencí byl přibližně roven 1 057 (viz tabulka 2). Porovnáním výsledků z tohoto experimentu s výsledky předchozího vyšlo najevo, že pokud použijeme širší homologní vyhledávání a aplikujeme navrženou filtraci, obdržíme výsledky s podstatně vyšší variabilitou, než kdybychom použili užší homologní vyhledávání. Filtrované výsledky totiž obsahují více než 20 % jiných putativních sekvencí s nižší mírou podobnosti, než výsledky z triviálního experimentu.

Časová náročnost filtrace a další anotace výsledků je ale enormní. Nejdéle trvá fáze homologního modelování a to výrazně oproti ostatním – průměrně zabere 30 minut procesorového času na jednu putativní dehalogenázu, což je při jejich tisícovém počtu nezvládnutelné v rozumném čase na běžných strojích. Ostatní fáze v součtu nezaberou více než 10 minut na jednu dehalogenázu a proto je celková časová náročnost v řádu stovek hodin.

Tabulka 2. Vliv E-value na počet výsledků BLAST vyhledávání.

E-value	10^{-20}	10^{-30}	10^{-40}	10^{-50}	10^{-60}	10^{-70}
Celkem	10 274	2 151	1 494	1 384	1 188	1 087
Umělé	51	39	39	39	39	39

6. Shrnutí

Ve spolupráci s Loschmidtovými laboratořemi jsem navrhl a implementoval nový nástroj pro vyhledávání příbuzných enzymů s modifikovanou funkcí v biologických databázích, který kombinuje hledání homologních sekvencí s predikcí funkce proteinu. Charakteristiku nástroje jsem ověřil na případu hledání příbuzných dehalogenáz. Ukázalo se, že navržené filtrační kroky vedou k významné redukci počtu výsledků při zachování jejich různorodosti, což bylo cílem a jednou z podmínek užitečnosti takového nástroje.

Vytvořil jsem tedy silný prostředek, který proteinovým inženýrům umožní – v případě, že mají k dispozici všechny potřebné vstupní informace – ve stále rostoucích databázích identifikovat nové zajímavé sekvence enzymů, které mohou být vhodné pro jejich další výzkum.

Zároveň si uvědomuji, že je vhodné, aby nástroj v budoucnu byl veřejně dostupný prostřednictvím webového rozhraní, jak to bývá u bioinformatických řešení zvykem, aby jej mohl bez komplikací využít širší okruh uživatelů. Tomuto ale prozatím brání příliš vysoká výpočetní náročnost filtračních kroků a tím i provozní náklady takové služby. Na druhou stranu se otevírá prostor pro zamyšlení nad vhodným obchodním modelem, který by vyhovoval jak potřebám a možnostem akademických pracovníku, tak soukromým společnostem.

Poděkování

Děkuji za pomoc a odborný přístup celému týmu Loschmidtových laboratoří a obzvláště paní Mgr. Evě Šebestové, Ph.D. za náměty týkající se filtračních kroků. Velice si vážím času, který mi věnovali.

Výpočetní prostředky byly poskytnuty MetaCentrem v rámci programu LM2010005 a organizací CERIT-

SC v rámci programu Centre CERIT Scientific Cloud, který je součástí Operačního programu výzkum a vývoj pro inovace, reg. č. CZ.1.05/3.2.00/08.0144.

Literatura

[1] J. James and B. K. Simpson. Application of enzymes in food processing. *Crit Rev Food Sci Nutr*, 36(5):437–463, May 1996.

[2] O. Kirk, T. V. Borchert, and C. C. Fuglsang. Industrial enzyme applications. *Curr. Opin. Biotechnol.*, 13(4):345–351, Aug 2002.

[3] C. C. Akoh, S. W. Chang, G. C. Lee, and J. F. Shaw. Biocatalysis for the production of industrial products and functional foods from rice and other agricultural produce. *J. Agric. Food Chem.*, 56(22):10445–10451, Nov 2008.

[4] Sylvie Le Borgne and Rodolfo Quintero. Biotechnological processes for the refining of petroleum. *Fuel Processing Technology*, 81(2):155–169, 2003.

[5] M. Ayala, M. A. Pickard, and R. Vazquez-Duhalt. Fungal enzymes for environmental purposes, a molecular biology challenge. *J. Mol. Microbiol. Biotechnol.*, 15(2-3):172–180, 2008.

[6] I Zafir-Lavie, Y Michaeli, and Y Reiter. Novel antibodies as anticancer agents. *Oncogene*, 26(25):3714–3733, 2007.

[7] T. Olafsen and A. M. Wu. Antibody vectors for imaging. *Semin Nucl Med*, 40(3):167–181, May 2010.

[8] Bernd H. A. Rehm. Bacterial polymers. *Nature Reviews Microbiology*, 8(8):578–592, 2010.

[9] S. Banta, I. R. Wheeldon, and M. Blenner. Protein engineering in the development of functional hydrogels. *Annu Rev Biomed Eng*, 12:167–186, Aug 2010.

[10] Mehmet Sarikaya, Candan Tamerler, Alex K. Y. Jen, Klaus Schulten, and François Baneyx. Molecular biomimetics. *Nature Materials*, 2(9):577–585, 2003.

[11] C. Tamerler, D. Khatayevich, M. Gungormus, T. Kacar, E. E. Oren, M. Hnilova, and M. Sarikaya. Molecular biomimetics: GEPI-based biological routes to technology. *Biopolymers*, 94(1):78–94, 2010.

[12] Growth of GenBank and WGS, 2015.

[13] Animesh Sarker, Marufa Nasreen, Md. Rafiad Islam, G M Mahmud Arif Pavel, and Al Amin. Computational approach to search for plant homologues of human heat shock protein. *Journal of Computer Science & Systems Biology*, 6(3):099–105, 2013.

[14] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[15] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J. Mol. Biol.*, 147(1):195–197, Mar 1981.

[16] B. Ma, J. Tromp, and M. Li. Patternhunter: faster and more sensitive homology search. *Bioinformatics*, 18(3):440–445, 2002.

[17] Michael Remmert, Andreas Biegert, Andreas Hauser, and Johannes Söding. HHblits. *Nature Methods*, 9(2):173–175, 2011.

[18] S. R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, 1998.

[19] C. J. Sigrist, L. Cerutti, N. Hulo, A. Gattiker, L. Falquet, M. Pagni, A. Bairoch, and P. Bucher. PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinformatics*, 3(3):265–274, Sep 2002.

[20] R. A. Laskowski, J. D. Watson, and J. M. Thornton. ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.*, 33(Web Server issue):89–93, Jul 2005.

[21] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res.*, 28(1):235–242, Jan 2000.

[22] Jianyi Yang, Renxiang Yan, Ambrish Roy, Dong Xu, Jonathan Poisson, and Yang Zhang. The I-TASSER Suite. *Nature Methods*, 12(1):7–8, 2014.

[23] Matheus P Postigo, Renata Krogh, Marcela F Terni, Humberto M Pereira, Glaucius Oliva, Marcelo S Castilho, and Adriano D Andricopulo. Enzyme kinetics, structural analysis and molecular modeling studies on a series of Schistosoma mansoni PNP inhibitors. *Journal of the Brazilian Chemical Society*, 22(3):583–591, 2011.

[24] K. D. Pruitt and D. R. Maglott. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, 29(1):137–140, Jan 2001.

[25] A. Bairoch and B. Boeckmann. The SWISS-PROT protein sequence data bank: current sta-

tus. *Nucleic Acids Res.*, 22(17):3578–3580, Sep 1994.

[26] W. C. Barker, D. G. George, L. T. Hunt, and J. S. Garavelli. The PIR protein sequence database. *Nucleic Acids Res.*, 19 Suppl:2231–2236, Apr 1991.

[27] Y. Loewenstein, E. Portugaly, M. Fromer, and M. Linial. Efficient algorithms for accurate hierarchical clustering of huge datasets. *Bioinformatics*, 24(13):i41–i49, 2008.

[28] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Soding, J. D. Thompson, and D. G. Higgins. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7(1):539–539, 2011.

[29] I. N. Shindyalov and P. E. Bourne. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, 11(9):739–747, 1998.

[30] T. A. Binkowski, S. Naghibzadeh, and J. Liang. CASTp: Computed Atlas of Surface Topography of proteins. *Nucleic Acids Res.*, 31(13):3352–3355, 2003.

[31] B. Kozlikova, E. Sebestova, V. Sustr, J. Brezovsky, O. Strnad, L. Daniel, D. Bednar, A. Pavelka, M. Manak, M. Bezdeka, P. Benes, M. Kotry, A. Gora, J. Damborsky, and J. Sochor. CAVER Analyst 1.0. *Bioinformatics*, 30(18):2684–2685, 2014.