Masarykova univerzita
Fakulta informatiky

Softwarový nástroj pro automatický návrh mutací a chytrých knihoven pro proteinové inženýrství

Diplomová práce

Bc. Jan Štourač

Brno, jaro 2017

Prohlášení

Prohlašuji, že tato diplomová práce je mým původním autorským dílem, které jsem vypracoval samostatně. Všechny zdroje, prameny a literaturu, které jsem při vypracování používal nebo z nich čerpal, v práci řádně cituji s uvedením úplného odkazu na příslušný zdroj.

Bc. Jan Štourač

Vedoucí práce: Mgr. Jan Brezovský, Ph.D.
Konzultant: RNDr. Martin Maška, Ph.D.

Poděkování

Rád bych poděkoval svému vedoucímu Mgr. Janu Brezovskému, Ph.D. za cenné rady a velkou podporu nejen při sepisování práce, ale především při návrhu nástroje a realizaci bioinformatických analýz. Dále mé velké poděkování patří také Ing. Jaroslavu Bendlovi, Ph.D. za vynikající tandemovou spolupráci ve všech fázích projektu a také psychickou podporu během jeho dokončování. V neposlední řadě bych rád poděkoval RNDr. Martinu Maškovi, Ph.D. za cenné připomínky k obsahu a textové části mé práce.

Shrnutí

HotSpot Wizard je interaktivní webový portál určený pro identifikaci aminokyselinových reziduí (tzv. hot spotů) vhodných pro mutagenezi za účelem zlepšení katalytických vlastností nebo termostability proteinů. Tyto pozice jsou vytipovány na základě kombinace strukturních, funkčních a evolučních informací získaných ze tří bioinformatických databází a 18 výpočetních nástrojů. Nástroj HotSpot Wizard je určen především uživatelům bez větších bioinformatických zkušeností, a proto byl kladen velký důraz na jeho snadné používání i bez nutnosti detailních informací o studovaném systému. Pro zaručení široké použitelnosti a robustnosti byly všechny výchozí hodnoty parametrů zvoleny na základě rozsáhlých bionformatických analýz, což by mělo minimalizovat případy vyžadující jejich modifikaci. Výsledky pro každou proteinovou strukturu je možné analyzovat s využitím čtyř zavedených strategií proteinového inženýrství: (i) identifikace evolučně variabilních reziduí umístěných v relevantních přístupových tunelech či katalytických kapsách s vysokou pravděpodobností ovlivnění katalytických vlastností, (ii) detekce vysoce flexibilních regionů, jejichž rigidifikací je možné zvýšit termostabilitu, (iii) identifikace pozic s dominujícím evolučním výskytem jiné aminokyseliny, jejíž zpětná mutace může protein stabilizovat a (iv) identifikace vzájemně evolučně korelovaných reziduí. Vybrané hot spoty lze dále využít pro návrh chytrých knihoven pro experimentální charakterizaci s podporou automatického výběru vhodných aminokyselin a degenerovaných kodonů. Nástroj je volně dostupný všem nekomerčním uživatelům na adrese `http://loschmidt.chemi.muni.cz/hotspotwizard`.

Klíčová slova

protein, enzym, mutace, proteinové inženýrství, hot spot, HotSpot Wizard, chytrá knihovna, výpočetní mutageneze, semi-racionální design

Obsah

1 Úvod

Proteiny neboli bílkoviny jsou jedny z nejdůležitějších makromolekul ve všech živých organismech. Slouží totiž nejen jako základní stavební kameny buněk, ale podílí se též na téměř všech biologických procesech, jako je katalýza reakcí nebo transport malých molekul. Proto již bylo proteinům věnováno za posledních několik desetiletí velké množství času za účelem pochopit jejich biologickou funkci a vysvětlit tak jejich význam pro život organismů. Bohužel přes významné pokroky v oblasti experimentálních technik a dostupného vybavení není stále možné je přímo sledovat a studovat jejich konkrétní chování a je tedy nutné hledat alternativní cesty pro jejich pochopení. Jednou z těchto cest se vydalo také proteinové inženýrství, jehož snahou je návrh a konstrukce mutantních proteinů s upravenými vlastnostmi. Díky tomu je tak možné ověřit nebo vyvrátit mnoho hypotéz o jejich mechanismu, ale také konstruovat vylepšené proteiny. Zvláště druhý cíl je pak výrazně katalyzován zájmem průmyslu, neboť proteiny jsou dnes již běžnou součástí velkého množství technologických procesů, které ale často limitují svojí nízkou teplotní stabilitou nebo katalytickou aktivitou.

Ačkoliv samotná konstrukce upravených proteinů je již triviální, stále nevyřešeným problémem je návrh optimální sady mutací, které upraví požadovanou vlastnost. I když je již mnoho let známé, že proteiny a všechny jejich vlastnosti jsou přímo určeny jejich jednodimenzionálním zápisem pomocí aminokyselinové sekvence, stále nikdo nedokázal vytvořit metodu, která by to spolehlivě dokázala. Proto vzniklo velké množství nástrojů a strategií, které na základě náhody nebo dostupných informací pomáhají proteinovým inženýrům s výběrem vhodných kandidátů. Jedním z nich je i HotSpot Wizard, který na základě vhodné kombinace výpočetních nástrojů vybírá rezidua vhodná pro mutování (tzv. hot spoty). Tato diplomová práce se zabývá návrhem a vývojem jeho druhé verze.

Práce je strukturovaná do osmi kapitol. První obsahuje základní úvod do problematiky. Přímo na ni navazuje druhá a třetí kapitola popisující základy proteinů a jejich struktury, a také proteinové inženýrství. Čtvrtá kapitola je již věnována návrhu nástroje HotSpot Wizard a jednotlivých kroků jeho výpočtu na základě kladených požadavků na uživatelskou přívětivost a vstupních a výstupních dat. V páté kapitole jsou pak popsány bioinformatické analýzy, které bylo nutné provést za účelem zvýšení robustnosti a použitelnosti nástroje. Šestá kapitola je věnována implementačním detailům výpočetního jádra a představení grafického uživatelského rozhraní. Sedmá kapitola obsahuje validaci hot spotů nalezených jednotlivými inženýrskými strategiemi vůči dostupným experimentálním datům. V osmé kapitole je práce shrnuta a naznačen plánovaný vývoj do následující verze. Práce dále obsahuje dvě přílohy – seznam elektronických příloh a také plné znění odborného článku věnovaného nástroji HotSpot Wizard.

2 Proteiny

Proteiny jsou makromolekuly vyskytující se ve všech živých organismech. Z biochemického hlediska se jedná o vysokomolekulární přírodní biopolymery složené z aminokyselin vzájemně spojených peptidickou vazbou do polypeptidového řetězce. Proteiny zajišťují mnoho důležitých funkcí a účastní se téměř všech biochemických procesů v buňkách. Slouží totiž nejenom jako stavební kameny buněk (kolagen, elastin, keratin), ale také zajišťují transport a skladování malých molekul (hemoglobin, transferin), pohyb buňky (aktin, myosin), katalýzu reakcí (enzymy), řízení a regulaci (hormony, receptory) nebo zvyšují obranyschopnost organismu (imunoglobulin, fibrin, fibrinogen) [1].

2.1 Struktura proteinů

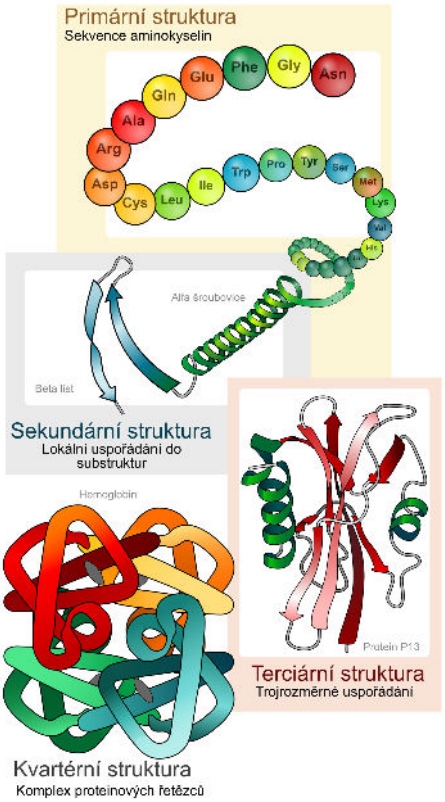
V přírodě se všechny proteiny vyskytují v jedné z možných prostorových konformací, díky které mohou plnit svůj účel. Stavební proteiny tak nejčastěji nabývají protáhlých vláknitých tvarů, které jsou nerozpustné ve vodě. Naproti tomu enzymy nebo hormony tvoří většinou kulovitou (tzv. globulární) konformaci, která jim umožňuje vhodně zformovat aktivní místo pro katalýzu reakcí, ale lze je snadno poškodit působením vyšší teploty nebo změnou pH [2].

Strukturu proteinů pak lze lépe charakterizovat s využitím čtyř úrovní strukturních elementů (obrázek 2.1):

- **Primární struktura** – jedná se o jednorozměrnou reprezentaci proteinu pouze pomocí pořadí aminokyselin.
- **Sekundární struktura** – sekundární strukturou se rozumí lokální geometrické uspořádání polypeptidového řetězce do několika dobře popsaných stavebních motivů – alfa šroubovice, beta skládaného listu a smyčky.
- **Terciární struktura** – označuje prostorové uspořádání celého polypeptidového řetězce do funkční podoby.
- **Kvartérní struktura** – struktura popisující vzájemné prostorové uspořádání několika polypeptidických řetězců. Vyskytuje se pouze u omezené množiny známých proteinů.

2.2 Enzymy

Jednou z nejvýznamnějších skupin proteinů jsou enzymy. Ty svojí schopností katalýzy chemických reakcí tvoří přírodní mechanismus pro konstrukci a dekonstrukci široké škály důležitých malých molekul a de facto tak řídí většinu biochemických procesů v živých organismech. Ze strukturního hlediska tvoří enzymy globulární tvary a jejich společným znakem je přítomnost alespoň jednoho aktivního místa. Aktivní místo si lze zjednodušeně představit jako vhodné uspořádání vybraných aminokyselin proteinu do tvaru formujícího dutinu, v níž následně probíhá reakce. Princip funkce enzymů je podobný ostatním katalyzátorům a jeho podstatou je snižování nutné aktivační (Gibbsovy) energie reakce, díky čemuž se provedení reakce značně zjednodušuje. Snížení této energie lze docílit několika různými mechanismy. Prvním z nich je vhodné přiblížení a změna prostorové orientace substrátů. Díky tomu vznikne



Obrázek 2.1: Vizuální reprezentace jednotlivých úrovní struktur (převzato z [3]).

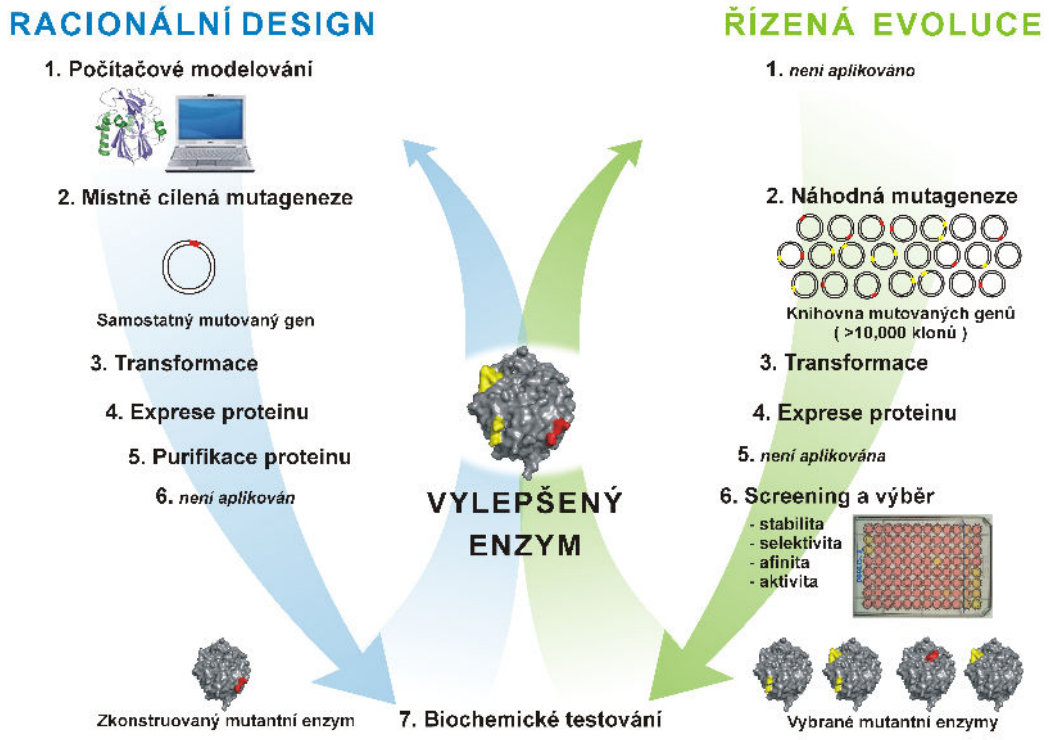
vhodné lokální uspořádání a reakce tak probíhá snáze. V jiném případě vytváří aktivní místo vlivem fyzikálně-chemických vlastností okolních aminokyselin vhodné prostředí (např. kyselé nebo zásadité) nutné pro proběhnutí reakce. Dalším častým mechanismem je navázání malé molekuly na enzym, díky čemuž dojde k natažení a tedy i oslabení vazeb mezi jeho atomy, které jsou pak snadněji přerušeny. Samotné aktivní místo se pak může nacházet jak na povrchu, tak i zanořené uvnitř struktury [2].

3 Proteinové inženýrství

Jak již bylo zmíněno, proteinové inženýrství je mladá vědní oblast zabývající se studiem proteinů, jejich funkce a především návrhem a konstrukcí jejich mutantních variant s upravenými vlastnostmi. Tyto upravené proteiny nacházejí své uplatnění nejenom v dalším vědeckém zkoumání, ale často také v rozličných biotechnologiích. Metody proteinového inženýrství jsou ohraničené dvěma základními přístupy podle experimentální a výpočetní náročnosti (obrázek 3.1) [4].

3.1 Racionální návrh

První a starší metodou proteinového inženýrství je racionální návrh. Jeho vznik se datuje do devadesátých let a je založen na masivním využívání výpočetních zdrojů, matematických modelů a dat z databází. Své uplatnění tak zde nachází různé vizualizační techniky, metody studování geometrických vlastností struktury, analýzy evolučního signálu, predikce škodlivosti mutací, výpočty vazebných a interakčních energií ligandů, simulace interakcí jednotlivých atomů a mnoho dalších. Data z jednotlivých analýz jsou následně zkombinovány a na jejich základě je vybráno několik vhodných mutací. Ty jsou následně zaneseny do sekvence s využitím místně cílené mutageneze [6]. Díky tomu je možné výrazně zredukovat experimentální náročnost návrhu až na jednotky kandidátů se zachováním vysoké úspěšnosti a výrazně tak zredukovat nutné množství času a finančních prostředků. Na druhé straně je tento přístup extrémně závislý



Obrázek 3.1: Obvyklý pracovní postup obou metod proteinového inženýrství (zdroj: [5]).

na dostupnosti kvalitních dat o studovaném systému, jeho struktuře a funkčním mechanismu. V neposlední řadě je také velmi citlivý na správné nastavení výpočetních nástrojů a limitován jejich omezenou přesností.

3.2 Řízená evoluce

O něco mladší je pak cílená řízená evoluce. Na rozdíl od racionálního designu se jedná o čistě experimentální metodu, která se svým principem snaží uměle napodobit evoluční mechanismus selekce a využít ho pro konstrukci upravených proteinů se zlepšenými konkrétními vlastnostmi. Jejím základem je konstrukce knihoven s využitím náhodné mutageneze, jejich experimentální charakterizace a výběr kandidátních sekvencí. Poté se celá konstrukce opakuje s využitím kandidátních sekvencí tak dlouho, dokud není získán protein s požadovanými vlastnostmi. Z principu metody je jasné, že její největší výhodou jsou nízké vstupní znalosti. Pro její provedení tak není nutné mít k dispozici ani žádné detailní informace a ani žádné výpočetní nástroje. Nevýhoda pak pochopitelně tkví v nízké úspěšnosti, kdy je obvykle nutné provést charakterizaci velkého množství kandidátů v několika po sobě jdoucích kolech. S tím se samozřejmě pojí také vysoká časová a hlavně finanční náročnost celého procesu [7].

4 HotSpot Wizard

Na pomezí mezi racionálním designem a řízenou evolucí se nacházejí tzv. hybridní metody, jejichž cílem je vhodnou kombinací obou metod snížit jejich nevýhody a zároveň těžit z jejich výhod. Do této skupiny patří také HotSpot Wizard. Jeho hlavní myšlenkou je snížení experimentální náročnosti řízené evoluce konstrukcí „chytrých“ knihoven. Ty jsou založeny na racionálně vybraných reziduích (tzv. hot spotech), která jsou náhodně nahrazena aminokyselinami z často také racionálně vybrané podskupiny, čímž je maximalizována pravděpodobnost ovlivnění cílené funkce, a díky tomu mohou být knihovny výrazně menší než při náhodném mutování.

Tím pádem je problém s experimentální náročností nahrazen problémem správného výběru hot spotů. Za dobu existence proteinového inženýrství byla pochopitelně navržena a používána široká škála strategií, které se od sebe odlišují nejenom cílenou skupinou vlastností proteinů a úhlem pohledu, z nějž ho studují, ale často také množstvím vyžadovaných vstupních dat a počtem využívaných nástrojů. Bohužel však žádný z těchto přístupů není jednoznačně nejlepší, a proto je v reálných studiích často nutné kombinovat výsledky z více různých strategií. Dalším limitujícím faktorem bývá i to, že mnohé z nich jsou dostupné pouze jako metodiky bez automatizovaných implementací, a je tedy nutné manuálně propojit výsledky i desítek různých nástrojů, všechny vhodně nastavit a v neposlední řadě správně interpretovat jejich výsledky. Tato komplikace se pak často stává nepřekonatelnou bariérou pro mnoho proteinových inženýrů, kteří nemají v týmu zkušeného bioinformatika nebo sami takovými znalostmi nedisponují. Proto vzniká přirozená poptávka po nástrojích nebo meta-nástrojích, které poskytují ucelený a snadný způsob identifikace těchto hot spotů. S touto motivací byl vyvinut i nástroj HotSpot Wizard. Oproti první verzi [8] orientující se pouze na identifikaci funkčních hot spotů, byl v rámci této diplomové práce výrazně rozšířen integrací dalších tří inženýrských strategií a rozšířením záběru i na teplotní stabilitu. Stal se tak v současné době jedinou platformou zahrnující více populárních strategií pro identifikaci hot spotů. Při jeho návrhu byl kladen důraz především jeho přívětivost vůči méně zkušeným uživatelům, kdy díky minimalizaci požadovaných vstupů a intenzivní optimalizaci výchozích nastavení jednotlivých kroků, jej lze využít téměř bez jakékoliv předchozí znalosti studovaného systému. Dalším klíčovým požadavkem byla možnost výsledky snadno zobrazit a detailněji analyzovat přímo v rámci nástroje, bez nutnosti zapojení dalších externích nástrojů.

4.1 Cílené vlastnosti

Aplikacemi vhodných strategií proteinového inženýrství je samozřejmě možné modifikovat širokou škálu vlastností enzymů. Pro nástroj HotSpot Wizard byly z této množiny vybrány dvě velmi často cílené skupiny vlastností.

Katalytické vlastnosti

Katalytické vlastnosti enzymů jsou klíčem k jejich biologické funkci. Jejich úpravami tak je možné výrazným způsobem zasáhnout do reakčního mechanismu celého enzymu a to nejen po stránce rychlosti a efektivity, ale též v množství akceptovaných malých molekul (substrátů) a jejich prostorového uspořádání. V krajním případě je dokonce možné celý reakční

mechanismus kompletně změnit. Hot spoty pro tento typ vlastností se nejčastěji nacházejí v blízkém okolí aktivního místa a transportních tunelů a bývají spojené s mechanismem transportu a vázání substrátu, stabilizace tranzitního stavu nebo uvolňováním produktu. Pro jejich detekci se obvykle používají nástroje určené pro analýzu interakcí enzymu se substrátem [9]–[11] nebo nástroje pro analýzu aktivního místa, katalytické kapsy a transportních tunelů [12]–[14]. HotSpot Wizard se v současné době zaměřuje především na druhé jmenované.

Stabilita

Další klíčovou vlastností enzymů je jejich stabilita, kterou se rozumí odolnost struktury enzymů vůči zhoršeným podmínkám, jako je například vyšší teplota nebo přítomnost organických rozpouštědel. Při překročení jejich limitu totiž ve většině případů dochází k postupnému rozpadu struktury a tudíž i ke snížení nebo úplné ztrátě funkce [15]. Stabilnější varianty enzymů pak bývají často žádané především v průmyslu nebo zdravotnictví, kde je většina procesů navržena pro vyšší teploty. Podobně jako pro katalytické vlastnosti i pro stabilitu existuje nespočet různých strategií jako například omezení pohyblivosti velmi mobilních regionů (rigidifikace), zaplňování kavit, modifikace přístupových tunelů, konsenzuální a ancestrální rekonstrukce nebo úpravy povrchových nábojů [16]–[18]. Ačkoliv jsou pro enzymy důležité oba typy stabilit, v rámci nástroje HotSpot Wizard je aktuálně adresována pouze stabilita teplotní (termostabilita) a to s využitím rigidifikace flexibilních regionů a konsenzuální rekonstrukce.

4.2 Vstupní data

Jediným povinným vstupem nástroje je pouze proteinová struktura. Ostatní vstupní data jsou zjišťována buď automaticky, nebo pro ně byly na základě rozsáhlých bioinformatických analýz vybrány takové výchozí hodnoty, které by měly být co nejuniverzálnější a které by většina uživatelů neměla mít potřebu měnit.

Proteinová struktura

Proteinovou strukturu je možné zadat dvěma různými způsoby. Prvním z nich je unikátní identifikátor v celosvětové databázi struktur Protein Data Bank [19] a druhým pak nahrání vlastního souboru ve formátu PDB [20]. Tento formát byl vybrán s ohledem na svou popularitu a širokou podporu komunity a dalších nástrojů. K mnoha strukturám (především z databáze) jsou kromě koordinát atomů také k dispozici další metadata popisující protein, jeho strukturní charakteristiky a další vlastnosti. Tato metadata je pak možné dále zpracovávat a využít je pro zkvalitnění analýz. V rámci nástroje HotSpot Wizard jsou pak dále zpracovávány dvě z nich.

Symetrie. S ohledem na princip rentgenové krystalografie je mnoho proteinů v databázi deponovaných ve strukturním uspořádání nazvaném asymetrická jednotka. Tato kvartérní struktura však reprezentuje prostorové uspořádání podjednotek proteinu v experimentálním krystalu a tedy nikoliv jeho biologicky funkční variantu, jejíž znalost je naprosto klíčová pro relevanci prováděných strukturních analýz [21]. Z tohoto důvodu jsou do metadat struktury také ukládány matice geometrických transformací, jejichž aplikací na atomy lze získat biologicky

relevantní konformaci [22]. HotSpot Wizard existenci této informace automaticky detekuje a s využitím nástroje MakeMultimer [23] provede automatickou konverzi.

Sekvence. Dalším problémem experimentálního měření struktur jsou časté nepřesnosti, díky kterým pak ve výsledné struktuře mohou chybět informace o poloze některých atomů nebo i celých reziduí (především na začátku a na konci řetězců) [24]. Ačkoliv je tento problém významný (a bohužel ne snadno řešitelný) především pro strukturní analýzy, při rekonstrukci sekvence proteinu ze struktury by mohl způsobovat nežádoucí šum i v analýzách sekvenčních. Naštěstí existují i jiné metody určování sekvence a pro většinu proteinů je tudíž známá daleko před strukturou. Pro úplnost je proto také často uvedena v metadatech struktury [24] a je tedy možné ji snadno extrahovat a využít pro sekvenční větev analýz, které tak budou robustnější a lépe odstíněné od experimentálního šumu.

Esenciální rezidua

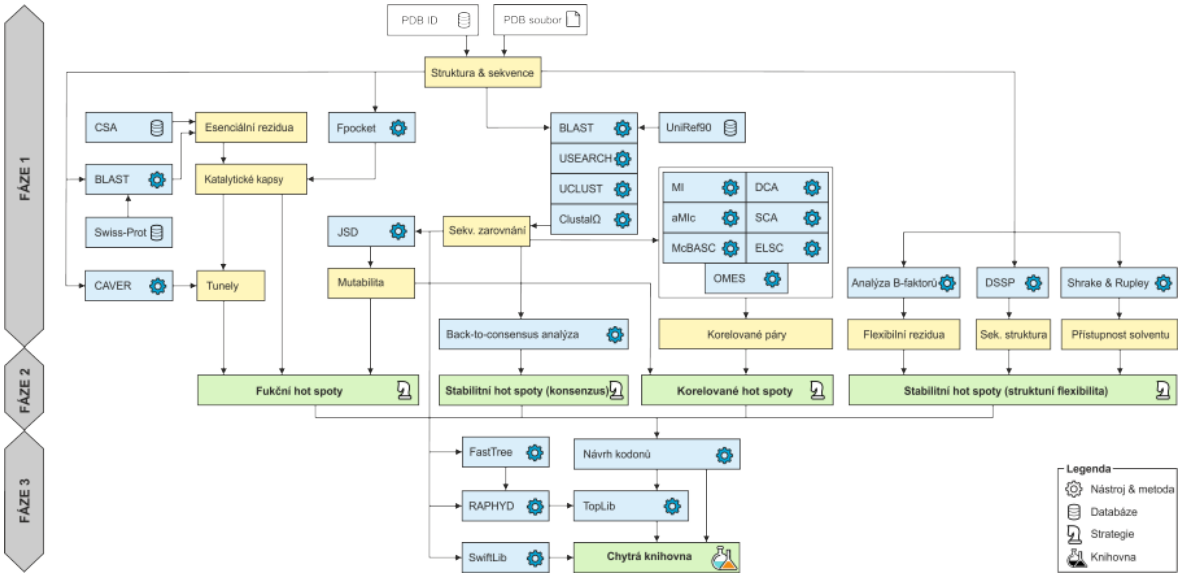
Dalším klíčovým vstupem jsou esenciální rezidua. Ty se většinou nachází v aktivním místě enzymu a přímo se podílí na jeho funkci [1]. Jejich znalost je významná především proto, že jejich neuváženou mutací je velmi snadné narušit nebo kompletně zničit funkci celého enzymu, a proto je silně doporučené tato rezidua neoznačovat jako hot spoty. Aby nebylo nutné tyto rezidua dopředu znát nebo si je složitě dohledávat z literatury, bylo do nástroje zaintegrováno jejich automatické získávání z databází Catalytic Site Atlas 2.0 [25] a SwissProt [26]. Všechna nalezená rezidua je pak samozřejmě možné ručně upravovat – odstraňovat nebo přidávat nová. Rezidua z databáze Catalytic Site Atlas jsou dostupná pouze pro proteiny se známým PDB kódem (tedy nikoliv pro uživatelské struktury), zatímco v případě databáze SwissProt probíhá jejich odvození na základě podobných sekvencí. Aby byla zajištěna jejich kvalita, bylo do nástroje HotSpot Wizard integrováno několik mechanismů, které odstraňují rezidua odvozená na základě nízké sekvenční podobnosti nebo ty, na jejichž zarovnaných pozicích se vyskytuje jiná aminokyselina.

Parametry výpočtu

Posledním volitelným vstupem jsou parametry jednotlivých kroků analýz. S ohledem na pečlivou optimalizaci výchozích hodnot se pak jejich upravování doporučuje především pokročilejším uživatelům.

4.3 Výpočet

HotSpot Wizard se na enzym dívá ze dvou základních úhlů. První z nich se na enzym dívá ze strany sekvence a její evoluce, která popisuje samovolný a dlouhodobý proces jejího vývoje a diverzifikace. Tento proces je realizován zanášením náhodných mutací do genomu, které se následně přeloží do proteinů a které tak mohou mít jiné vlastnosti. Ačkoliv většina těchto změn nic výrazně nemění, některé z nich dokáží protein výrazně zlepšit nebo naopak poškodit. Jelikož je přirozeně vyvíjen tlak na životaschopnost organismu a přežití těch nejlepších, jsou pochopitelně zachovány především lepší či neutrální varianty, zatímco ty poškozené rychle mizí. Díky tomu nám může pohled do historického vývoje a příbuzenstva studovaného enzymu poskytnout velmi cenné informace. Druhý úhel se zaměřuje na analýzu enzymu z pohledu jeho



Obrázek 4.1: Workflow nástroje HotSpot Wizard.

struktury a vzájemných interakcí. Tím se do značné míry obchází neschopnost predikovat jeho funkci a vlastnosti pouze na základě sekvence. Vhodnými algoritmy tak lze zkoumat jeho tvar, povrch, kapsy, vnitřní dutiny a transportní tunely. Energetickými výpočty pak lze studovat jeho interakce s malými molekulami, okolními proteiny nebo interakce samotných atomů.

Níže jsou popsány jednotlivé kroky výpočty, na jejichž základě se následně identifikují vhodné hot spoty.

Predikce sekundárních struktur

Jednou ze základních a zároveň triviálních analýz je přiřazení sekundárních strukturních motivů jednotlivým úsekům proteinu. Tato analýza probíhá pouze na základě sekvence s využitím populárního algoritmu DSSP [27], který k jednotlivým úsekům sekvence asociuje jednu z osmi podporovaných typů sekundárních struktur.

Analýza flexibility

Další rychlou analýzou je identifikace flexibilních regionů v rámci proteinové struktury. Pro nalezení takových regionů se využívá informace o teplotních faktorech (též B-faktorech), které lze získat během rentgenové krystalografie a které vyjadřují přirozené vibrace atomů jako faktor rozptylu naměřené elektronové hustoty [28]. Kromě flexibility se tyto faktory používají také pro odhalení chyb v rekonstrukci 3D koordinátů atomů struktury [29].

Analýza relativní polohy reziduí

Poslední z triviálních analýz je analýza relativní polohy reziduí v rámci struktury. Jejím prvním krokem je rekonstrukce povrchu proteinu dostupného rozpouštědlu (accessible surface area) a to s využitím algoritmu Shrake-Rupley [30]. Na základě dopočítaného povrchu se následně určí relativní plocha každého rezidua, která může přijít do kontaktu s rozpouštědlem. Tyto

hodnoty jsou následně převedeny na binární hodnotu vyjadřující, jestli se dané reziduum vyskytuje na povrchu proteinu nebo ne [31].

Konstrukce mnohonásobného sekvenčního zarovnání

Prvním krokem sekvenční větve je konstrukce mnohonásobného sekvenčního zarovnání (MSA). To si lze v případě proteinů představit jako matici, kde sloupce reprezentují jednotlivá rezidua jdoucí za sebou a řádky potom příbuzné sekvence. Všechny sekvence jsou navíc zarovnány tak, aby se ve sloupcích maximalizovala pravděpodobnost, že jde o stejné pozice [32]. Kvalita zarovnání je přitom naprosto zásadní, neboť na jeho základě jsou prováděny všechny další sekvenční analýzy. Prvním krokem konstrukce je extrakce příbuzných sekvencí z databáze. Jako databáze sekvencí byla vybrána UniProt Reference Clusters (UniRef) [33], která obsahuje sekvence z databází UniProt a UniParc. Tato databáze byla vybrána na úkor standardní nr (non-redundant) databáze kvůli tomu, že je k dispozici její varianta UniRef90 [33], která je předshlukovaná nástrojem CD-HIT [34] na 90% sekvenční identitu. Díky tomuto předshlukování již dopředu dojde k odfiltrování příliš podobných sekvencí, které většinou nepřináší přidanou hodnotu a způsobují nechtěný šum v evoluční informaci. Zároveň je ale zachováno dostatečné pokrytí sekvenčního prostoru. Dalším důležitým kritériem je její velikost, která je v případě této varianty zhruba 2,5krát menší, a tudíž je její prohledávání daleko rychlejší. Samotné vyhledání podobných sekvencí zajišťuje program blastp z balíku NCBI-BLAST+ [35], který je v této oblasti zlatým standardem. Druhým krokem je filtrace sekvencí s příliš nízkou globální identitou. Tento krok byl zařazen z důvodu, že nástroj BLAST kvůli rychlosti výpočtu používá pro srovnání sekvencí pouze lokální podobnost, v níž se hodnotí pouze vysoce podobné části sekvence. To může být problematické především v situaci, kdy se v sekvenci proteinu nachází některý z často se vyskytujících motivů nebo domén. Poté totiž snadno může dojít k zahrnutí i velmi vzdálených proteinů se stejným motivem do vybraných sekvencí. Odstraněním sekvencí s příliš nízkou globální podobností se tento problém efektivně řeší. Pro tento účel byl vybrán nástroj USEARCH [36]. Dalším filtračním krokem je odstranění příliš podobných sekvencí. Ačkoliv je databáze sekvencí předshlukovaná, přesto se občas stává, že se v ní nachází sekvence se vzájemnou podobností vyšší než 90 %. Jejich zahrnutí do kandidátních sekvencí pak opět způsobuje zbytečný šum, a proto je provedeno ještě jedno kolo filtrování nástrojem USEARCH a tyto sekvence jsou odstraněny. Ze zbývajících sekvencí je následně vybrán definovaný počet podle pořadí určeného nástrojem blastp a tyto sekvence jsou zarovnány nástrojem ClustalO [37]. Na úplný závěr je provedena rekonstrukce fylogenetického stromu nástrojem FastTree [38].

Konzervovanost pozic

Analýza konzervovanosti vychází z evolučního předpokladu, že při samovolném mutování reziduí důležitých pro funkci nebo stabilitu nebude protein korektně fungovat a dojde k úhynu daného jedince. Díky tomu se tak na této pozici v průběhu vyskytuje většinou pouze omezené množství variant. Tuto informaci lze snadno statisticky analyzovat a odhadnout, nakolik je konkrétní pozice významná pro správnou funkci nebo stabilitu proteinu, a tak se vyhnout neopatrnému zanášení mutací na tyto pozice [39]. Tato analýza je prováděna na základě mnohonásobného sekvenčního zarovnání s využitím algoritmu Jensen-Shannon divergence (JSD) [40]. Ačkoliv se jedná o obecnou metodu, na bioinformatických datech dosahuje velmi

Algoritmus 4.1 Algoritmus pro konverzi skóre JSD na stupeň mutability.

Vstup: pole obsahující JSD skóre pro jednotlivé sloupce MSA
Výstup: pole obsahující stupeň mutability (9 – nejvyšší, 1 – nejnižší) pro jednotlivé sloupce MSA

```
1: procedure ConvertJSDScore(scores)
2:     intervals ← array(3)
3:     intervals[0] ← (max(scores) − mean(scores) − stddev(scores))/3
4:     intervals[1] ← (stddev(scores) ∗ 2)/3
5:     intervals[2] ← (mean(scores) − min(scores) − stddev(scores))/3
6:     grades ← array(length(scores))
7:     col ← 0
8:     while col < length(scores) do
9:         threshold ← min(scores)
10:        i ← 0
11:        while i ≤ 8 do
12:            threshold ← threshold + interval[floor(i/3)]
13:            i ← i + 1
14:            if scores[col] < threshold then
15:                break
16:            end if
17:        end while
18:        grades[col] ← 9 − i + 1
19:        col ← col + 1
20:    end while
21:    return grades
22: end procedure
```

dobrých výsledků. Jelikož je skóre vypočítané algoritmem pouze abstraktní metrikou a jeho absolutní hodnoty a rozpětí se liší protein od proteinu, byl na základě analýzy jeho distribuce navržen algoritmus převádějící ho na celočíselnou škálu od 1 do 9 (algoritmus 4.1) převzatou z nástroje Rate4Site [41]. Tato škála je v komunitě populární a tudíž pro mnoho uživatelů dobře uchopitelná.

Konsenzuální mutace

Další analýzou vycházející z mnohonásobného sekvenčního zarovnání je analýza konsenzuálních mutací (tzv. back-to-consensus). Podobně jako konzervovanost vychází z principů evoluce, ale z jiného úhlu. Ačkoliv je mutování konzervovaných pozic obecně rizikové, v přírodě se přesto děje. Pokud nyní odmyslíme plně nefunkční mutanty, může mít daná mutace na výsledný protein pouze mírně negativní, neutrální nebo dokonce i pozitivní vliv. A právě na mírně negativní varianty cílí tato analýza. Na základě evoluční informace se identifikují pozice, které jsou obecně konzervované a ve zkoumaném proteinu se na nich vyskytuje jiná aminokyselina než ty nejčastěji zastoupené v zarovnání. Právě jejich zpětná záměna za konzervované varianty může přinést kýžené zlepšení vlastností [42]. Pro tuto analýzu jsou aktuálně používána dvě rozdílná kritéria. Prvním z nich je prostá dominance, kdy je za konsenzuální aminokyselinu

považována ta, která se na dané pozici vyskytuje nejčastěji. Jelikož je tento přístup příliš zjednodušující a necitlivý k pozicím, na nichž se může vyskytovat více různých aminokyselin, bylo přidáno též druhé kritérium, které je poměrové. To znamená, že kromě absolutního výskytu je též měřeno, zda se vyskytuje alespoň n-krát častěji než aktuální aminokyselina. Díky tomuto filtru je možné snížit hranici absolutního výskytu.

Predikce škodlivosti mutací

Další sekvenční analýzou, která však necílí na identifikaci hot spot reziduí, je predikce škodlivosti mutací. Shodně s ostatními vychází z evolučního signálu mnohonásobného zarovnání, ale tentokrát na jeho základě a také na vybraných fyzikálně-chemických parametrech vyskytujících se aminokyselin dokáže predikovat pravděpodobnost, s jakou by záměna aktuální aminokyseliny za jinou ovlivnila funkci tohoto enzymu [43]. Tato informace je cenná především při návrhu knihoven, kdy se lze efektivně vyvarovat nevhodným záměnám a ušetřit tak experimentální náročnost. Ačkoliv nástrojů na tento účel existuje opět velké množství, pro nástroj HotSpot Wizard byl nově vyvinut RAPHYD [44], který je vylepšenou verzí nástroje MAPP [45].

Korelované páry

Poslední ze sekvenčních analýz je detekce korelovaných párů. Vychází z myšlenky interakčních sítí v proteinech, které musí být pro zachování správných interakcí mutované buď společně, nebo alespoň s ohledem na ostatní. I když mohou být interakční sítě rozsáhlé, jejich přesná detekce je velmi obtížná a neexistuje dostatečně spolehlivá metoda. Lepších výsledků dosahují nástroje, pokud se omezíme pouze na dvojice, na něž se zaměřuje také HotSpot Wizard [46]. Pro jejich detekci existuje velké množství statistických metod a nástrojů. Z jejich srovnání ale vyplynulo, že žádný z nich svojí přesností nedominuje, a proto byl pro účel nástroje HotSpot Wizard sestrojen metaprediktor kombinující výsledky sedmi nejběžněji používaný metod – DCA [47], ELSC [48], McBASC [49], MI [50], aMIc [51], OMES [52] a SCA [53].

Detekce katalytických kapes

Trojrozměrné struktury proteinů se často vyznačují svým nepravidelným tvarem a povrchem. Díky tomu je možné na jejich povrchu detekovat velké množství různě velkých a tvarovaných prohlubní (tzv. kapes). Obdobně komplexní bývá volný prostor uvnitř proteinové struktury, kde často vzniká spleť různých dutinek. Většina z dutin a kapes je pouze geometrickým reliktem bez většího významu, ale zvláště u enzymů mohou některé z nich hrát důležitou roli. Tyto kapsy a dutiny se nazývají katalytické a svým vhodným tvarem a umístěním aminokyselin formují aktivní místo a vytvářejí tak vhodné podmínky pro průběh katalyzované chemické reakce [1], [54]. Tyto prostory, jejich vlastnosti a zvláště blízké aminokyseliny jsou vynikajícím zdrojem informací pro detekci hot spotů [55], [56]. Detekce kapes a dutin je většinou realizována s využitím geometrických algoritmů. Dřívější z nich byly většinou založeny na principu mřížek, které ale byly postupem času vytlačeny efektivnější reprezentací pomocí Voroného diagramu a odvozenými alfa tvary. Pro nástroj HotSpot Wizard byl vybrán nástroj Fpocket [57], který je na tomto principu založený a kromě detekce dokáže každé kapse přiřadit i odhad tzv. „druggability“, který reprezentuje její vhodnost k navázání substrátu [58].

Jelikož proteiny většinou obsahují jednotky až desítky nezajímavých kapes, je nutné z nich správně vybrat ty katalytické. Jejich určování v rámci nástroje HotSpot Wizard probíhá ve třech krocích. V prvním z nich je pro každý řetězec provedeno přiřazení esenciálních reziduí ke kapsám a následně je jako katalytická kapsa označena ta s jejich největším počtem. Záměrně je zmíněn největší počet, neboť se nezřídka stává, že kapsa neobsahuje všechna esenciální rezidua. To může být způsobeno nejenom nepřesností geometrických algoritmů, ale též biologickým významem, neboť některá stabilizační rezidua mohou být umístěny až v přístupových tunelech a tudíž ve větší vzdálenosti od aktivního místa. Pokud je nalezeno více kapes se stejným počtem reziduí, vybere se ta s vyšším skóre „druggability“, které relativně dobře koreluje s významem kapsy (viz kapitola 5.2). V případě, že v prvním kroku není nalezena žádná katalytická kapsa, je pro celou strukturu vybrána ta s nejvyšším skóre „druggability“. Tento výběr již není pro každý řetězec, neboť bez znalosti esenciálních reziduí (a tudíž informací o funkčním významu řetězce) je tento výběr zbytečně rizikový. V posledním kroku je ještě provedeno přiřazení kapes na základě homologie řetězců. Pokud totiž mají dva či více řetězců úplně shodné sekvence, mají také i shodnou biologickou funkci. Proto je celkem bezpečné provést mapování, když na základě reziduí v katalytické kapse původního řetězce je v dalších shodných řetězcích nalezena taková kapsa, jejíž průnik reziduí je co největší. Opět se především kvůli možným nepřesnostem v geometrii bere kapsa pouze s největším a nikoliv s úplným průnikem.

Identifikace tunelů

Jelikož mnoho enzymů má své aktivní místo zanořené hlouběji ve struktuře, musí existovat způsob transportu substrátů a dalších malých molekul dovnitř a ven. K tomuto účelu slouží tunely, které si lze představit jako souvislý prázdný prostor mezi atomy proteinu spojující jeho povrch a aktivní místo [59]. Podobně jako u kapes, hrají tunel a jeho vlastnosti jako jeho délka, tvar, šířka nebo aminokyseliny, jež ho tvoří, velkou roli pro vykonávanou biologické funkci [60]. Pro jejich hledání se podobně jako pro kapsy používají geometrické algoritmy, které jsou v současnosti nejčastěji postavené na analýze Voroného diagramu. Pro nalezení relevantních tunelů je většinou potřeba správně nastavit startovací bod výpočtu. Zde je možné vhodně zužitkovat informace o katalytických kapsách, neboť nás zajímají především transportní cesty do tohoto místa. V nástroji HotSpot Wizard je tento výpočet realizován nástrojem CAVER 3 [61], který v této oblasti patří mezi nejpoužívanější.

4.4 Analýza výsledků

Aby bylo možné získaná data využít pro návrh mutantních proteinů, je nutné je vhodně zkombinovat a dát jim tak kvalitativní význam. Podobně jako v každém oboru i v proteinovém inženýrství došlo postupem času k ustálení mnoha různých strategií se zaměřením na různé vlastnosti proteinů. Pro nástroj HotSpot Wizard byly vybrány čtyři zavedené a oblíbené strategie, kterým budou věnovány dalších odstavce.

Funkční hot spoty

Analýza funkčních hot spotů je původní strategií navrženou pro první verzi nástroje. Je určena především k modifikaci biologické funkce enzymu, a to jak z pohledu aktivity, tak i substrátové

specifity a stereoselektivity. Strategie je založena na hledání reziduí nacházejících se v blízkém okolí aktivního místa (resp. katalytické kapsy) a jeho přístupových tunelů, jejichž modifikace má velikou šanci ovlivnit tyto vlastnosti, které však nejsou ani konzervované a ani anotované jako esenciální, a tudíž relativně bezpečné pro modifikaci.

Flexibilní hot spoty

Flexibilní regiony zvyšují entropii celé struktury a mohou tak negativně ovlivňovat její stabilitu. Proto je jednou z metod stabilizace rigidifikace flexibilních regionů a snižování entropie [62]. Pro tento účel se využívá analýza teplotních faktorů struktury s využitím informace o sekundárních strukturách a umístění rezidua na povrchu.

Back-to-consensus hot spoty

Tato strategie se shodně s druhou zaměřuje také na stabilitní hot spoty, ale tentokrát z pohledu sekvence. Jak již bylo zmíněno dříve, konzervovaná rezidua jsou často spojeny s nějakou důležitou funkcí v rámci enzymu. A jednou z nich může být i teplotní stabilita. Ačkoliv není možné s jistotou rozhodnout, pro kterou funkci je reziduum důležité, přesto se zpětným mutováním na evolucí preferované aminokyseliny podařilo v mnoha studiích enzym výrazně stabilizovat [63]–[66]. Proto je tato strategie také součástí nástroje.

Korelované páry

Poslední strategie je založena na předpokladu interakčních sítí v proteinech z pohledu dvojic, které jsou v rámci evoluce mutovány společně. Díky tomu je možné se vyhnout nevhodným mutacím pouze jednoho z páru a naopak vhodným mutováním zlepšit katalytické vlastnosti celého enzymu [67]–[70].

4.5 Návrh chytrých knihoven

Posledním krokem před experimentální charakterizací navržených variant je návrh nového genu, který bude schopen kódovat vybrané mutanty a bude možné jej využít pro jejich syntézu. Jelikož většinou je potřeba otestovat větší množství variant, je nutné mít k dispozici biochemický nástroj schopný automaticky generovat všechny kombinace mutací. Pro tento účel jsou využívány tzv. degenerované kodony [71], které dokáží kódovat více aminokyselin a jimiž jsou v původním genu nahrazeny kodony originální. Jelikož existuje více různých degenerovaných kodonů, je nutné vždy vybrat ten nejvhodnější pro vybranou podmnožinu aminokyselin. Další úskalí se skrývá v matematice, neboť počet možných kombinací roste exponenciálně a pro zajištění dostatečného pokrytí by bylo nutné provést nerealizovatelně velké množství měření. Tudíž je nutné vhodně vyvážit počet reziduí a substituovaných aminokyselin.

Nástroj HotSpot Wizard pro tento účel obsahuje podporu pro omezení testovaných aminokyselin na základě kritérií relevantních pro danou strategii a také automatický návrh degenerovaných kodonů, které nejlépe zachycují množinu požadovaných aminokyselin a zároveň obsahující co nejmenší počet stop kodonů. Jejich seznam s krátkým popisem je uveden v tabulce 4.1. Alternativně lze využít metodu SwiftLib [72], která navrhuje degenerované kodony na základě maximálního limitu diverzity. Ačkoliv takto navržená knihovna nemusí

Tabulka 4.1: Metody pro výběr cílových aminokyselin dostupné při návrhu chytrých knihoven.

Metoda výběru	Strategie	Popis
Frekvence aminokyselin	FUNC, FLEX	výběr aminokyselin splňujících kritérium minimálního výskytu na dané pozici v rámci mnohonásobného zarovnání
Škodlivost mutací	FUNC, FLEX	výběr aminokyselin splňujících kritérium minimální pravděpodobnosti zachování proteinové funkce
Sekvenční konsenzus	CONS	výběr aminokyselin vybraných pro danou pozici na základě sekvenčního konsenzu
Korelované pozice	COREL	výběr aminokyselin splňujících kritérium minimálního výskytu na korelované pozici
Manuální režim	všechny	manuální výběr aminokyselin

FUNC – strategie funkčních hot spotů; FLEX – strategie stabilitních hot spotů s využitím flexibility; CONS – strategie stabilitních hot spotů na základě sekvenčního konsenzu (back-to-consensus); COREL – strategie korelovaných pozic

obsahovat všechny varianty, lze podobně jako v první metodě využít kritéria relevantní pro danou strategii k nastavení vah jednotlivých aminokyselin a tím snížit pravděpodobnost jejího vynechání. Pro obě varianty jsou také automaticky počítány základní statistiky výsledné knihovny, jako její velikost a očekávané pokrytí, s využitím nástroje TopLib [73], což umožňuje uživateli se rozhodnout, kterou knihovnu zvolit pro vlastni experimentální práci vzhledem k očekávané náročnosti.

5 Optimalizace výchozích parametrů

Aby bylo možné poskytnout uživatelům opravdu snadný způsob analýzy jejich proteinů bez nutnosti detailního nastavení nebo znalosti nástrojů, je nutné nejen vybrat správné nástroje, ale také vybrat jejich kvalitní a robustní nastavení. Jen pak bude možné nástroj spolehlivě aplikovat na co nejširší spektrum proteinů. Z tohoto důvodu bylo provedeno několik detailních bioinformatických analýz, jejichž úkolem bylo potvrdit kvalitu navrženého řešení a nalézt jejich vhodné nastavení.

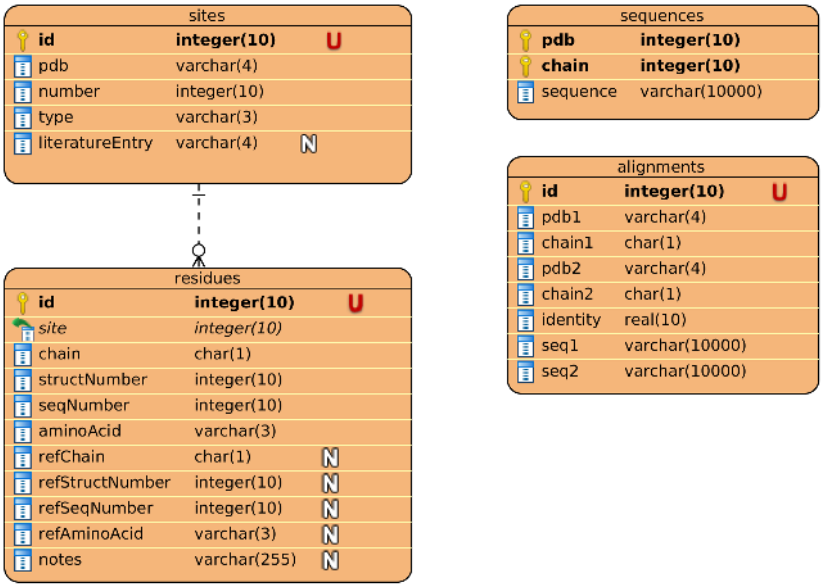
5.1 Rozšíření databáze Catalytic Site Atlas

Catalytic Site Atlas 2.0 (CSA) je databáze shromažďující informace o funkčních reziduích enzymů pro struktury dostupné v databázi Protein Data Bank. Její základ tvoří rezidua manuálně extrahovaná z odborné literatury, které byly následně namapovány na jiné struktury na základě jejich podobnosti. Dle našich testů však tento proces nebyl dostatečně restriktivní a tak se v databázi nachází mnoho irelevantních záznamů často založených na velmi nízké sekvenční podobnosti (výrazně pod obecně přijímaný práh 30 % [74]). To samozřejmě významným způsobem snižuje jejich biologickou relevanci a vede k falešně pozitivním výsledkům. Jelikož jsou kvalitou nalezených funkčních reziduí výrazněny ovlivněny sekvenční i strukturní analýzy nástroje HotSpot Wizard, je vhodné tyto falešně pozitivní rezidua eliminovat a do výpočtu je vůbec nezahrnout. Databáze však v plain-textové verzi neposkytuje vůbec žádné informace o sekvenční podobnosti nebo procesu mapování. Její SQL verze by tato data obsahovat mohla, nicméně její stažitelná verze není kompletní a autoři nereagovali na dotazy ani žádosti. Proto bylo rozhodnuto vytvořit novou verzi této databáze, která bude rozšířená o detailní informace o podobnosti sekvencí a jejich zarovnání. Tak bude možné v databázi interaktivně filtrovat a poskytovat pouze relevantní rezidua.

Struktura databáze

Struktura databáze je založena na tzv. katalytických místech. Pod tímto pojmem se rozumí množina reziduí, které by společně měly tvořit právě jedno funkční místo. Na enzym pak těchto míst může být samozřejmě více. Původní plain-textová verze databáze je uložena ve formátu CSV a obsahuje následující sloupce:

- **PDB ID** – PDB kód struktury
- **SITE NUMBER** – číslo katalytického místa
- **RESIDUE TYPE** – aminokyselina
- **CHAIN ID** – identifikátor řetězce
- **RESIDUE NUMBER** – číslo rezidua dle strukturního číslování
- **CHEMICAL FUNCTION** – informace jestli reziduum participuje na funkci svým bočním řetězcem (S) nebo konstrou (N)
- **EVIDENCE TYPE** – původ záznamu (LIT – literatura, HOM – podobnost)



Obrázek 5.1: Entitně-relační diagram databáze rozšířené verze Catalytic Site Atlas.

- **LITERATURE ENTRY** – reference na strukturu typu LIT, na jejímž základě bylo místo odvozeno

Podoba rozšířené databáze vychází přímo z originální verze a je určena pro standardní relační databázové servery s podporou SQL. Její struktura v podobě entitně-relačního diagramu je na obrázku 5.1.

Tabulka sites. Tabulka obsahuje seznam všech původních katalytických míst s jejich základními údaji jako je identifikátor struktury, číslo, původ záznamu a reference na zdrojovou strukturu.

Tabulka residues. Tato tabulka obsahuje informace o všech reziduích s příslušností k dané kapse. Kromě informací shodných s původní databází obsahuje také informace o sekvenčním indexu, který reprezentuje index daného rezidua v rámci sekvenci (vždy číslováno od nuly bez mezer), a které se bohužel často liší od indexu v rámci struktury. Pro záznamy odvozené na základě podobnosti také obsahuje stejné informace pro zarovnané reziduum z referenční sekvence, na jejímž základě bylo místo odvozeno.

Tabulka sequence. Tabulka obsahuje pouze sekvence pro relevantní struktury s informací o PDB kódu a řetězci.

Tabulka alignments. Tabulka obsahuje informace o zarovnání sekvencí a jejich identitě. Tabulka obsahuje pouze záznamy relevantní pro záznamy v rámci CSA, a tudíž nejde o všechny možné kombinace z tabulky sequence.

Získávání dat

Aby bylo možné databázi naplnit, bylo nutné pro každé aktivní místo zkonstruovat sekvenční zarovnání s referenčním proteinem. Aby se předešlo problémům lokálního zarovnání, pro konstrukci bylo využito globální zarovnání s využitím algoritmu Needleman-Wunsch [75]. Před konstrukcí zarovnání bylo samozřejmě nutné získat sekvence pro všechny proteiny. To probíhalo stažením struktury z PDB databáze a extrakce sekvence z jeho metadata. Dále bylo provedeno její zarovnání se sekvencí získanou ze struktury (pouhé spojení po sobě jdoucích aminokyselin), aby se zjistilo mapování strukturních indexů na sekvenční.

V dalším kroku by tedy mělo následovat zarovnání sekvence s referenční. Bohužel však databáze neposkytuje informaci o řetězci referenčního proteinu a mnoho proteinů má řetězců více, bylo nutné provést zarovnání na všech a vybrat z nich ten s nejvyšší identitou. Poté již bylo možné provést zarovnání sekvencí a všechny důležité informace zapsat do databáze.

Shrnutí

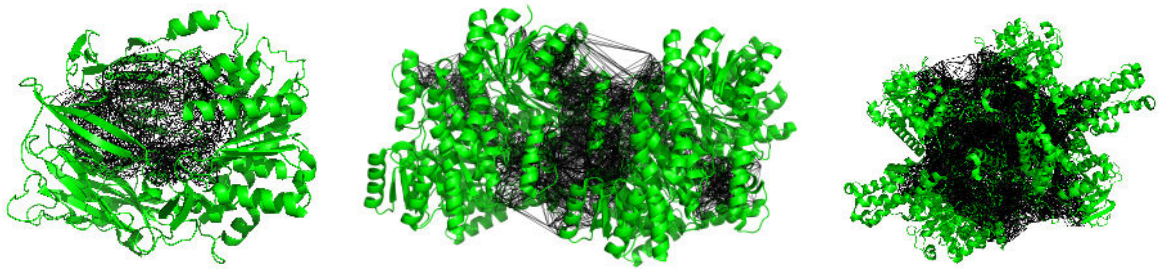
V rámci této analýzy byla sestrojena rozšířená verze databáze Catalytic Site Atlas 2.0. Ta na rozdíl od původní byla obohacena informací o globálním párovém sekvenčním zarovnání s referenčním záznamem. Díky tomu je možné se vyhnout práci s irelevantními záznamy se sekvenční identitou pod 30 %, kterých je v databázi 33 %.

5.2 Identifikace kapes

Jak již bylo dříve zmíněno, detekce kapes je velmi významnou součástí strukturní analýzy a je tedy nutné identifikovat katalytické kapsy s vysokou geometrickou i biologickou relevancí. V původní verzi nástroje HotSpot Wizard byl pro detekci kapes použit nástroj CASTp [76] s výchozím nastavením. Jelikož však tento nástroj na mnoha proteinech poskytuje nepřesné výsledky (obrázek 5.2 z pohledu minimálního a maximálního objemu nalezených kapes (běžná velikost by neměla přesáhnout 1 000 $Å^3$[77]), neumožňuje vizualizaci nalezených kapes, nemá téměř žádné možnosti nastavení a není již aktivně vyvíjen, bylo rozhodnuto o jeho nahrazení. Jako vhodná náhrada byl vybrán nástroj Fpocket. Ten tedy bylo nutné srovnat se stávajícím řešením a v případě potřeby provést optimalizaci jeho vstupních parametrů.

Fpocket

Fpocket je podobně jako jiné nástroje pro hledání kapes založen na konceptu reprezentace struktury pomocí Voroného digramu a následné detekce tzv. alfa sfér. Ty jsou definovány jako kontaktní sféry, které se dotýkají právě 4 různých atomů v prostoru a zároveň v sobě neobsahují žádný jiný atom [78]. Tento geometrický koncept je vhodný pro proteiny také díky jejich charakteristickému uspořádání jednotlivých atomů v prostoru, které se musí řídit fyzikálními zákony. Díky tomu je tak možné snadno rozlišovat mezi povrchem, kapsami a vnitřním prostorem proteinů [79]. Algoritmus nástroje Fpocket se skládá ze tří hlavních kroků. V prvním je provedena konstrukce Voroného diagramu s využitím knihovny Qhull [80] a výpočet alfa sfér. V dalším kroku je provedeno postupné shlukování alfa sfér až do finálních kapes. Toto shlukování je prováděno kvůli rychlosti ve třech různých úrovních. V první úrovni je provedeno hrubé sloučení alfa sfér s využitím triviálního vzdálenostního kritéria



(a) Hydroláza **(b)** Oxydoreduktáza **(c)** Oxydoreduktáza

Obrázek 5.2: Kapsy nalezené nástrojem CASTp se biologicky nerelevantními geometrickými parametry. Proteiny jsou vizualizovány zelenou barvou stylem „cartoon“ a kapsy jsou reprezentovány jako síť čtyřstěnů v černé barvě. (a) Katalytická kapsa hydrolázy (PDB kód: `3GOC`) s objemem 15 079 Å^3, (b) katalytická kapsa oxydoreduktázy (PDB kód: `1GEG`) s objemem 31 795 Å^{3}a (c) katalytická kapsa oxydoreduktázy (PDB kód: `1HWZ`) s objemem 80 502 Å^3.

do tzv. iniciálních kapes a také zahození všech velkých shluků složených pouze z jedné sféry (jedná se o vnější prostor mimo protein). V druhé úrovni se sloučí iniciální kapsy s blízkým těžištěm. Díky čemuž dojde ke sloučení především malých povrchových shluků do jednoho většího. V poslední úrovni je pak provedeno shlukování s využitím multiple-linkage přístupu do výsledných kapes. Třetím krokem je pak ohodnocení kapes, výpočet jejich „druggability“ a uložení do výstupních souborů.

Datová sada

Jako srovnávací datová sada byly využity unikátní struktury použité jako vstupy pro HotSpot Wizard 1.7 za celou dobu jeho existence. Díky tomu by měla datová sada obsahovat především proteiny zajímavé pro cílovou skupinu uživatelů a tedy bude srovnání a optimalizace parametrů prováděna na skutečně relevantních datech. Do datové sady byly zkombinovány jak struktury zadané pomocí PDB kódu, tak i ty uživatelsky nahrané. Detaily o jejich funkčních reziduích byly extrahovány z výsledkových souborů z nástroje HotSpot Wizard 1.7, který je také získává z databází SwissProt a Catalytic Site Atlas, bez dalšího extra vyhledávání. Po odfiltrování duplicitních a očividně nekorektních struktur a proteinů bez známých funkčních reziduí zůstalo 1 670 z celkem 4 055 struktur.

Srovnání výchozích parametrů s původním přístupem

Srovnání výsledků obou nástrojů s výchozími parametry ukázalo, že Fpocket detekuje v průměru menší počet kapes na strukturu s lepšími geometrickými vlastnostmi (tabulka 5.1). Zajímavé také je, že většina detekovaných katalytických kapes se umístila na prvním místě v interním řazení nástroje Fpocket podle „druggability“, což koresponduje s autory deklarovanou kvalitou této metriky. Bohužel však nedokázal detekovat katalytické kapsy ve stejném množství struktur jako původní CASTp. Ačkoliv nelze výkon CASTp s ohledem na jeho tendenci hledat příliš velké kapsy považovat za směrodatný, neschopnost testovaného nástroje nalézt katalytickou kapsu u 127 struktur (8 %) bylo motivací k provedení pokusu o optimalizaci jeho vstupních parametrů. Parametry k optimalizaci byly vybírány na základě manuálu

Tabulka 5.1: Výsledky nástrojů CASTp a Fpocket s použitím výchozího nastavení.

Nástroj	Dokončených úloh	Nalezených katalytických kapes	Počet kapes		Medián pořadí	Objem katalytických kapes		
			Medián	Max		Medián	Min	Max
CASTp	2 847	1 627	53	1 125	48	727	11	472 939
Fpocket	2 787	1 500	12	311	1	807	123	10 691

k nástroji a krátkého iniciálního testování, kterým byl potvrzen jejich vliv na nalezené kapsy. Testované hodnoty byly úmyslně vybírány v blízkém okolí výchozí hodnoty a to na základě jejich očekávaného chování a v takovém počtu, aby bylo možné analýzu výpočetně provést.

Optimalizace parametrů

Optimalizace parametrů probíhala v několik po sobě jdoucích krocích.

Krok 1: Minimální poloměr alfa sféry. V prvním kroku optimalizace byl studován efekt minimálního poloměru alfa sféry nacházející se ve vazebném místě (`-m`, výchozí hodnota: 3 Å). Předpokladem bylo, že nižší minimální poloměr povede k vyššímu počtu nalezených kapes a tudíž i většímu počtu správně označených katalytických reziduí. Pro testování bylo vybráno rozpětí parametru 2,0–2,9 Å s krokem 0,1 Å. Analýzou výsledků uvedených v tabulce 5.2 byl tento předpoklad potvrzen a bylo nalezeno větší množství katalytických kapes. Nicméně se snižujícím se poloměrem začalo pochopitelně docházet k postupnému zhoršení jejich geometrických vlastností (jednotlivé kapsy a dutiny se slučovaly do extrémně rozlehlých celků), a to dokonce hluboko pod kvalitu původního CASTp. Pro další krok byl tedy vybrán interval 2,5–2,7 Å, který vykazuje asi nejlepší poměr mezi množstvím nalezených katalytických kapes a jejich geometrickými vlastnostmi.

Krok 2: Minimální počet alfa sfér. Druhý krok analýzy mířil na zmenšení objemu nalezených kapes neboť jejich přehnaná velkost by pak vedla k návrhu velkého množství falešně pozitivních hot spot reziduí bez většího vlivu na funkci proteinů. V tomto kroku byl optimalizován parametr udávající minimální počet alfa sfér nutných pro sloučení dvou kapes v rámci poslední úrovně shlukování (`-n`, výchozí hodnota: 2). Zvyšováním této hodnoty by tak mělo docházet k detekci většího množství menších kapes a tím odstranění irelevantních spojení a snížení jejich objemu. V prvním kole byly testovány hodnoty 3, 4 a 5. Z tabulky 5.3 je patrné, že ke zlepšování geometrických parametrů katalytických kapes skutečně dochází. Bohužel i tak výsledky bylo zlepšení relativně malé a zvláště s ohledem na hodnotu minimálního poloměru alfa sféry 2,5 Å vykazující nejvyšší počet katalytických kapes nebyl zcela uspokojivý. Proto byl ještě proveden další test, kdyz pro hodnotu minimálního poloměru alfa sféry 2,5 Å byla rapidně zvýšena hodnota minimálního počtu alfa sfér až na 10. Tímto zvýšením opět došlo k mírnému zlepšení, ale bohužel stále ne dostatečnému – medián přes 1 600 Å^3 s maximem téměř 100 000 Å^3 ani zdaleka neodpovídá biologické relevanci. Kvůli nevelkému přínosu tohoto parametru bylo jeho další optimalizování odloženo a další kroky tak byly zaměřeny na další vhodné parametry.

Tabulka 5.2: Srovnání výsledků s různými hodnotami minimálního poloměru alfa sféry.

-m	Dokončených úloh	Nalezených katalytických kapes	Počet kapes		Medián pořadí	Objem katalytických kapes		
			Medián	Max		Medián	Min	Max
CASTp	2 847	1 627	53	1125	48	727	11	472 939
3,0	2 787	1 500	12	311	1	807	123	10 691
2,0	2 847	1 641	3	174	1	22 575	299	593 665
2,1	2 847	1 648	3	157	1	21 261	185	595 989
2,2*	–	–	–	–	–	–	–	–
2,3	2 846	1 648	6	129	1	16 933	97	500 538
2,4	2 846	1 638	10	158	1	13 355	58	508 747
2,5	2 845	1 634	16	219	1	8 268	49	370 316
2,6	2 835	1 621	22	405	1	2 078	58	100 062
2,7	2 829	1 605	21	534	1	1 320	56	36 789
2,8	2 824	1 582	18	475	1	1 052	54	19 974
2,9	2 812	1 531	15	387	1	854	83	10 711

* – z důvodu chyby při spuštění výpočet nedoběhl; s ohledem na irelevanci okolních hodnot nebylo dopočítáváno

Tabulka 5.3: Srovnání výsledků s různými kombinacemi hodnot minimálního poloměru alfa sféry s minimálním počtem alfa sfér.

-m	-n	Dokončených úloh	Počet kapes		Medián pořadí	Objem katalytických kapes		
			Nalezených katalytických kapes	Medián	Max	Medián	Min	Max
CASTp		2 847	1 627	53	1 125	48	727	11	472 939
3,0	2	2 787	1 500	12	311	1	807	123	10 691
2,5	3	2 845	1 634	18	262	1	6 967	50	326 101
2,5	4	2 845	1 632	21	295	1	5 496	42	155 688
2,5	5	2 844	1 632	23	354	1	4 337	46	132 304
2,6	3	2 835	1 620	24	480	1	1 753	64	66 691
2,6	4	2 834	1 619	25	557	1	1 421	65	65 908
2,6	5	2 832	1 619	27	639	2	1 220	59	51 210
2,7	3	2 827	1 604	22	580	1	1 186	53	30 382
2,7	4	2 826	1 603	23	626	1	1 076	56	25 879
2,7	5	2 825	1 603	24	658	1	961	54	20 986
2,5	10	2 844	1 632	31	599	2	1 623	46	97 341

Krok 3: Maximální vzdálenost těžišť a alfa sfér. Třetí krok podobně jako druhý cílí na parametry ovlivňující shlukování kapes se shodným cílem – snížením objemu nalezených kapes při zachování jejich počtu. Hodnota parametru minimálního poloměru alfa sféry byla zafixována na hodnotě 2,5 Å při které bylo dosaženo nejvyššího počtu kapes a hodnota minimálního počtu alfa sfér byla nastavena na 10, kdy dokáže alespoň částečně kompenzovat jejich geometrii. Pro optimalizaci pak byly vybrány dva další parametry. První z nich určuje maximální vzdálenost těžišť iniciálních kapes, které budou sloučeny v rámci druhé úrovně shlukování (`-r`, výchozí hodnota: 4,5 Å) a byly pro něj testovány hodnoty 3,5, 4,0 a 4,5 Å.

Tabulka 5.4: Srovnání výsledků s různými kombinacemi hodnot maximální vzdálenosti těžišť shluků a maximální vzdálenosti nejbližších alfa sfér sousedních shluků.

			Počet kapes			Objem katalytických kapes			
-r	**-s**	**Dokončených úloh**	**Nalezených katalytických kapes**	**Medián**	**Max**	**Medián pořadí**	**Medián**	**Min**	**Max**
CASTp		2 847	1 627	53	1 125	48	727	11	472 939
4,5	2,5	2 787	1 500	12	311	1	807	123	10 691
3,5	2,0	2 832	1 620	36	931	3	842	47	23 777
3,5	2,5	2 832	1 624	28	585	2	1 356	49	77 879
4,0	2,0	2 833	1 626	38	989	3	849	40	25 004
4,0	2,5	2 833	1 628	29	601	2	1 396	47	92 242
4,5	2,0	2 844	1 632	43	1 076	3	849	50	25 311

Druhým parametrem je maximální vzdálenost dvou nejbližších alfa sfér kapes, které budou sloučeny v rámci třetí úrovně shlukování (`-s`, výchozí hodnota: 2,5 Å). Pro něj byly testovány hodnoty 2,0 a 2,5 Å. Z tabulky 5.4 je očividné, že modifikace maximální vzdálenosti těžišť nemá na výsledky kapsy prakticky žádný vliv. Naproti tomu snižování vzdálenosti alfa sfér má vliv velmi výrazný a to dokonce tak moc, že jeho snížení vede téměř až k polovičním hodnotám a to vše při zachování téměř shodného počtu nalezených katalytických kapes. Jediným překvapivým výsledkem je snížení mediánu pořadí kapsy ve výsledkovém souboru a tím narušení relevance skóre „druggability“. Bohužel i přes tyto výrazné změny v geometrii byly nalezené kapsy stále příliš velké a často zasahovaly až do irelevantních částí proteinu. Jelikož by další agresivní úprava tohoto parametru mohla opět oslabit význam skóre „druggability“, bylo rozhodnuto slevit z aktuální hodnoty minimálního poloměru alfa sféry a zvýšit ji na hodnotu 2,8 Å. Ačkoliv je tato hodnota výrazným skokem od původní uvažované, byla vybrána jako kompromis mezi počtem nalezených katalytických kapes (o 52 méně než 2,5 Å, ale stále o 82 více než původní 3,0 Å) a jejich rozumnou velikostí (v mediánu o 7 186 $Å^3$ oproti 2,5 Å a jen o 245 $Å^3$ oproti 3,0 Å).

Krok 4: Maximální vzdálenost alfa sfér pro vyšší poloměr. Zvýšení poloměru se ukázalo jako dobré rozhodnutí, neboť se objem katalytických kapes najednou výrazně zmenšil a to až na akceptovatelné hodnoty. Podobně došlo k napravení mediánu pořadí ve výsledcích. Hodnota minimálního počtu alfa sfér byla opět ponechána na hodně 10, protože dle výsledků z kroku 2 pozitivně ovlivňuje geometrii při zachování shodného počtu nalezených kapes. S ohledem na výrazný přínos upraveného parametru maximální vzdálenosti nejbližších alfa sfér v předchozím kroku bylo provedeno jeho opětovné otestování s vyšším poloměrem. Překvapivě však byl jeho přínos prakticky zanedbatelný a parametry nalezených kapes zůstávaly velmi podobné (tabulka 5.5). Při porovnání s výsledky CASTp lze vidět výrazně lepší geometrické parametry nalezených kapes, které ale byly vykoupeny o trochu nižší úspěšností v celkovém počtu (asi o 50). Nicméně toto snížení je s ohledem na kvalitu nalezených kapes akceptovatelnou cenou. Tímto krokem se tedy podařilo dosáhnout dostatečně kvalitních parametrů, a tudíž žádný další optimalizační krok již nebyl proveden.

Tabulka 5.5: Srovnání výsledků s různými hodnotami maximální vzdálenosti nejbližších alfa sfér sousedních shluků.

			Počet kapes			Objem katalytických kapes		
-s	**Dokončených úloh**	**Nalezených katalytických kapes**	**Medián**	**Max**	**Medián pořadí**	**Medián**	**Min**	**Max**
CASTp	2 847	1 627	53	1 125	48	727	11	472 939
2,5	2 787	1 500	12	311	1	807	123	10 691
2,0	2 822	1 572	21	648	2	606	49	5 733
2,5	2 822	1 573	21	598	1	734	47	7 008

Tabulka 5.6: Parametry katalytických kapes nalezených v testovací datové sadě haloalkandehalogenáz.

PDB ID	**Řetězec**	**Počet esenciálních reziduí**	**Pořadí kapsy**	**Skóre**	**Objem [Å^3]**	**Esenciálních reziduí v kapse**
1B6G	A	5	4	16,310	134	4
1IZ7	A	5	1	35,957	467	4
2PSD	A	5	1	38,625	561	4
2QVB	A	3	1	34,587	572	2
	B	3	3	31,303	463	2
2XT0	A	5	1	33,260	1 144	5
3A2M	A	3	1	36,993	1 260	2
	B	3	2	32,762	1 022	2
3U1T	A	5	3	25,661	564	4
	B	5	4	20,016	403	4
3WI7	A	3	3	15,710	418	2
	B	3	2	18,556	431	2
4C6H	A	3	1	37,629	732	1
4E46	A	3	1	34,634	526	2
4K2A	A	3	5	19,228	234	2
	B	3	7	16,535	264	2
	C	3	11	13,327	225	2
	D	3	2	27,503	1 245	2

Validace

Ověření relevance dosažených výsledků bylo provedeno manuálním porovnáním nalezených katalytických kapes (tabulka 5.6) se známými kapsami několika reprezentativních struktur z rodiny enzymů halogenalkandehalogenáz (analýzu provedli odborníci z Loschmidtových laboratoří). Tato rodina je dlouhodobě zkoumána Loschmidtovými laboratořemi a je dobře znám jejich reakční mechanismu, lokace aktivního místa i geometrické parametry katalytické kapsy. Umístění a geometrie nalezených kapes u všech struktur dobře korelovaly s očekávaným výsledkem.

Tabulka 5.7: Výsledné hodnoty parametrů.

Název	Parametr	Původní hodnota	Optim. hodnota
minimální poloměr alfa sféry	`-m`	3	2,8
minimální počet společných alfa sfér	`-n`	2	10
maximální vzdálenost těžišť	`-r`	4,5	4,5
maximální vzdálenost nejbližších alfa sfér	`-s`	2,5	2,5

Shrnutí

V rámci této analýzy byla provedena optimalizace vstupních parametrů nástroje Fpocket a srovnána s původním nástrojem CASTp. V průběhu analýzy bylo postupně otestováno více než 25 různých kombinací hodnot čtyř parametrů na sadě 1 670 proteinů, z nichž byla vybrána nejlepší kombinace uvedená v tabulce 5.7. Tato kombinace zajišťuje nalezení katalytické kapsy ve většině struktur se zachováním jejich velmi dobrých geometrických vlastností. Dále byl potvrzen význam skóre „druggability“, kdy byla jako katalytická označována většinou kapsa s jeho nejvyšší hodnotou, a tudíž je možné ho využít jako kritérium pro její výběr.

5.3 Optimalizace dalších analýz

K většině dalších netriviálních částí workflow byly také provedeny bioinformatické analýzy a optimalizace jejich parametrů. Ačkoliv mají srovnatelný význam s výše rozebíranými, byly prováděny dalšími spoluautory nástroje, a proto jim v této práci bude věnován pouze omezený prostor.

Konstrukce mnohonásobného sekvenčního zarovnání

Autor: Jaroslav Bendl, Fakulta informačních technologií, Vysoké učení technické v Brně

V rámci této analýzy byl optimalizován proces konstrukce mnohonásobného zarovnání a navazující analýzy konzervovanosti tak, aby zarovnání obsahovalo dostatečný počet sekvencí pro získání očekávané distribuce skóre konzervovanosti v rámci sekvence. Konkrétněji byl testován vliv e-value na hledání podobných sekvencí a zejména pak finální velikost zarovnání. V případě algoritmu Jensen-Shanonovy divergence (JSD) použitého pro výpočet konzervovanosti pak bylo studováno nastavení penalizace mezer v zarovnání a také bylo experimentováno s využitím plovoucího okna při počítání konzervovanosti.

Výběr nástrojů pro identifikace korelovaných párů

Autor: Ondřej Vávra, Přírodovědecká fakulta, Masarykova Univerzita

Nástrojů schopných na základě mnohonásobného sekvenčního zarovnání identifikovat korelované páry existuje celá řada. Cílem této analýzy tak bylo sestrojit nezávislou datovou sadu a na jejím základě ověřit přesnost vybraných nástrojů a vybrat ten nejlepší. Jelikož se však ukázalo, že žádný z nich nepodává jednoznačně nejlepší výsledky, bylo navrženo jejich spojení v metaprediktor na základě z-skóre.

Predikce škodlivosti mutací

Autor: Miloš Musil, Fakulta informačních technologií, Vysoké učení technické v Brně

Predikce možné škodlivosti zvažovaných mutací je cenným zdrojem informací pro výběr pozic a cílových aminokyselin do experimentální knihovny. Původně zvažovaným kandidátem na integraci byl nástroj MAPP, který funguje velmi rychle a dosahuje dostatečné přesnosti. Jeho limitací ale byla restriktivní licence, která znemožňovala jeho snadnou integraci a velmi limitovaná sada používaných fyzikálně chemických rysů jednotlivých aminokyselin. Proto bylo rozhodnuto vzít jeho základní myšlenku, rozšířit ji a vytvořit tak nástroj vlastní s ideálně lepší přesností. Navržený algoritmus byl testován na rozsáhlé sadě téměř 80 tisíc mutací a dosahoval lepších výsledků než nástroj MAPP a téměř srovnatelných výsledků s mnohonásobně pomalejším nástrojem SNAP [81].

6 Implementace

Nástroj HotSpot Wizard je založen shodně s dalšími portály vyvinutými v Loschmidtových laboratořích na platformě Loschmidt Core [82], která tak výrazným způsobem ovlivňuje jeho architekturu. Ačkoliv je celá platforma napsaná v Javě, je HotSpot Wizard vyvíjen a testován výhradně pro operační systémy založené na Linuxu a to především kvůli nedostupnosti velkého počtu integrovaných nástrojů na dalších platformách. Samotný vývoj byl rozdělen na dvě nezávislé komponenty (výpočetní jádro a grafické uživatelské rozhraní), které budou blíže popsány v této kapitole.

6.1 Platforma Loschmidt Core

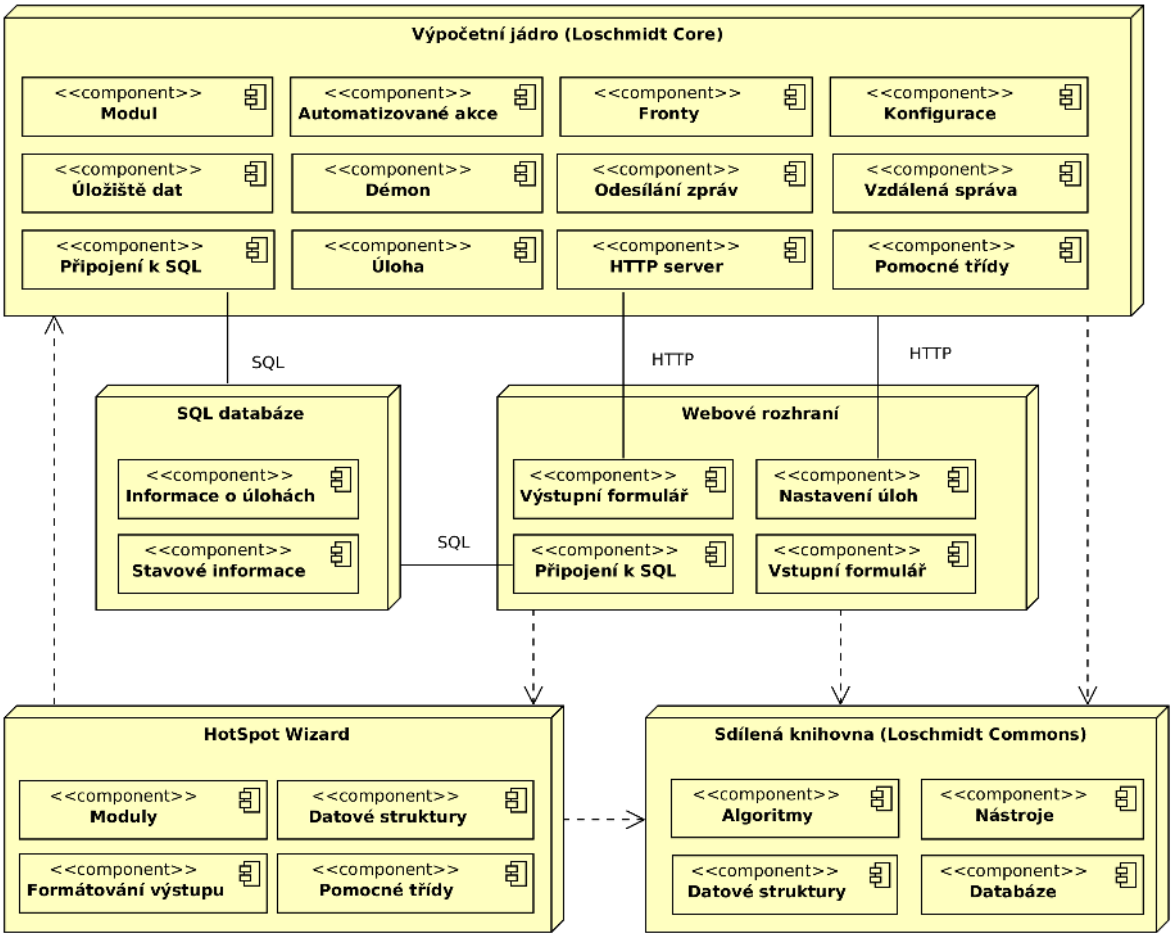
Loschmidt Core [82] je specializovaná platforma v jazyce Java navržená přímo pro vývoj výpočetních jader bioinformatických nástrojů. Jejím hlavním cílem je poskytnout robustní a spolehlivý základ pro vlastní aplikace a umožnit vývojáři soustředit se primárně na návrh a implementaci bioinformatického workflow aplikace bez nutnosti řešení „servisních“ věcí typu zpracování úloh nebo řízení zdrojů. Celá platforma je dobře konfigurovatelná a navržena modulárně, což umožňuje snadné nahrazení většiny komponent vlastní implementací. Kromě samotné platformy je její součástí také rozsáhlá kolekce bioinformatických tříd (Loschmidt Commons) zahrnujících implementace vybraných algoritmů, parsery výstupních formátů externích aplikací, rozhraní pro přístup k několika bionformatickým databázím a také datové modely běžně využívaných elementů.

6.2 Architektura aplikace

Architektura nástroje (obrázek 6.1) je převzata z doporučení pro platformu Loschmidt Core. To znamená, že výpočetní jádro je postavené přímo na platformě a jeho vývoj je striktně oddělen (kromě datové struktury) od vývoje grafického uživatelského rozhraní. Komunikace mezi jednotlivými komponentami je zajišťována pomocí databáze s metadaty a standardního HTTP protokolu pro přenos datových souborů. Díky tomu je jejich vzájemná závislost velmi malá, což pak umožňuje vývoj, kdy je možné každou komponentu vyvíjet prakticky nezávisle nebo ji snadno naradit. V neposlední řadě je výhodou možnost provozovat komponenty odděleně, čímž se výrazně zvyšuje škálovatelnost a také možnost správy, kdy je možné více uživatelských rozhraní provozovat na stejné instanci webového serveru bez vzájemného omezování výpočetního výkonu na úlohy, zatímco jejich výpočetní jádra mít podle potřeb rozložena na více strojů.

6.3 Datová struktura

Pro snadné ukládání a manipulaci s výsledky jednotlivých analýz, byla navržena specializovaná datová struktura (obrázek 6.2). Jejím cílem je co nejpřirozeněji reprezentovat protein a jeho jednotlivé komponenty v bioinformatickém pohledu v hierarchii protein → řetězec → reziduum. Díky tomu je možné výsledky analýz mapovat přímo na vhodné komponenty. Asi největší komplikací při návrhu datové struktury byla nutnost podpory dvou různých typů indexů reziduí – sekvenčního a strukturního. Zatímco sekvenční index určuje pozici rezidua v sekvenci

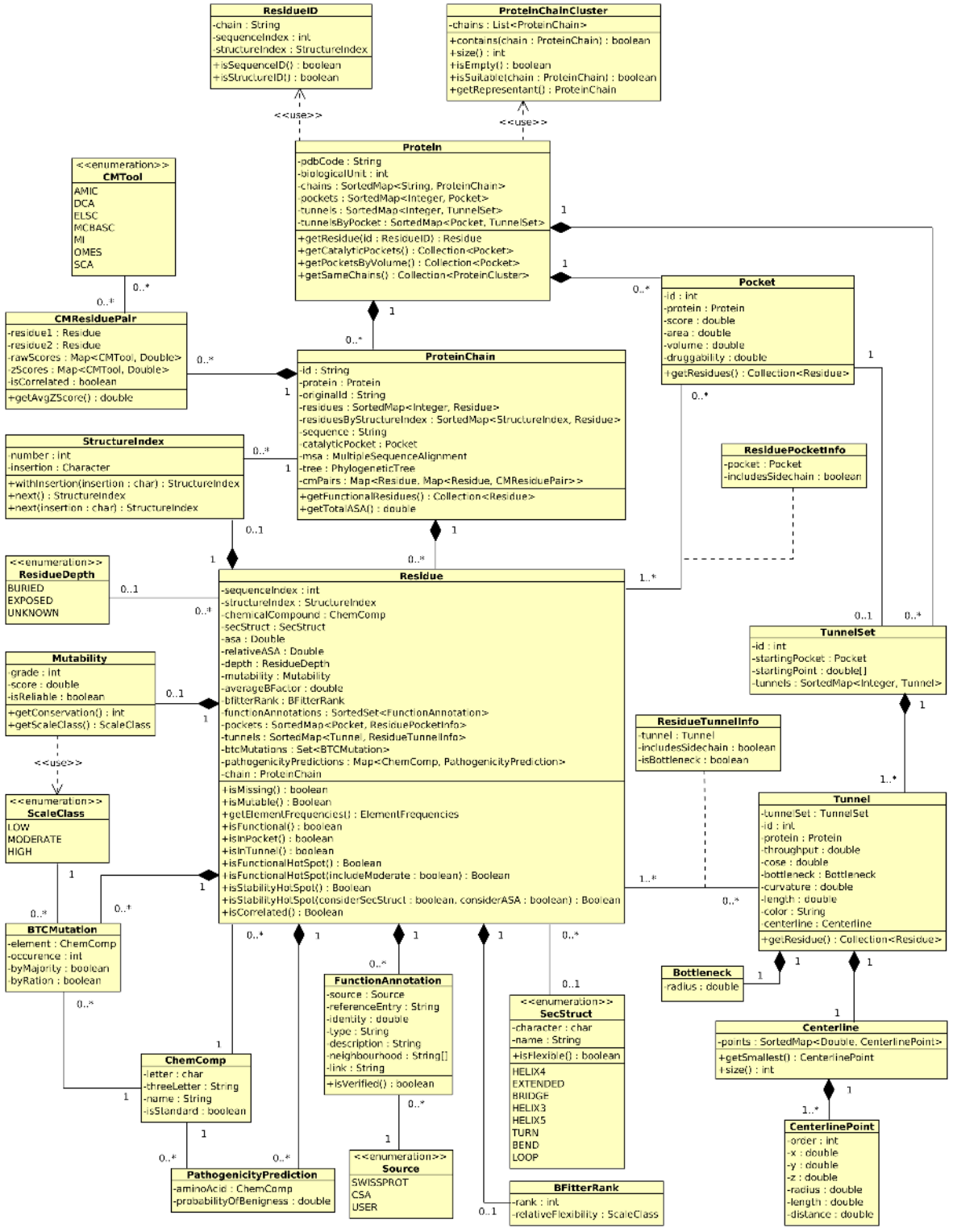


Obrázek 6.1: Diagram aktivit popisující architekturu nástroje HotSpot Wizard v kontextu platformy Loschmdit Core.

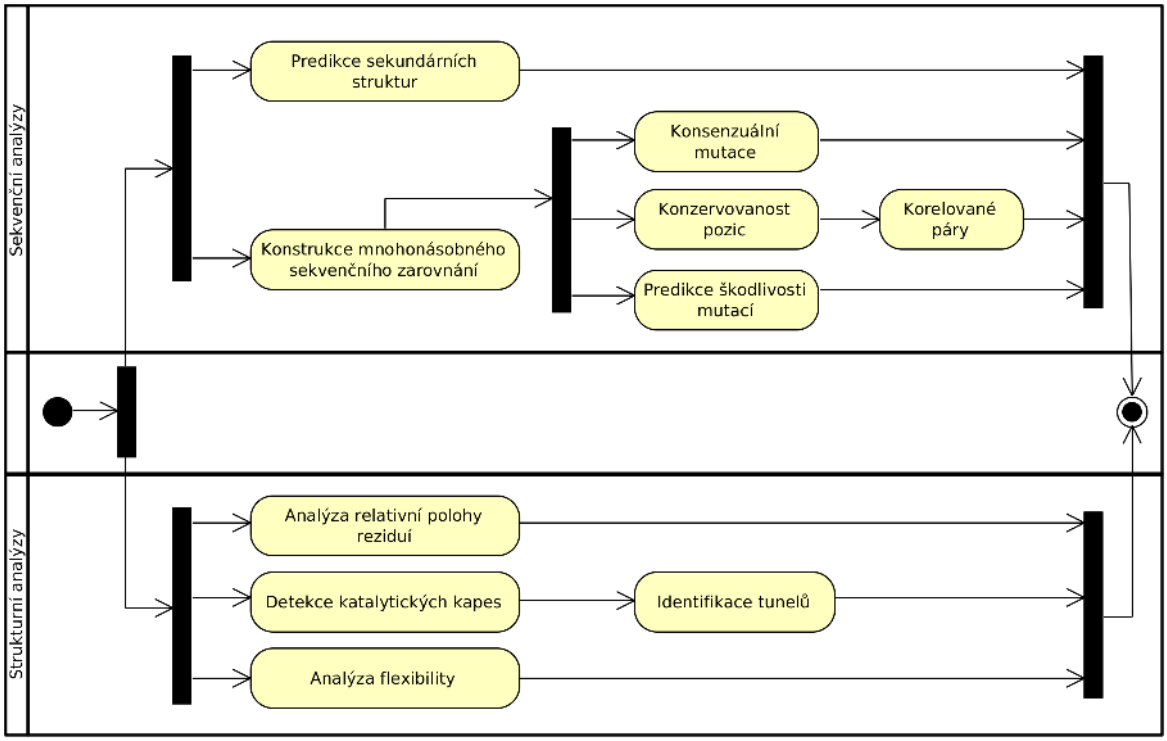
a jde vždy nepřerušovaně od jedničky, index strukturní reprezentuje identifikátor v rámci struktury. Kvůli chybovosti a omezení experimentálních technik tak tento index nejenže nemusí začínat od jedničky, ale také často obsahuje mezery nebo není sekvenční. Podpora obou variant pak byla vynucena jednotlivými nástroji, které většinou dokáží pracovat pouze s jedním typem. Pro účely přenositelnosti a další zpracovatelnosti byla také implementována podpora pro její serializaci do formátu XML. Tato serializace je napsaná manuálně bez využití knihoven a to kvůli potřebě pokročilé filtrace ukládaných informací. Formát XML pak byl zvolen s ohledem na svou snadnou strojovou zpracovatelnost poskytující uživatelům možnost s výstupy nadále pracovat.

6.4 Výpočetní jádro

Implementace výpočetního jádra je plně podřízena požadavkům platformy Loschmidt Core. Z externích knihoven pak HotSpot Wizard využívá především open-source projektu BioJava [83], a to především robustní implementaci parseru PDB formátu a také algoritmy pro predikci sekundárních struktur a povrchu proteinu. Dále jsou používány knihovny UniProtJAPI [84]



Obrázek 6.2: Diagram tříd reprezentující datovou strukturu `Protein`. Zobrazeny jsou všechny třídy, jejich atributy a vybrané metody (`get`/`set` metody pro atributy byly vynechány).



Obrázek 6.3: Diagram aktivit zobrazující průběh jednotlivých kroků analýz, které jsou vizuálně rozdělené na *sekvenční* a *strukturní* větev.

pro programatický přístup k databázi SwissProt a knihovna `covariance` [85] obsahující implementace algoritmů MI, McBASC, OMES, SCA, ELSC pro identifikaci korelovaných párů.

Na jeho vstupu je XML soubor definující úlohu, její nastavení a odkazy na další soubory, kterými jsou proteinová struktura v PDB formátu a také serializovaná datová struktura `Protein` s anotacemi funkčních míst. Tento soubor je zpracován a na jeho základě sestavena úloha obsahující nakonfigurované a instancované jednotlivé kroky výpočtu (tzv. moduly). Těch je celkem 10 a přesně reprezentují jednotlivé analýzy popsané v kapitole Výpočet. Mezi kroky také byly definovány nutné závislosti, zajišťující správnou logickou návaznost výpočtu a zároveň poskytující možnost paralelního spouštění výpočtů, což výrazně zvyšuje propustnost celého nástroje. Jednotlivé kroky a jejich závislosti jsou znázorněny pomocí diagramu aktivit na obrázku 6.3. Pro vzájemné sdílení a ukládání dat napříč moduly slouží datová struktura `Protein` popsaná v kapitole 6.3.

Aby bylo zajištěno vhodné vytížení výpočetních zdrojů, jsou výpočetní kroky řazeny do dvou front. První z nich je určena pouze na konstrukci mnohonásobného zarovnání a umožňuje paralelní běh pouze dvou výpočtů. Tato fronta byla vytvořena především kvůli vysoké diskové náročnosti prohledávání sekvenční databáze nástrojem BLAST, kdy při vyšším počtu souběžných instancí dochází vlivem neefektivního čtení z disku k výraznému poklesu celkového výkonu. Druhá fronta pak slouží pro ostatní výpočty a umožňuje paralelní běh až čtyř analýz. Obě fronty jsou aktuálně bez pokročilého řazení, neboť současné vytížení nástroje to nevyžaduje.

Hlavním výstupem výpočtu je serializovaná datová struktura `Protein` v XML souboru. Ta obsahuje vybranou relevantní podmnožinu všech napočítaných dat nutných pro anotaci proteinu a analýzu výsledků ve webovém rozhraní nebo alternativním zpracováním uživatelem. Dále jsou do jednoho archivu zabaleny všechny výstupní soubory integrovaných nástrojů, aby měl uživatel k dispozici všechna surová data z mezikroků pro případnou detailní analýzu či kontrolu výsledků.

6.5 Časová náročnost výpočtu

Pro ilustraci časové náročnosti výpočtu byla provedena analýza nástrojem s výchozím nastavením pro 9 různých proteinů. Výpočet probíhal na provozní verzi nástroje s prázdnou frontou (úloha tedy nečekala v žádné frontě). Naměřené hodnoty jsou shrnuty v tabulce 6.1. Z hodnot lze vidět, že časová náročnost závisí nejen na absolutní velikosti proteinu, ale také na jeho oligomerním stavu. To je způsobeno optimalizací časově nejnáročnějších sekvenčních analýz, které se pro shodné řetězce (tzn. 100% sekvenční identita) provádí pouze jednou a výsledky se rozkopírují.

Tabulka 6.1: Časová náročnost výpočtu pro vybrané proteiny.

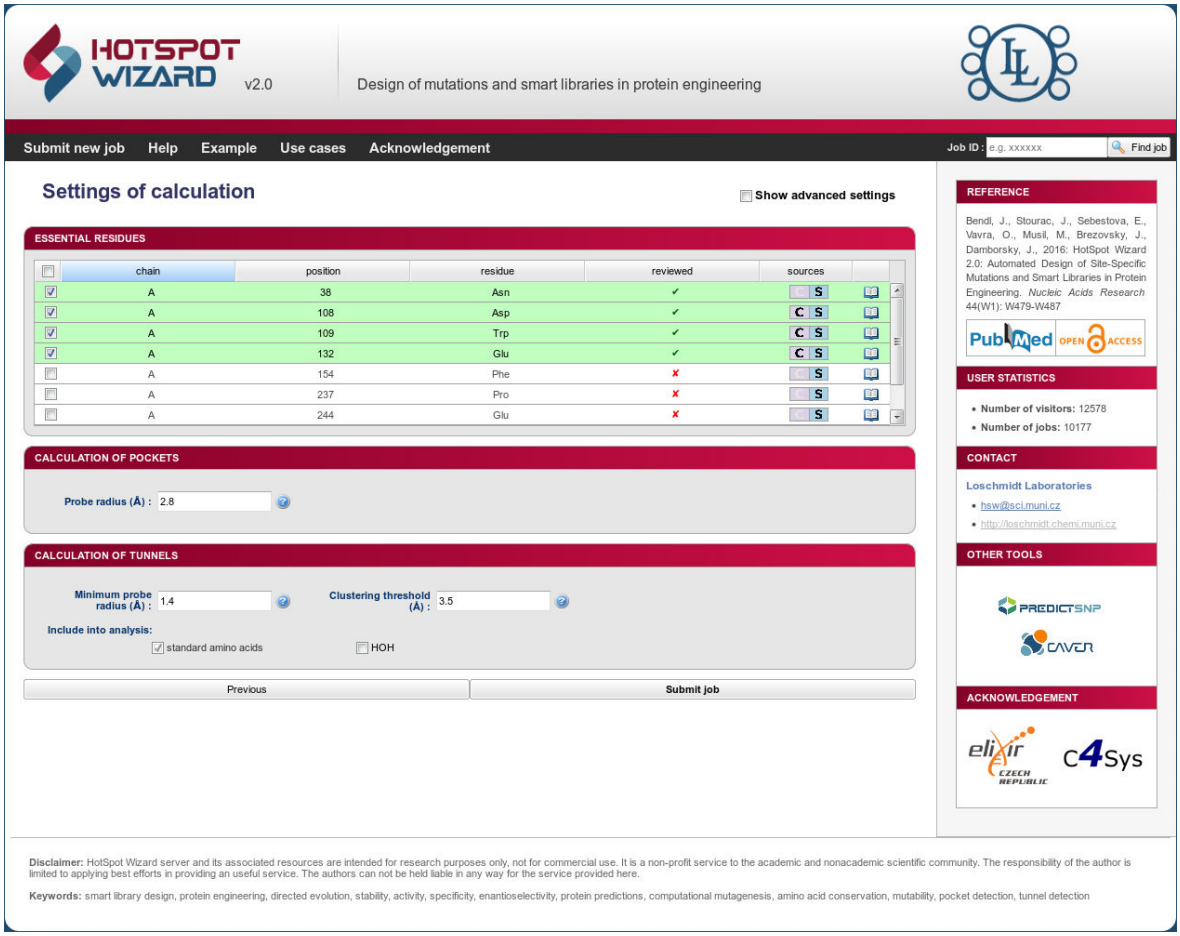
PDB kód	Oligomerní stav	Počet aminokyselin	Počet atomů	Délka výpočtu
`4OEE`	monomer	155	1 068	~4 minuty
`2OMB`	homodimer	434	6 387	~6 minut
`4E46`	monomer	296	2 491	~10 minut
`4LRJ`	heterodimer	296	2 258	~10 minut
`5K3A`	homodimer	590	4 707	~10 minut
`1CFJ`	monomer	535	4 244	~30 minut
`5HQ4`	monomer	669	5 344	~40 minut
`5KSK`	homodimer	1 496	11 095	~55 minut
`2AKA`	heterodimer	1 075	8 625	~65 minut

6.6 Grafické uživatelské rozhraní

Podoba uživatelského rozhraní vychází ze šablony využívané pro všechny webové portály vyvíjené v rámci Loschmidtových laboratoří. Při jeho návrhu bylo věnováno největší úsilí směrem k přehlednosti a interaktivitě, díky čemuž je možné pohodlně provádět všechny analýzy přímo ve webovém rozhraní bez nutnosti externích nástrojů. Webové rozhraní je implementováno stejně jako jádro v programovacím jazyce Java a následně transpilováno do optimalizovaného JavaScriptu s využitím technologie Google Web Toolkit [86]. Většina interaktivních komponent pak pochází z platformy Smart GWT [87]. Důvodem pro volbu této technologie byla její robustnost a široká paleta dostupných komponent, díky čemuž se vývoj výrazně zrychlil a zjednodušil. V neposlední řadě je výhodou možnost vyvíjet přímo v jazyce Java.

Vstupní formulář

Vstupní formulář se skládá ze dvou po sobě jdoucích kroků. V prvním je uživatel vyzván k zadání nebo nahrání proteinové struktury a následně vybrání vhodné biologické jednotky



Obrázek 6.4: Uživatelské rozhraní druhého kroku vstupního formuláře.

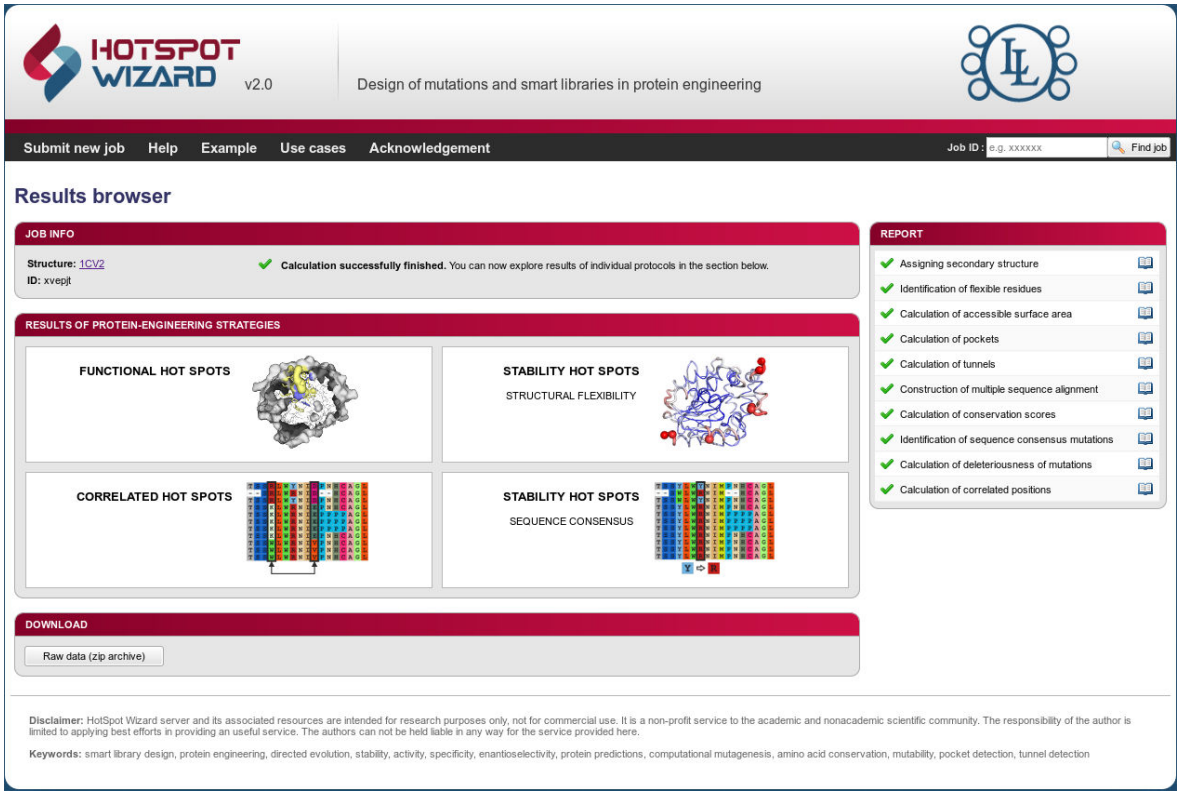
či řetězců. Dále je možné zadat emailovou adresu pro notifikační zprávy o stavu úlohy a vlastní název úlohy. V druhém (obrázek 6.4) je pak možné zkontrolovat a upravovat automaticky nalezená esenciální rezidua a také měnit konfigurační parametry jednotlivých kroků výpočtu. Pro lepší přehlednost formuláře pro méně zkušené uživatele je většina konfiguračních voleb implicitně skryta a je možné je zobrazit přepnutím na pokročilý mód. Poté je možné úlohu rovnou odeslat ke zpracování a uživatel je ihned přesměrován na stránku se stavem úlohy.

Shrnutí úlohy

Tato stránka (obrázek 6.5) je středobodem celé analýzy výsledků. V případě korektně dokončené úlohy se skládá z následujících komponent:

- informační panel obsahující základní informace o úloze (identifikátor, název a identifikátor proteinu) a globální shrnutí stavu úlohy,

- panel detailních informací o průběhu výpočtu, kde jsou vidět detailní informace o stavu jednotlivých kroků výpočtu a je možné si zobrazit jejich stavové zprávy,



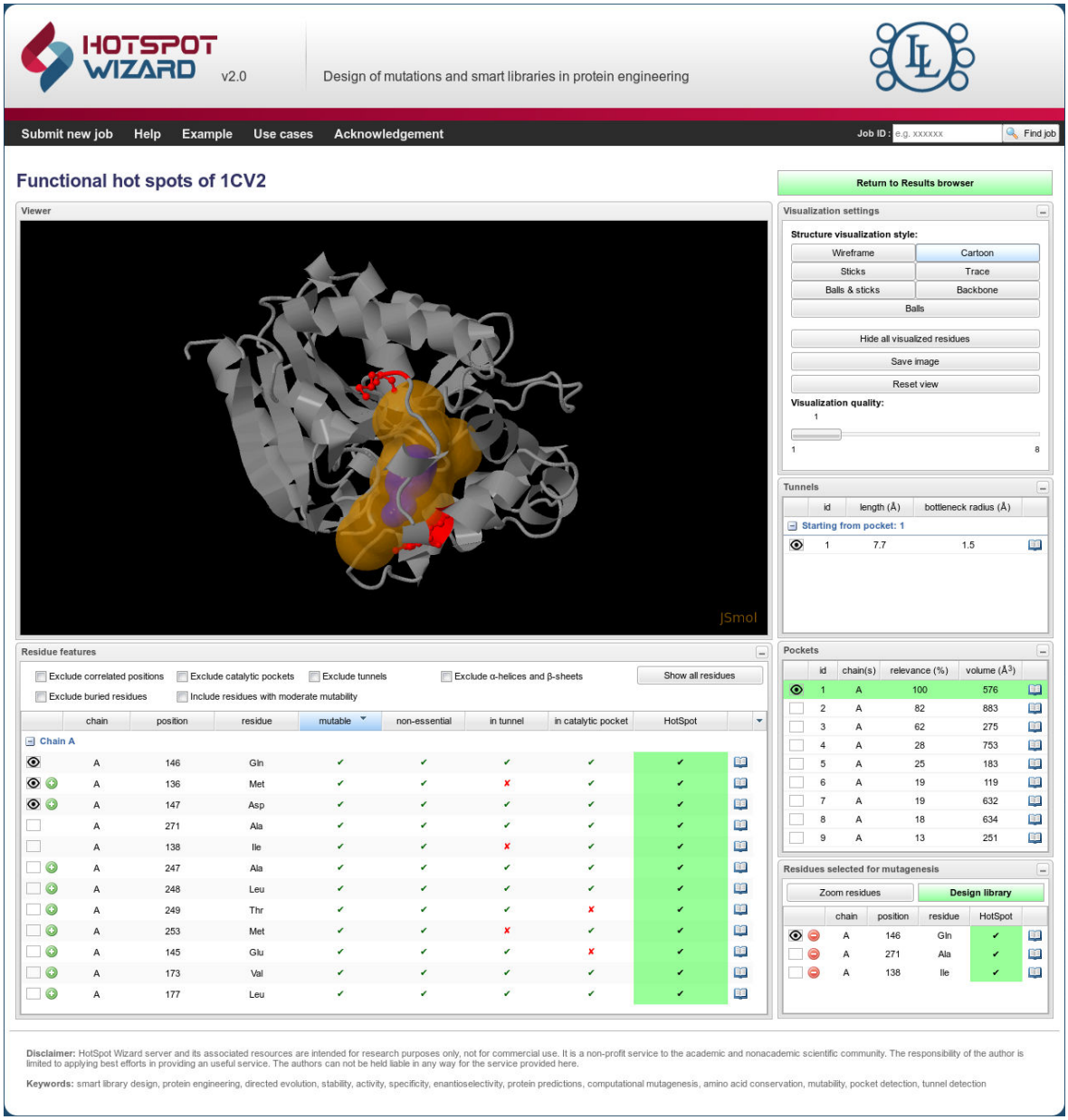
Obrázek 6.5: Uživatelské rozhraní shrnutí úlohy.

- rozcestník strategií, pomocí kterého je možné zobrazit výsledky analýzy dat pomocí vybrané strategie a
- panel stahování, kde je možné stáhnout archiv obsahující výstupní soubory jednotlivých nástrojů a výsledkový XML soubor.

Tato stránka však nabývá i mírně odlišných podob v závislosti na aktuálním stavu celé úlohy. V případě nedokončené úlohy se pochopitelně nezobrazuje rozcestník strategií a panel stahování, ale naopak se zobrazuje aktuální délka úlohy. Pokud se nepodaří dokončit některou ze strategií, rozcestník se zobrazí, ale odkazy na výstupní analýzy jsou neaktivní a nelze s nimi dále pracovat. U neúspěšně dokončené úlohy se pak nezobrazuje pouze rozcestník strategií – panel stahování je zachován pro případ, že by se uživatel chtěl podívat na dosavadní výstupy analýz (včetně výstupu modulu, který selhal).

Analýza výsledků

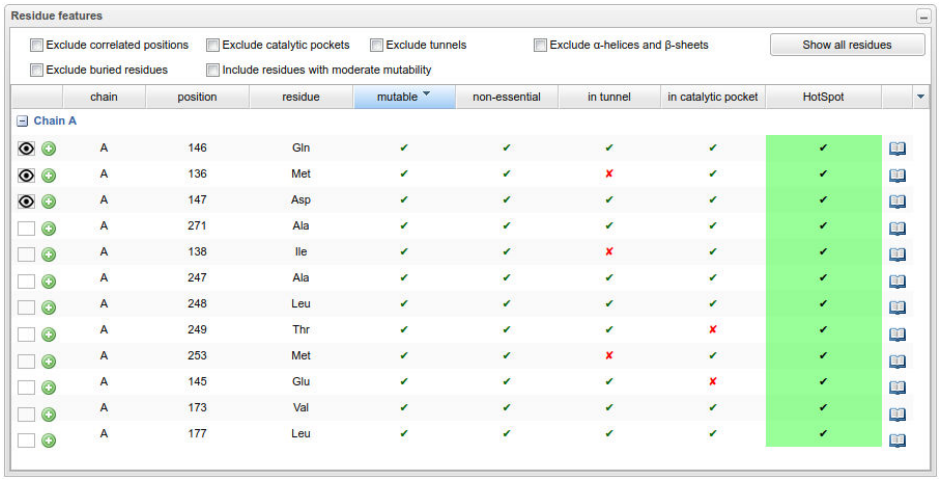
Klíčovou stránkou celého rozhraní je analýza výsledků vybranou strategií. Ačkoliv má každá strategie mírně odlišný způsob prezentace, základní rozložení je zachováno. Stránce tedy dominuje především panel vizualizace struktury a také přehledová tabulka výsledků. Vizualizace je realizována s využitím aplikace JSmol [88] a je velmi úzce a interaktivně propojená s ostatními komponentami na stránce. Díky tomu může uživatel okamžitě vizuálně zhodnotit



Obrázek 6.6: Uživatelské rozhraní pro analýzu výsledků strategií funkčních hot spotů.

umístění vybraného rezidua nebo i geometrii katalytických kapes a tunelů bez nutnosti otevírat strukturu v externím vizualizačním programu.

Druhou dominantou stránky je pak tabulka výsledků (obrázek 6.7) umístěná pod vizualizačním oknem. Její sloupce se mění v závislosti na vybrané strategii a ve výchozí podobě obsahuje pouze informace o nalezených hot spotech. Tyto hot spoty je dále možné filtrovat s využitím předpřipravených filtrů na několik strukturních charakteristik reziduí. Pokud by uživateli nestačil přehled hot spotů, má možnost přepnout do plného zobrazení všech reziduí proteinu. V neposlední řadě je možné rezidua vybírat pro vizualizaci (ikona oka), pro návrh chytrých knihoven (ikona plus) nebo zobrazit jejich detailní informace (ikona knihy).



Obrázek 6.7: Tabulka výsledků pro vybrané hot spoty.

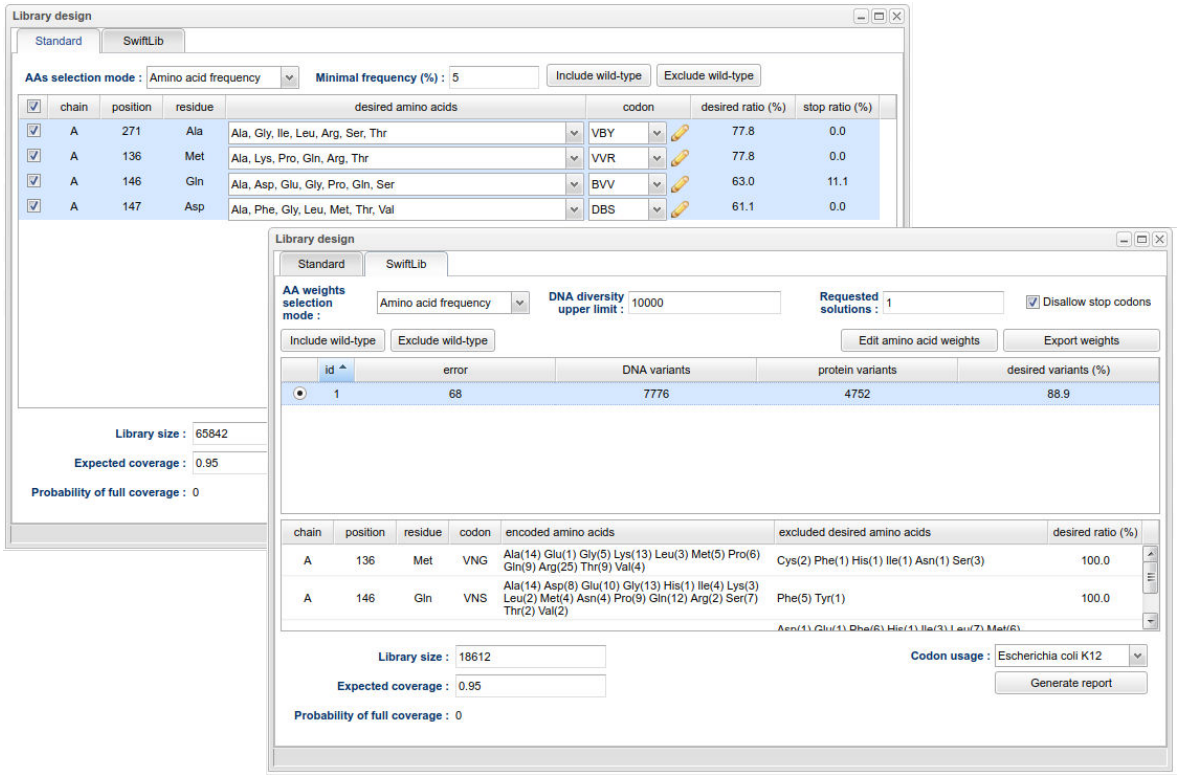
Okno detailních informací shrnuje všechny známé informace získané v průběhu výpočtu o daném reziduu bez ohledu na zvolenou strategii a poskytuje tak uživateli snadný způsob globální analýzy rezidua bez nutnosti přepínání strategií. Dále se na stránce nachází panel určený k manipulaci s vizualizací, panel se seznamem nalezených katalytických kapes, panel se seznamem nalezených tunelů a jejich asociacemi ke kapsám a také panel obsahující seznam reziduí vybraných pro návrh chytrých knihoven.

Návrh chytrých knihoven

Po dokončení výběru slibných reziduí, je možné otevřít okno určené pro návrh chytrých knihoven. Toto okno je rozděleno do dvou záložek podle použité strategie.

První „standardní“ záložka (obrázek 6.8) umožňuje výběr cílových aminokyselin na základě několika přístupů jako například jejich frekvence v mnohonásobném zarovnání nebo predikovaná patogenicita s nastavitelnými prahy. V případě zájmu lze samozřejmě zvolit i manuální režim, kde probíhá výběr aminokyselin ručně. Pro vybrané aminokyseliny je automaticky nalezen vhodný degenerovaný kodon s možností jeho snadné záměny výběrem z nabídky nebo ručním zadáním. S každou provedenou změnou se také automaticky přepočítávají statistiky knihovny. Druhá záložka (obrázek 6.8) nabízí odlišný pohled na návrh knihoven reprezentovaný metodou SwiftLib, kdy uživatel nezadává přímo cílové aminokyseliny, ale pouze jejich váhy, a především pak maximální velikost knihoven. Algoritmus metody pak automaticky nalezne vhodné degenerované kodony tak, aby upřednostnil aminokyseliny s vyšší váhou a zároveň splnil kritérium velikosti. Návrh tohoto rozhraní byl z pragmatických důvodů do velké míry inspirován původním rozhraním metody SwiftLib.

Na úplný závěr je možné z obou záložek provést export navržené knihovny. Tento export je proveden na základě zvolené tabulky použití kodonů charakteristické pro vybraný expresní systém (výchozí je bakteriální systém *Escherichia Coli*) a obsahuje informace o úloze, vybraných mutacích a degenerovaných kodonech a také zpětně přeložené sekvence genů vygenerované nástrojem backtranseq z balíku European Molecular Biology Open Software Suite (EMBOSS) [89], které je možné přímo využít pro objednávku knihovny. Reporty jsou k dispozici ve formá-



Obrázek 6.8: Rozhraní pro návrh chytrých knihoven. V pozadí vlevo je zobrazen standardní mód a v popředí vpravo pak metoda SwiftLib.

tech HTML a PDF generované programem LaTeX a každý z nich dostává unikátní permanentní odkaz, aby se bylo možné na něj kdykoliv vrátit.

Návod a tutoriál

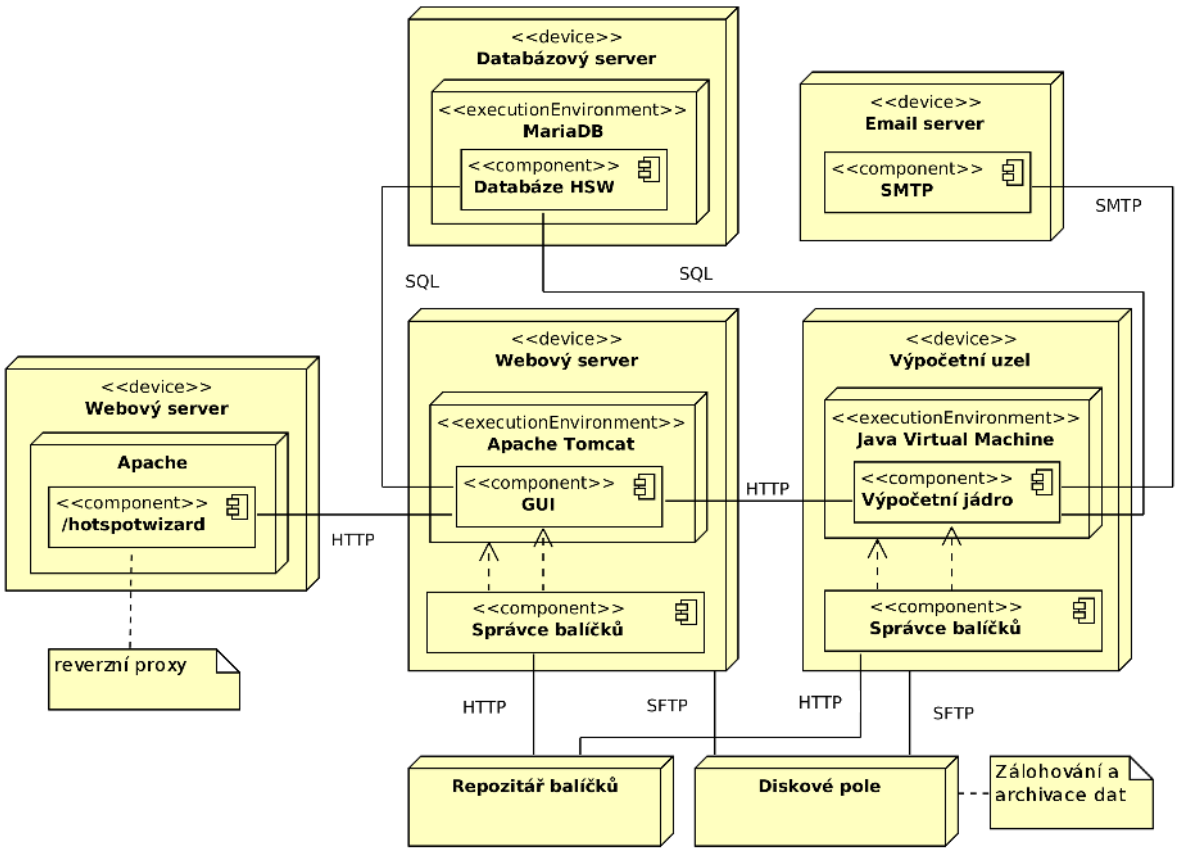
Z dalších stránek stojí za zmínku především Help a Example. První jmenovaná je cílena na poskytnutí všeobecných informací o nástroji a zejména pak vstupních parametrech a jejich významu. Stránka Example je psána formou tutoriálu, kdy je detailně rozebrána a vysvětlena analýza výsledků pomocí všech strategií včetně návrhu knihovny na modelovém příkladu mikrobiálního enzymu LinB z rodiny halogenalkandehalogenáz.

Další stránky

Kromě dříve zmíněných obsahuje portál také stránku shrnující vybrané publikace, v nichž byl nástroj využit, a také stránku s informacemi o integrovaných metodách a databázích s krátkým popisem a odkazy na jejich domovské stránky.

6.7 Nasazení

Jako operační systém pro produkční prostředí byl zvolen Debian 8, který patří mezi konzervativní distribuce s delší podporou a zároveň poskytuje velký počet předpřipravených aplikací



Obrázek 6.9: Diagram aktuálního nasazení produkční verze nástroje HotSpot Wizard.

včetně mnoha bioinformatických nástrojů. Pro zajištění snadného a bezpečného nasazování novějších verzí, byl jako hlavní distribuční formát zvolen standardní formát balíčkovacího systému distribuce Debian. Díky možnostem, které tento systém poskytuje tak není nutné řešit vedlejší efekty (jako chybná knihovna nebo špatné nastavení) způsobené ručním nasazováním nových verzí. Pro účel distribuce byly samotné binární soubory nástroje doplněny o další pomocné skripty zajišťující automatické spouštění výpočetního jádra jako systémové služby nebo další údržbu během nasazování nové verze. Z globálního pohledu (obrázek 6.9) pak byla navržená architektura rozšířena především o reverzní proxy, která poskytuje možnost provozovat jednotlivé služby na různých strojích a také jejich snadnou migraci.

7 Validace

K získání důvěry uživatelů bylo nutné demonstrovat, že je nástroj schopen nalézt relevantní hot spoty pro experimenty řízené evoluce. Za tímto účelem byly pro každou ze čtyř implementovaných strategií proteinového inženýrství nalezené vědecké studie, které návrhem experimentu rámcově odpovídaly implementaci v HotSpot Wizardu. Cílem bylo ověřit, zda je HotSpot Wizard schopen identifikovat ty kombinace reziduí, které byly v daných studiích vyhodnoceny jako nejvíce relevantní, tj. zvyšující aktivitu, stabilitu nebo měnící substrátovou specificitu. Ačkoliv málokdy bylo vytipováno přesně tolik hot spotů, kolik bylo experimentálně validováno, neznamená to nutně špatnou predikci přebývajících hot spotů, ale pouze fakt, že dané reziduum nebylo autory studie vybráno pro ohodnocení, a tudíž k němu nejsou dostupná experimentální data.

7.1 Funkční hot spoty

Halogenalkandehalogenáza

Halogenalkandehalogenázy jsou rodina mikrobiálních enzymů katalyzujících hydrolitický rozklad haloalkanů na primární alkohol a halogenový aniont. V mikroorganismech bývají součástí několika degradačních metabolických drah a v průmyslu jsou využitelné jako biodegradátory [90].

Pro validaci byl využit protein s PDB kódem `1BN6`, který pochází z organismu kmene *Rhodococcus* a byla u něj vylepšována aktivita. HotSpot Wizard identifikoval 15 hot spotů, z nichž k šesti existuje experimentální ověření (obrázek 7.1a) [91]–[93].

Fosfotriesteráza

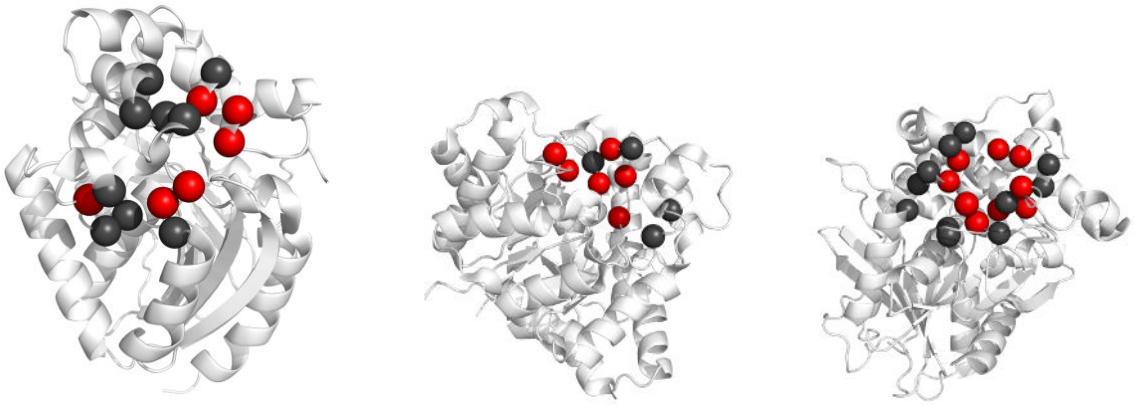
Fosfotriesterázy patří mezi bakteriální rodinu enzymů schopných degradovat velmi toxické fosfotriestery. Z pohledu proteinového inženýrství je tento enzym zajímavý především kvůli svému potenciálnímu využití pro degradaci bojových látek jako je VX, soman a sarin nebo pro degradaci pesticidů jako je parathion [94].

Pro validaci byl využit protein s PDB kódem `1IOD`, který pochází z organismu *Pseudomonas diminuta* a byla u něj modifikována aktivita a enantioselektivita. HotSpot Wizard identifikoval 10 hot spotů, z nichž k šesti existuje experimentální ověření (obrázek 7.1b) [95]–[98].

Epoxid hydroláza

Epoxid hydrolázy jsou třídou enzymů důležitých pro detoxifikaci genotoxických látek. Tyto látky dokáží narušovat genetickou informaci v buňkách nebo blokovat její přirozenou schopnost samoopravy a tak způsobit nezvratné poškození nebo i rozvoj rakoviny [99].

Pro validaci byl využit protein s PDB kódem `1Q07`, který pochází z organismu *Aspergillus niger* a byla u něj modifikována enantioselektivita. HotSpot Wizard identifikoval čtyři hot spoty, z nichž k jednomu existuje experimentální ověření (obrázek 7.1c) [62], [100].



(a) Halogenalkandehalogenáza **(b)** Fosfotriesteráza **(c)** Epoxid hydroláza

Obrázek 7.1: Proteiny s vyznačenými ověřenými (červeně) a dalšími (černě) funkčními hot spoty.

7.2 Stabilitní hot spoty (flexibilita)

Sucrose fosforyláza

Tento enzym hraje důležitou roli v metabolismu sacharózy a regulaci dalších metabolických meziproduktů. V jeho reakčním mechanismu dochází k navázání sacharózy na enzym a její konverzi na fruktózu [101].

Pro validaci byl využit protein s PDB kódem `1R7A`, který pochází z organismu *Bifidobacterium adolescentis*. HotSpot Wizard identifikoval 10 hot spotů, z nichž ke dvěma existuje experimentální ověření (obrázek 7.2a) [102].

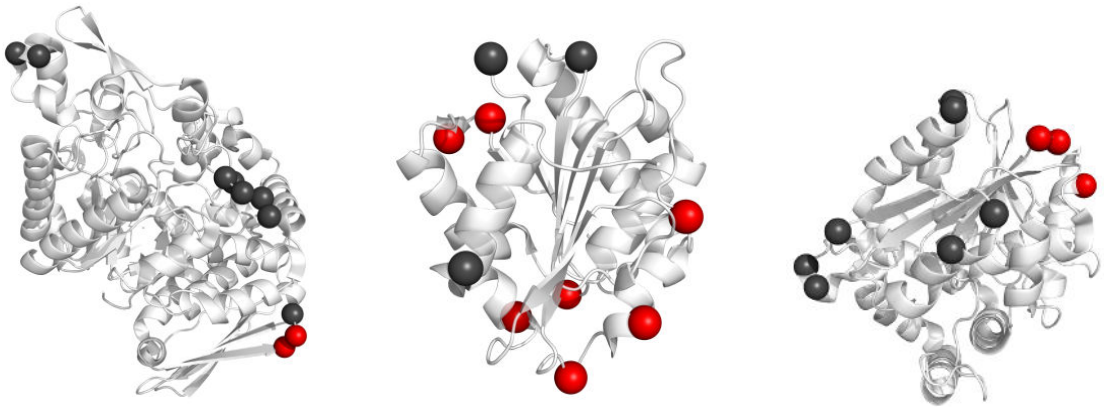
Lipáza

Lipázy jsou skupina enzymů schopných katalyzovat hydrolýzu tuků. Tato schopnost je esenciální pro většinu organismů a tak se lipázy významně podílí na celém procesu trávení, transportování a zpracování tuků nebo olejů. V průmyslu jsou lipázy hojně využívány v procesech spojených s pekařstvím, výrobou pracích prášků nebo dokonce při konvertování rostlinných olejů na palivo [103]. V lékářství jsou pak využívány například při léčbě metodou PERT (pancreatic enzyme replacement therapy) [104].

Pro validaci byl využit protein s PDB kódem `1ISP`, který pochází z organismu *Bacillus subtilis*. HotSpot Wizard identifikoval 10 hot spotů, z nichž k sedmi existuje experimentální ověření (obrázek 7.2b) [102].

Esteráza

Esterázy jsou enzymy schopné hydrolyticky štěpit estery na kyselinu a alkohol. V přírodě existuje mnoho různých esteráz podílejících se na odlišných reakcích. Pro validaci byla vybrána varianta štěpící fenylacetát.



(a) Sucrose fosforyláza **(b)** Lipáza **(c)** Esteráza

Obrázek 7.2: Proteiny s vyznačenými ověřenými (červeně) a dalšími (černě) flexibilitními hot spoty.

Pro validaci byl využit protein s PDB kódem `1VA4`, který pochází z organismu *Pseudomonas fluorescens*. HotSpot Wizard identifikoval 10 hot spotů, z nichž ke třem existuje experimentální ověření (obrázek 7.2c) [105].

7.3 Stabilitní hot spoty (back-to-consensus)

Triosephosphate isomeráza

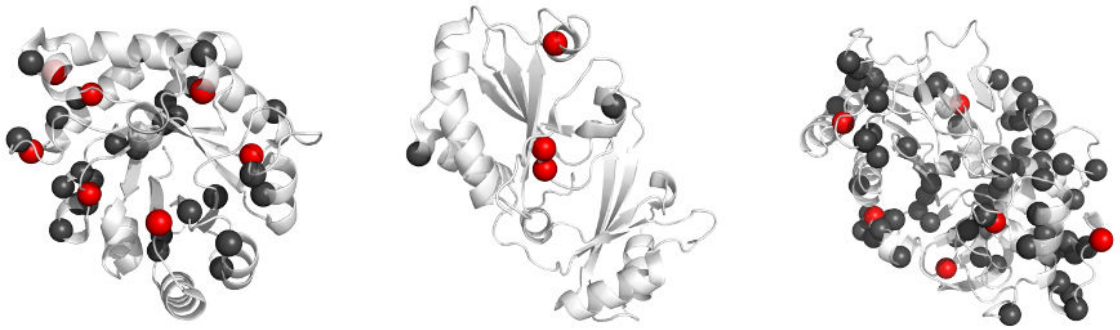
Triosephosphate isomerázy hrají důležitou roli v mechanismu glykolýzy a jsou důležité i pro efektivní výrobu energie v buňce. Tyto enzymy je možné nalézt prakticky ve všech organismech včetně savců, hmyzu, hub, rostlin nebo bakterií. Zajímavostí je, že katalytická efektivita těchto enzymů je považována za perfektní – je totiž limitována pouze rychlostí, s níž se substrát dokáže pohybovat dovnitř a ven z aktivního místa [106].

Pro validaci byl využit protein s PDB kódem `1YPI`, který pochází z organismu *Saccharomyces cerevisiae*. HotSpot Wizard identifikoval 37 hot spotů, z nichž k sedmi existuje experimentální ověření (obrázek 7.3a) [64].

Thioredoxin

Thireoxiny jsou proteiny sloužící jako antioxidanty schopné katalyzovat redukci ostatních proteinů pomocí cystein thiol-disulfidické výměny. Nacházejí se prakticky ve všech známých organismech a účastní se mnoha důležitých procesů, jako je redoxní signalizace. U člověka jsou tyto proteiny čím dále častěji spojovány s lékařstvím díky svým reakcím s látkami obsahujícími volné radikály kyslíku [107].

Pro validaci byl využit protein s PDB kódem `3DXB`, který pochází z organismu *Escherichia coli*. HotSpot Wizard identifikoval pět hot spotů, z nichž ke třem existuje experimentální ověření (obrázek 7.3b) [65].



(a) Triosephosphate isomeráza **(b)** Thioredoxin **(c)** β-laktamáza

Obrázek 7.3: Proteiny s vyznačenými ověřenými (červeně) a dalšími (černě) konsensuálními hot spoty.

β-laktamáza

β-laktamázy jsou enzymy produkované některými druhy bakterií a jsou zodpovědná za jejich rezistenci vůči beta-laktamovým antibiotikům jakou jsou peniciliny a cefamyciny. Tyto enzymy dokáží otevírat čtyřatomový kruh (zvaný beta-laktam) antibiotik a tím deaktivovat jejich antibakteriální vlastnosti [108].

Pro validaci byl využit protein s PDB kódem `1BLS`, který pochází z organismu *Enterobacter cloacae*. HotSpot Wizard identifikoval pět hot spotů, z nichž ke třem existuje experimentální ověření (obrázek 7.3c) [63].

7.4 Korelované hot spoty

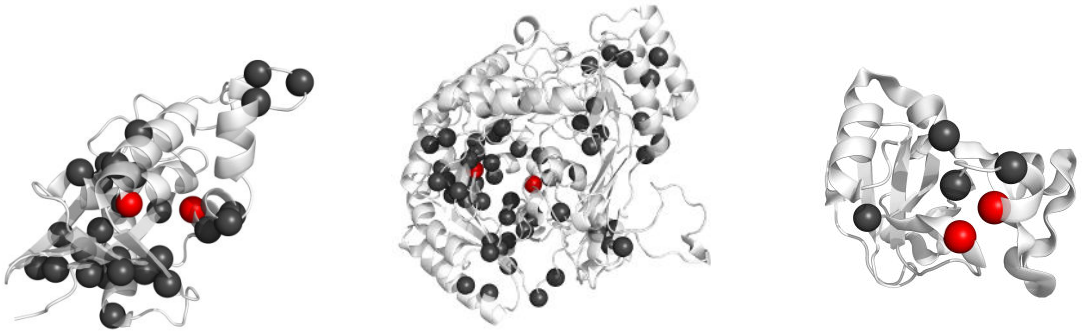
Ribonukleáza

Ribonukleázy dokáží hydrolyticky štěpit fosfodiesterovou vazbu v rámci ribonukleové kyseliny. V organismech plní tyto enzymy širokou škálu funkcí – od sestřihů pre-rRNA, přes interferonové imutní odpovědi až po rozklad ribonukleových kyselin přítomných v potravě. Ve výzkumu jsou často využívány k odstraňování přebytečného RNA při práci s DNA [109].

Pro validaci byl využit protein s PDB kódem `2RN2`, který pochází z organismu *Escherichia coli*. HotSpot Wizard identifikoval 29 hot spotů, z nichž ke dvěma (třináctý pár reziduí z 20 korelovaných párů) existuje experimentální ověření (obrázek 7.4a) [110].

Vanyllyl-alkohol oxidáza

Enzym patřící do skupiny oxidoreduktáz schopný katalyzovat reverzibilní chemickou reakci konverze vanillyl alkoholu s kyslíkem na vanillin a peroxid vodíku. Bohužel její význam stále není plně pochopen a popsán [111].



(a) Ribonukleáza **(b)** Vanyllyl-alkohol oxidáza **(c)** Barnáza

Obrázek 7.4: Proteiny s vyznačenými ověřenými (červeně) a dalšími (černě) korelovanými hot spoty.

Pro validaci byl využit protein s PDB kódem `1AHU`, který pochází z organismu *Penicillium simplicissimum*. HotSpot Wizard identifikoval 57 hot spotů, z nichž ke dvěma (třicátý devátý pár reziduí z 85 korelovaných párů) existuje experimentální ověření (obrázek 7.4b) [112].

Barnáza

Barnáza je bakteriální protein vykazující ribonukleázovou aktivitu. Jejím specifikem je, že musí být exprimována společně se svým inhibitorem, jinak je velmi nebezpečná pro všechnu syntetizovanou RNA v buňce. Její komplex s inhibitorem je netradičně úzký [113].

Pro validaci byl využit protein s PDB kódem `1BRS`, který pochází z organismu *Bacillus amyloliquefaciens*. HotSpot Wizard identifikoval šest hot spotů, z nichž ke dvěma (první pár reziduí ze 14 korelovaných párů) existuje experimentální ověření (obrázek 7.4c) [114].

8 Závěr

Hlavním cílem této práce byl návrh a vývoj nové verze webového portálu HotSpot Wizard určeného pro automatickou identifikaci mutačních hot spotů v proteinech. Oproti původní verzi byly kromě dílčích zlepšení stávajících analýz integrovány další tři populární inženýrské strategie, které výrazně rozšiřují oblast použití nástroje a umožňují se uživateli zaměřit na jiné zajímavé vlastnosti proteinů. V neposlední řadě byla přidána možnost přímého návrhu chytrých knihoven na základě identifikovaných hot spotů. Podobně jako původní verze nástroj cílí primárně na uživatele bez větších znalostí bioinformatiky, a proto kromě intuitivního uživatelského rozhraní byl kladen důraz také na vhodnou volbu výchozího nastavení jednotlivých analýz. Za tímto účelem bylo provedeno několik rozsáhlých bioinformatických analýz zaměřených na jejich optimalizaci tak, aby se zajistila vysoká robustnost výpočtu a univerzální použitelnost portálu pro široké spektrum běžně studovaných proteinů. Celý portál je postaven na platformě Loschmidt Core, která implicitně zajišťuje většinu režijních úkonů a poskytuje tak solidní a robustní základ pro celé výpočetní jádro. Dále bylo kompletně přepracováno uživatelské grafické rozhraní, které je nyní více interaktivní a poskytuje komfortnější způsob analýzy nalezených hot spotů. Relevance výsledků pro jednotlivé strategie byla validována na základě vybraných příkladů z literatury s dostupnými experimentálními daty. Nástroj HotSpot Wizard vytvořený v rámci této práce byl publikován v prestižním mezinárodním časopise Nucleic Acids Research (příloha B) a od svého vydání v červenci 2016 byl již dle Web of Science několikrát citován v časopisech z prvního kvartilu (Q1) IF. Celkově jej navštívilo více než 2 700 unikátních uživatelů z celého světa, kteří s jeho pomocí analyzovali 1 524 různých proteinových struktur. Vývoj nástroje i nadále pokračuje a v blízké budoucnosti do něj bude integrováno několik nových funkcí. První z nich je přidání podpory pro konstrukci konkrétních mnohonásobných mutantů a jejich energetického ohodnocení s využitím nástrojů FoldX [115] a Rosetta [116]. Dalším významným vylepšením bude integrace metod homologního modelování, díky němuž bude možné začínat úlohy přímo ze sekvence bez znalosti struktury.

Literatura

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts a P. Walter, *Molecular Biology of the Cell*, 6. vyd. New York: Garland Science, 2002, 1465 s., ISBN: 0-8153-3218-1; 0-8153-4072-9.

[2] J. M. Berg, J. L. Tymoczko a L. Stryer, *Biochemistry*, 5. vyd. New York: W H Freeman, 2002, ISBN: 0-7167-3051-0.

[3] Villarreal, Mariana Ruiz, *Main protein structures levels*, 2008. WWW: `https://en.wikipedia.org/wiki/File:Main_protein_structure_levels_en.svg` (cit. 14. 05. 2017).

[4] U. T. Bornscheuer, G. W. Huisman, R. J. Kazlauskas, S. Lutz, J. C. Moore a K. Robins, "Engineering the third wave of biocatalysis", *Nature*, roč. 485, č. 7397, s. 185–194, květ. 2012. DOI: `10.1038/nature11117`.

[5] Z. Prokop a J. Damborsky, *Srovnání metod proteinového inženýrství*, 2009.

[6] D. Shortle, D. DiMaio a D. Nathans, "Directed Mutagenesis", *Annual Review of Genetics*, roč. 15, s. 265–294, 1981. DOI: `10.1146/annurev.ge.15.120181.001405`.

[7] M. T. Reetz, *Directed Evolution of Selective Enzymes: Catalysts for Organic Chemistry and Biotechnology*, 1. vyd. Weinheim: Wiley-VCH, 2016, ISBN: 978-3-527-31660-1.

[8] A. Pavelka, E. Chovancova a J. Damborsky, "HotSpot Wizard: a web server for identification of hot spots in protein engineering", *Nucleic Acids Research*, roč. 37, W376, 2009. DOI: `10.1093/nar/gkp410`.

[9] Z.-R. Xie a M.-J. Hwang, "Methods for Predicting Protein–Ligand Binding Sites", in *Molecular Modeling of Proteins*, A. Kukol, ed. New York, NY: Springer New York, 2015, s. 383–398, ISBN: 978-1-4939-1465-4. DOI: `10.1007/978-1-4939-1465-4_17`.

[10] Y. Yaxia, P. Jianfeng a L. Luhua, "Binding Site Detection and Druggability Prediction of Protein Targets for Structure- Based Drug Design", *Current Pharmaceutical Design*, roč. 19, č. 12, s. 2326–2333, 2013. DOI: `10.2174/1381612811319120019`.

[11] A. Lavecchia a C. Di Giovanni, "Virtual Screening Strategies in Drug Discovery: A Critical Review", *Current Medicinal Chemistry*, roč. 20, č. 23, s. 2839–2860, 2013. DOI: `10.2174/09298673113209990001`.

[12] E. Sebestova, J. Bendl, J. Brezovsky a J. Damborsky, "Computational Tools for Designing Smart Libraries", in *Directed Evolution Library Creation: Methods and Protocols*, E. M. Gillam, J. N. Copp a D. Ackerley, ed. New York, NY: Springer New York, 2014, s. 291–314, ISBN: 978-1-4939-1053-3. DOI: `10.1007/978-1-4939-1053-3_20`.

[13] J. Brezovsky, E. Chovancova, A. Gora, A. Pavelka, L. Biedermannova a J. Damborsky, "Software tools for identification, visualization and analysis of protein tunnels and channels", *Biotechnology Advances*, roč. 31, č. 1, s. 38–49, 2013. DOI: `10.1016/j.biotechadv.2012.02.002`.

[14] Z. Zhang, Y. Li, B. Lin, M. Schroeder a B. Huang, "Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction", *Bioinformatics*, roč. 27, č. 15, s. 2083, 2011. DOI: 10.1093/bioinformatics/btr331.

[15] C. Ó'Fágáin, "Engineering Protein Stability", in *Protein Chromatography: Methods and Protocols*, D. Walls a S. T. Loughran, ed. Totowa, NJ: Humana Press, 2011, s. 103–136, ISBN: 978-1-60761-913-0. DOI: 10.1007/978-1-60761-913-0_7.

[16] A. S. Bommarius a M. F. Paye, "Stabilizing biocatalysts", *Chem. Soc. Rev.*, roč. 42, s. 6534–6565, 15 2013. DOI: 10.1039/C3CS60137D.

[17] H. J. Wijma, R. J. Floor a D. B. Janssen, "Structure- and sequence-analysis inspired engineering of proteins for enhanced thermostability", *Current Opinion in Structural Biology*, roč. 23, č. 4, s. 588–594, 2013. DOI: 10.1016/j.sbi.2013.04.008.

[18] H. Yu a H. Huang, "Engineering proteins for thermostability through rigidifying flexible sites", *Biotechnology Advances*, roč. 32, č. 2, s. 308–315, 2014. DOI: 10.1016/j.biotechadv.2013.10.012.

[19] H. Berman, K. Henrick a H. Nakamura, "Announcing the worldwide Protein Data Bank", *Nat Struct Mol Biol*, roč. 10, č. 12, s. 980–980, pros. 2003. DOI: 10.1038/nsb1203-980.

[20] wwPDB, *Protein Data Bank Contents Guide: Atomic Coordinate Entry Format Description*, English, ver. Version 3.3, wwPDB, 21. lis. 2012, 194 s., 2012-11-21.

[21] E. Krissinel a K. Henrick, "Inference of Macromolecular Assemblies from Crystalline State", *Journal of Molecular Biology*, roč. 372, č. 3, s. 774–797, 2007. DOI: 10.1016/j.jmb.2007.05.022.

[22] RCSB, *Introduction to Biological Assemblies and the PDB Archive*. WWW: https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/biological-assemblies (cit. 14. 05. 2017).

[23] M. Palmer, *MakeMultimer.py*. WWW: http://watcut.uwaterloo.ca/tools/makemultimer/ (cit. 14. 05. 2017).

[24] RCSB, *Primary Sequences and the PDB Format*. WWW: https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/primary-sequences-and-the-pdb-format (cit. 14. 05. 2017).

[25] N. Furnham, G. L. Holliday, T. A. P. de Beer, J. O. B. Jacobsen, W. R. Pearson a J. M. Thornton, "The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes", *Nucleic Acids Research*, roč. 42, č. D1, s. D485, 2014. DOI: 10.1093/nar/gkt1243.

[26] U. Consortium, "UniProt: the universal protein knowledgebase", *Nucleic Acids Research*, roč. 45, č. D1, s. D158, 2017. DOI: 10.1093/nar/gkw1099.

[27] W. Kabsch a C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features", *Biopolymers*, roč. 22, č. 12, s. 2577–2637, 1983. DOI: 10.1002/bip.360221211.

[28] A. L. Morris, M. W. MacArthur, E. G. Hutchinson a J. M. Thornton, "Stereochemical quality of protein structure coordinates", *Proteins: Structure, Function, and Bioinformatics*, roč. 12, č. 4, s. 345–364, 1992. DOI: 10.1002/prot.340120407.

[29] S. Parthasarathy a M. Murthy, "Analysis of temperature factor distribution in high-resolution protein structures", *Protein Science*, roč. 6, č. 12, s. 2561–2567, 1997. DOI: 10.1002/pro.5560061208.

[30] A. Shrake a J. Rupley, "Environment and exposure to solvent of protein atoms. Lysozyme and insulin", *Journal of Molecular Biology*, roč. 79, č. 2, s. 351–371, 1973. DOI: 10.1016/0022-2836(73)90011-9.

[31] S. Chakravarty a R. Varadarajan, "Residue depth: a novel parameter for the analysis of protein structure and stability", *Structure*, roč. 7, č. 7, s. 723–732, 1999.

[32] D. Mount, *Bioinformatics: Sequence and Genome Analysis*, 2. vyd. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 2004, ISBN: 978-0879697129.

[33] B. E. Suzek, H. Huang, P. McGarvey, R. Mazumder a C. H. Wu, "UniRef: comprehensive and non-redundant UniProt reference clusters", *Bioinformatics*, roč. 23, č. 10, s. 1282, 2007. DOI: 10.1093/bioinformatics/btm098.

[34] L. Fu, B. Niu, Z. Zhu, S. Wu a W. Li, "CD-HIT: accelerated for clustering the next-generation sequencing data", *Bioinformatics*, roč. 28, č. 23, s. 3150, 2012. DOI: 10.1093/bioinformatics/bts565.

[35] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer a T. L. Madden, "BLAST+: architecture and applications", *BMC Bioinformatics*, roč. 10, č. 1, s. 421, 2009. DOI: 10.1186/1471-2105-10-421.

[36] R. C. Edgar, "Search and clustering orders of magnitude faster than BLAST", *Bioinformatics*, roč. 26, č. 19, s. 2460, 2010. DOI: 10.1093/bioinformatics/btq461.

[37] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson a D. G. Higgins, "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega", *Molecular Systems Biology*, roč. 7, č. 1, 2011. DOI: 10.1038/msb.2011.75.

[38] M. N. Price, P. S. Dehal a A. P. Arkin, "FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments", *PLOS ONE*, roč. 5, č. 3, s. 1–10, břez. 2010. DOI: 10.1371/journal.pone.0009490.

[39] W. S. Valdar, "Scoring residue conservation", *Proteins: Structure, Function, and Bioinformatics*, roč. 48, č. 2, s. 227–241, 2002. DOI: 10.1002/prot.10146.

[40] J. A. Capra a M. Singh, "Predicting functionally important residues from sequence conservation", *Bioinformatics*, roč. 23, č. 15, s. 1875, 2007. DOI: 10.1093/bioinformatics/btm270.

[41] T. Pupko, R. E. Bell, I. Mayrose, F. Glaser a N. Ben-Tal, "Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues", *Bioinformatics*, roč. 18, S71, 2002. DOI: 10.1093/bioinformatics/18.suppl_1.S71.

[42] M. Lehmann, C. Loch, A. Middendorf, D. Studer, S. F. Lassen, L. Pasamontes, A. P. van Loon a M. Wyss, "The consensus concept for thermostability engineering of proteins: further proof of concept", *Protein Engineering, Design and Selection*, roč. 15, č. 5, s. 403, 2002. DOI: 10.1093/protein/15.5.403.

[43] P. C. Ng a S. Henikoff, "Predicting the Effects of Amino Acid Substitutions on Protein Function", *Annual Review of Genomics and Human Genetics*, roč. 7, s. 61–80, 2006. DOI: 10.1146/annurev.genom.7.080505.115630.

[44] M. Musil, "RAPHYD: Predictor of the Effect of Amino Acid Substitutions on Protein Function.", in *Studentská konference Excel@FIT*, FIT VUT Brno, 2016.

[45] E. A. Stone a A. Sidow, "Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity", *Genome Research*, roč. 15, č. 7, s. 978–986, 2005. DOI: 10.1101/gr.3804205.

[46] U. Göbel, C. Sander, R. Schneider a A. Valencia, "Correlated mutations and residue contacts in proteins", *Proteins: Structure, Function, and Bioinformatics*, roč. 18, č. 4, s. 309–317, 1994. DOI: 10.1002/prot.340180402.

[47] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch a T. Hwa, "Identification of direct residue contacts in protein–protein interaction by message passing", *Proceedings of the National Academy of Sciences*, roč. 106, č. 1, s. 67–72, 2009. DOI: 10.1073/pnas.0805923106.

[48] J. P. Dekker, A. Fodor, R. W. Aldrich a G. Yellen, "A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments", *Bioinformatics*, roč. 20, č. 10, s. 1565, 2004. DOI: 10.1093/bioinformatics/bth128.

[49] O. Olmea, B. Rost a A. Valencia, "Effective use of sequence correlation and conservation in fold recognition1", *Journal of Molecular Biology*, roč. 293, č. 5, s. 1221–1239, 1999. DOI: 10.1006/jmbi.1999.3208.

[50] B. T. Korber, R. M. Farber, D. H. Wolpert a A. S. Lapedes, "Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis", *Proceedings of the National Academy of Sciences*, roč. 90, č. 15, s. 7176–7180, 1993.

[51] B.-C. Lee a D. Kim, "A new method for revealing correlated mutations under the structural and functional constraints in proteins", *Bioinformatics*, roč. 25, č. 19, s. 2506, 2009. DOI: 10.1093/bioinformatics/btp455.

[52] I. Kass a A. Horovitz, "Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations", *Proteins: Structure, Function, and Bioinformatics*, roč. 48, č. 4, s. 611–617, 2002. DOI: 10.1002/prot.10180.

[53] S. W. Lockless a R. Ranganathan, "Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families", *Science*, roč. 286, č. 5438, s. 295–299, 1999. DOI: 10.1126/science.286.5438.295.

[54] R. G. Coleman a K. A. Sharp, "Protein Pockets: Inventory, Shape, and Comparison", *Journal of Chemical Information and Modeling*, roč. 50, č. 4, s. 589–603, 2010. DOI: 10.1021/ci900397t.

[55] A. O. Belduz, E. J. Lee a J. G. Harman, "Mutagenesis of the cyclic AMP receptor protein of Escherichia coli targeting positions 72 and 82 of the cyclic nucleotide binding pocket", *Nucleic Acids Research*, roč. 21, č. 8, s. 1827, 1993. DOI: 10.1093/nar/21.8.1827.

[56] F. A. Dick, E. Sailhamer a N. J. Dyson, "Mutagenesis of the pRB Pocket Reveals that Cell Cycle Arrest Functions Are Separable from Binding to Viral Oncoproteins", *Molecular and Cellular Biology*, roč. 20, č. 10, s. 3715–3727, 2000. DOI: 10.1128/MCB.20.10.3715-3727.2000.

[57] V. Le Guilloux, P. Schmidtke a P. Tuffery, "Fpocket: An open source platform for ligand pocket detection", *BMC Bioinformatics*, roč. 10, č. 1, s. 168, 2009. DOI: 10.1186/1471-2105-10-168.

[58] P. Schmidtke a X. Barril, "Understanding and Predicting Druggability. A High-Throughput Method for Detection of Drug Binding Sites", *Journal of Medicinal Chemistry*, roč. 53, č. 15, s. 5858–5867, 2010. DOI: 10.1021/jm100574m.

[59] A. Svendsen, *Bioinformatics: Sequence and Genome Analysis*, 1. vyd. Stanford, CA: Pan Stanford, 2016, ISBN: 978-981-4669-32-0.

[60] J. Damborsky a J. Brezovsky, "Computational tools for designing and engineering biocatalysts", *Current Opinion in Chemical Biology*, roč. 13, č. 1, s. 26–34, 2009. DOI: 10.1016/j.cbpa.2009.02.021.

[61] E. Chovancova, A. Pavelka, P. Benes, O. Strnad, J. Brezovsky, B. Kozlikova, A. Gora, V. Sustr, M. Klvana, P. Medek, L. Biedermannova, J. Sochor a J. Damborsky, "CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures", *PLOS Computational Biology*, roč. 8, č. 10, s. 1–12, říj. 2012. DOI: 10.1371/journal.pcbi.1002708.

[62] M. T. Reetz, J. D. Carballeira a A. Vogel, "Iterative Saturation Mutagenesis on the Basis of B Factors as a Strategy for Increasing Protein Thermostability", *Angewandte Chemie International Edition*, roč. 45, č. 46, s. 7745–7751, 2006. DOI: 10.1002/anie.200602795.

[63] N. Amin, A. Liu, S. Ramer, W. Aehle, D. Meijer, M. Metin, S. Wong, P. Gualfetti a V. Schellenberger, "Construction of stabilized proteins by combinatorial consensus mutagenesis", *Protein Engineering, Design and Selection*, roč. 17, č. 11, s. 787, 2004. DOI: 10.1093/protein/gzh091.

[64] B. J. Sullivan, T. Nguyen, V. Durani, D. Mathur, S. Rojas, M. Thomas, T. Syu a T. J. Magliery, "Stabilizing Proteins from Sequence Statistics: The Interplay of Conservation and Correlation in Triosephosphate Isomerase Stability", *Journal of Molecular Biology*, roč. 420, č. 4–5, s. 384–399, 2012. DOI: 10.1016/j.jmb.2012.04.025.

[65] A. L. Pey, D. Rodriguez-Larrea, S. Bomke, S. Dammers, R. Godoy-Ruiz, M. M. Garcia-Mira a J. M. Sanchez-Ruiz, "Engineering proteins with tunable thermodynamic and kinetic stabilities", *Proteins: Structure, Function, and Bioinformatics*, roč. 71, č. 1, s. 165–174, 2008. DOI: 10.1002/prot.21670.

[66] M. Lehmann, L. Pasamontes, S. Lassen a M. Wyss, "The consensus concept for thermostability engineering of proteins", *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology*, roč. 1543, č. 2, s. 408–415, 2000. DOI: 10.1016/S0167-4838(00)00238-7.

[67] D. de Juan, F. Pazos a A. Valencia, "Emerging methods in protein co-evolution", *Nat Rev Genet*, roč. 14, č. 4, s. 249–261, dub. 2013. DOI: 10.1038/nrg3414.

[68] R. K. P. Kuipers, H.-J. Joosten, E. Verwiel, S. Paans, J. Akerboom, J. van der Oost, N. G. H. Leferink, W. J. H. van Berkel, G. Vriend a P. J. Schaap, "Correlated mutation analyses on super-family alignments reveal functionally important residues", *Proteins: Structure, Function, and Bioinformatics*, roč. 76, č. 3, s. 608–616, 2009. DOI: 10.1002/prot.22374.

[69] A. Nobili, Y. Tao, I. V. Pavlidis, T. van den Bergh, H.-J. Joosten, T. Tan a U. T. Bornscheuer, "Simultaneous Use of in Silico Design and a Correlated Mutation Network as a Tool To Efficiently Guide Enzyme Engineering", *ChemBioChem*, roč. 16, č. 5, s. 805–810, 2015. DOI: 10.1002/cbic.201402665.

[70] C. Wang, R. Huang, B. He a Q. Du, "Improving the thermostability of alpha-amylase by combinatorial coevolving-site saturation mutagenesis", *BMC Bioinformatics*, roč. 13, č. 1, s. 263, 2012. DOI: 10.1186/1471-2105-13-263.

[71] J. D. Watson, T. A. Baker, A. Gann, M. Levine a R. Losick, *Molecular Biology of the Gene*, 7. vyd. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 2013, ISBN: 978-0-321-76243-6.

[72] T. M. Jacobs, H. Yumerefendi, B. Kuhlman a A. Leaver-Fay, "SwiftLib: rapid degenerate-codon-library optimization through dynamic programming", *Nucleic Acids Research*, roč. 43, č. 5, e34, 2015. DOI: 10.1093/nar/gku1323.

[73] Y. Nov, "Probabilistic Methods in Directed Evolution: Library Size, Mutation Rate, and Diversity", in *Directed Evolution Library Creation: Methods and Protocols*, E. M. Gillam, J. N. Copp a D. Ackerley, ed. New York, NY: Springer New York, 2014, s. 261–278, ISBN: 978-1-4939-1053-3. DOI: 10.1007/978-1-4939-1053-3_18.

[74] B. Rost, "Twilight zone of protein sequence alignments", *Protein Engineering, Design and Selection*, roč. 12, č. 2, s. 85, 1999. DOI: 10.1093/protein/12.2.85.

[75] S. B. Needleman a C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins", *Journal of Molecular Biology*, roč. 48, č. 3, s. 443–453, 1970. DOI: 0.1016/0022-2836(70)90057-4.

[76] T. A. Binkowski, S. Naghibzadeh a J. Liang, "CASTp: Computed Atlas of Surface Topography of proteins", *Nucleic Acids Research*, roč. 31, č. 13, s. 3352, 2003. DOI: 10.1093/nar/gkg512.

[77] S. Pérot, O. Sperandio, M. A. Miteva, A.-C. Camproux a B. O. Villoutreix, "Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery", *Drug Discovery Today*, roč. 15, č. 15–16, s. 656–667, 2010. DOI: 10.1016/j.drudis.2010.05.015.

[78] H. Edelsbrunner a E. P. Mücke, "Three-dimensional Alpha Shapes", *ACM Trans. Graph.*, roč. 13, č. 1, s. 43–72, led. 1994. DOI: 10.1145/174462.156635.

[79] W. Zhou a H. Yan, "Alpha shape and Delaunay triangulation in studies of protein-related interactions", *Briefings in Bioinformatics*, roč. 15, č. 1, s. 54, 2014. DOI: 10.1093/bib/bbs077.

[80] C. B. Barber, D. P. Dobkin a H. Huhdanpaa, "The Quickhull Algorithm for Convex Hulls", *ACM Trans. Math. Softw.*, roč. 22, č. 4, s. 469–483, pros. 1996. DOI: 10.1145/235815.235821.

[81] Y. Bromberg a B. Rost, "SNAP: predict effect of non-synonymous polymorphisms on function", *Nucleic Acids Research*, roč. 35, č. 11, s. 3823, 2007. DOI: 10.1093/nar/gkm238.

[82] J. Štourač, "Systém na integraci bioinformatických nástrojů pro analýzu tunelů v proteinech", Bakalářská práce, Fakulta informatiky, Masarykova Univerzita, 2015.

[83] A. Prlić, A. Yates, S. E. Bliven, P. W. Rose, J. Jacobsen, P. V. Troshin, M. Chapman, J. Gao, C. H. Koh, S. Foisy, R. Holland, G. Rimša, M. L. Heuer, H. Brandstätter–Müller, P. E. Bourne a S. Willis, "BioJava: an open-source framework for bioinformatics in 2012", *Bioinformatics*, roč. 28, č. 20, s. 2693, 2012. DOI: 10.1093/bioinformatics/bts494.

[84] S. Patient, D. Wieser, M. Kleen, E. Kretschmann, M. Jesus Martin a R. Apweiler, "UniProtJAPI: a remote API for accessing UniProt data", *Bioinformatics*, roč. 24, č. 10, s. 1321, 2008. DOI: 10.1093/bioinformatics/btn122.

[85] A. A. Fodor a R. W. Aldrich, "Influence of conservation on calculations of amino acid covariance in multiple sequence alignments", *Proteins: Structure, Function, and Bioinformatics*, roč. 56, č. 2, s. 211–221, 2004. DOI: 10.1002/prot.20098.

[86] Google, *Google Web Toolkit*. (cit. 14. 05. 2017).

[87] I. Software, *Smart GWT*. (cit. 14. 05. 2015).

[88] R. M. Hanson, "Jmol – a paradigm shift in crystallographic visualization", *Journal of Applied Crystallography*, roč. 43, č. 5, s. 1250–1260, 1. říj. 2010. DOI: 10.1107/S0021889810030256.

[89] W. Li, A. Cowley, M. Uludag, T. Gur, H. McWilliam, S. Squizzato, Y. M. Park, N. Buso a R. Lopez, "The EMBL-EBI bioinformatics web and programmatic tools framework", *Nucleic Acids Research*, roč. 43, č. W1, W580, 2015. DOI: 10.1093/nar/gkv279.

[90] S. Fetzner a F. Lingens, "Bacterial dehalogenases: biochemistry, genetics, and biotechnological applications.", *Microbiological Reviews*, roč. 58, č. 4, s. 641–685, 1994.

[91] K. A. Gray, T. H. Richardson, K. Kretz, J. M. Short, F. Bartnek, R. Knowles, L. Kan, P. E. Swanson a D. E. Robertson, "Rapid Evolution of Reversible Denaturation and Elevated Melting Temperature in a Microbial Haloalkane Dehalogenase", *Advanced Synthesis & Catalysis*, roč. 343, č. 6-7, s. 607–617, 2001. DOI: 10.1002/1615-4169(200108)343:6/7<607::AID-ADSC607>3.0.CO;2-M.

[92] T. Bosma, J. Damborský, G. Stucki a D. B. Janssen, "Biodegradation of 1,2,3-Trichloropropane through Directed Evolution and Heterologous Expression of a Haloalkane Dehalogenase Gene", *Applied and Environmental Microbiology*, roč. 68, č. 7, s. 3582–3587, 2002. DOI: 10.1128/AEM.68.7.3582-3587.2002.

[93] M. Pavlova, M. Klvana, Z. Prokop, R. Chaloupkova, P. Banas, M. Otyepka, R. C. Wade, M. Tsuda, Y. Nagata a J. Damborsky, "Redesigning dehalogenase access tunnels as a strategy for degrading an anthropogenic substrate", *Nat Chem Biol*, roč. 5, č. 10, s. 727–733, říj. 2009. DOI: 10.1038/nchembio.205.

[94] A. N. Bigley a F. M. Raushel, "Catalytic mechanisms for phosphotriesterases", *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, roč. 1834, č. 1, s. 443–453, 2013. DOI: 10.1016/j.bbapap.2012.04.004.

[95] S. Gopal, V. Rastogi, W. Ashman a W. Mulbry, "Mutagenesis of Organophosphorus Hydrolase to Enhance Hydrolysis of the Nerve Agent VX", *Biochemical and Biophysical Research Communications*, roč. 279, č. 2, s. 516–519, 2000. DOI: 10.1006/bbrc.2000.4004.

[96] L. M. Watkins, H. J. Mahoney, J. K. McCulloch a F. M. Raushel, "Augmented Hydrolysis of Diisopropyl Fluorophosphate in Engineered Mutants of Phosphotriesterase", *Journal of Biological Chemistry*, roč. 272, č. 41, s. 25 596–25 601, 1997. DOI: 10.1074/jbc.272.41.25596.

[97] P.-C. Tsai, A. Bigley, Y. Li, E. Ghanem, C. L. Cadieux, S. A. Kasten, T. E. Reeves, D. M. Cerasoli a F. M. Raushel, "Stereoselective Hydrolysis of Organophosphate Nerve Agents by the Bacterial Phosphotriesterase", *Biochemistry*, roč. 49, č. 37, s. 7978–7987, 2010. DOI: 10.1021/bi101056m.

[98] L. Briseño-Roa, C. M. Timperley, A. D. Griffiths a A. R. Fersht, "Phosphotriesterase variants with high methylphosphonatase activity and strong negative trade-off against phosphotriesters", *Protein Engineering, Design and Selection*, roč. 24, č. 1-2, s. 151, 2011. DOI: 10.1093/protein/gzq076.

[99] M. Arand, A. Cronin, M. Adamska a F. Oesch, "Epoxide Hydrolases: Structure, Function, Mechanism, and Assay", in *Phase II Conjugation Enzymes and Transport Systems*, ř. Methods in Enzymology, sv. 400, Academic Press, 2005, s. 569–588. DOI: 10.1016/S0076-6879(05)00032-7.

[100] M. T. Reetz, C. Torre, A. Eipper, R. Lohmer, M. Hermes, B. Brunner, A. Maichele, M. Bocola, M. Arand, A. Cronin, Y. Genzel, A. Archelas a R. Furstoss, "Enhancing the Enantioselectivity of an Epoxide Hydrolase by Directed Evolution", *Organic Letters*, roč. 6, č. 2, s. 177–180, 2004. DOI: 10.1021/ol035898m.

[101] J. G. Voet a R. H. Abeles, "The Mechanism of Action of Sucrose Phosphorylase: Isolation and Properties of a Beta-Linked Covalent Glucose-Enzyme Complex", *Journal of Biological Chemistry*, roč. 245, č. 5, s. 1020–1031, 1970.

[102] A. Cerdobbel, K. De Winter, D. Aerts, R. Kuipers, H.-J. Joosten, W. Soetaert a T. Desmet, "Increasing the thermostability of sucrose phosphorylase by a combination of sequence- and structure-based mutagenesis", *Protein Engineering, Design and Selection*, roč. 24, č. 11, s. 829, 2011. DOI: 10.1093/protein/gzr042.

[103] A. Houde, A. Kademi a D. Leblanc, "Lipases and their industrial applications", *Applied Biochemistry and Biotechnology*, roč. 118, č. 1, s. 155–170, 2004. DOI: 10.1385/ABAB:118:1-3:155.

[104] A. Fieker, J. Philpott a M. Armand, "Enzyme replacement therapy for pancreatic insufficiency: present and future", *Clin Exp Gastroenterol*, roč. 4, s. 55–73, květ. 2011. DOI: 10.2147/CEG.S17634.

[105] H. Jochens, D. Aerts a U. T. Bornscheuer, "Thermostabilization of an esterase by alignment-guided focussed directed evolution", *Protein Engineering, Design and Selection*, roč. 23, č. 12, s. 903, 2010. DOI: 10.1093/protein/gzq071.

[106] S. J. Putman, A. F. W. Coulson, I. R. T. Farley, B. Riddleston a J. R. Knowles, "Specificity and kinetics of triose phosphate isomerase from chicken muscle", *Biochem J*, roč. 129, č. 2, s. 301–310, zář. 1972.

[107] E. S. J. Arnér a A. Holmgren, "Physiological functions of thioredoxin and thioredoxin reductase", *European Journal of Biochemistry*, roč. 267, č. 20, s. 6102–6109, 2000. DOI: 10.1046/j.1432-1327.2000.01701.x.

[108] F. K. Majiduddin, I. C. Materon a T. G. Palzkill, "Molecular analysis of beta-lactamase structure and function", *International Journal of Medical Microbiology*, roč. 292, č. 2, s. 127–137, 2002. DOI: 10.1078/1438-4221-00198.

[109] G. D'Alessio a J. Riordan, *Ribonucleases: Structures and Functions*. Elsevier Science, 1997, ISBN: 9780080540597.

[110] A. Akasako, M. Haruki, M. Oobatake a S. Kanaya, "Conformational Stabilities of Escherichia coli RNase HI Variants with a Series of Amino Acid Substitutions at a Cavity within the Hydrophobic Core", *Journal of Biological Chemistry*, roč. 272, č. 30, s. 18 686–18 693, 1997. DOI: 10.1074/jbc.272.30.18686.

[111] E. de Jong, W. J. H. van Berkel, R. P. van der Zwan a J. A. M. de Bont, "Purification and characterization of vanillyl-alcohol oxidase from Penicillium simplicissimum", *European Journal of Biochemistry*, roč. 208, č. 3, s. 651–657, 1992. DOI: 10.1111/j.1432-1033.1992.tb17231.x.

[112] R. H. H. van den Heuvel, M. W. Fraaije, M. Ferrer, A. Mattevi a W. J. H. van Berkel, "Inversion of stereospecificity of vanillyl-alcohol oxidase", *Proceedings of the National Academy of Sciences*, roč. 97, č. 17, s. 9455–9460, 2000. DOI: 10.1073/pnas.160175897.

[113] R. W. Hartley, "Barnase and barstar: two small proteins to fold and fit together", *Trends in Biochemical Sciences*, roč. 14, č. 11, s. 450–454, 1989. DOI: 10.1016/0968-0004(89)90104-7.

[114] T. R. Killick, S. M. Freund a A. R. Fersht, "Real-time NMR studies on a transient folding intermediate of barstar.", *Protein Sci*, roč. 8, č. 6, s. 1286–1291, červ. 1999.

[115] R. Guerois, J. E. Nielsen a L. Serrano, "Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations", *Journal of Molecular Biology*, roč. 320, č. 2, s. 369–387, DOI: 10.1016/S0022-2836(02)00442-4.

[116] E. H. Kellogg, A. Leaver-Fay a D. Baker, "Role of conformational sampling in computing mutation-induced changes in protein structure and stability", *Proteins: Structure, Function, and Bioinformatics*, roč. 79, č. 3, s. 830–838, 2011. DOI: 10.1002/prot.22921.

A Seznam elektronických příloh

- `hsw.tar.gz` – archiv obsahující zdrojový kód jádra (jako NetBeans projekt)
- `hsw-gui.tar.gz` – archiv obsahující zdrojový kód webového rozhraní (jako Eclipse GWT projekt[1])

1. Pro jeho správné načtení je třeba mít nainstalovaný „Google Plugin for Eclipse“ a „Google Web Toolkit SDK“. Návod pro instalaci je k dispozici zde: `https://developers.google.com/eclipse/docs/download`.

B Článek publikovaný v odborném časopise Nucleic Acids Research

J. Bendl, J. Štourač, E. Šebestová, O. Vávra, M. Musil, J. Brezovský a J. Damborský, "HotSpot Wizard 2: Automated Design of Site-Specific Mutations and Smart Libraries in Protein Engineering", *Nucleic Acids Research*, roč. 44, s. W479-W487, 2016. DOI: 10.1093/nar/gkw416.

Published online 12 May 2016 *Nucleic Acids Research, 2016, Vol. 44, Web Server issue* ***W479–W487***
doi: 10.1093/nar/gkw416

HotSpot Wizard 2.0: automated design of site-specific mutations and smart libraries in protein engineering

Jaroslav Bendl[1,2,3,†], Jan Stourac[1,†], Eva Sebestova[1], Ondrej Vavra[1], Milos Musil[1,2], Jan Brezovsky[1,3,*] and Jiri Damborsky[1,3,*]

[1]Loschmidt Laboratories, Department of Experimental Biology and Research Centre for Toxic Compounds in the Environment RECETOX, Masaryk University, 625 00 Brno, Czech Republic, [2]Department of Information Systems, Faculty of Information Technology, Brno University of Technology, 612 66 Brno, Czech Republic and [3]International Clinical Research Center, St. Anne's University Hospital Brno, Brno, Czech Republic

Received February 21, 2016; Revised April 22, 2016; Accepted May 03, 2016

ABSTRACT

HotSpot Wizard 2.0 is a web server for automated identification of hot spots and design of smart libraries for engineering proteins' stability, catalytic activity, substrate specificity and enantioselectivity. The server integrates sequence, structural and evolutionary information obtained from 3 databases and 20 computational tools. Users are guided through the processes of selecting hot spots using four different protein engineering strategies and optimizing the resulting library's size by narrowing down a set of substitutions at individual randomized positions. The only required input is a query protein structure. The results of the calculations are mapped onto the protein's structure and visualized with a JSmol applet. HotSpot Wizard lists annotated residues suitable for mutagenesis and can automatically design appropriate codons for each implemented strategy. Overall, HotSpot Wizard provides comprehensive annotations of protein structures and assists protein engineers with the rational design of site-specific mutations and focused libraries. It is freely available at http://loschmidt.chemi.muni.cz/hotspotwizard.

INTRODUCTION

The development of tailor-made enzymes for industrial applications is facilitated by understanding the molecular mechanisms of protein function. However, despite significant advances in recent decades, it is not yet clear how a protein's sequence encodes its function (1,2). Traditional directed evolution circumvents this problem by using repeated rounds of random mutagenesis and screening of large sequence libraries to explore the mutational landscape and find proteins with desired properties (2–5). This approach has the advantage of requiring no prior knowledge of the protein's structure or understanding of its structure–function relationships (6), but necessitates the laborious and costly screening of very large libraries (4). The efficiency of directed evolution experiments can be significantly improved by creating smaller, higher quality libraries that are more likely to yield positive results. Such 'smart' libraries can be generated by focusing mutagenesis on a limited number of 'hot spot' positions that are likely to affect the property of interest, or by selecting a limited set of substitutions (1–5).

The optimal strategy for identifying hot spots depends on the property being targeted. Catalytic properties such as activity, specificity and stereoselectivity are often related to amino acid residues that mediate substrate binding, transition-state stabilization or product release (7,8). Such residues can be identified using tools for predicting and analyzing enzyme-ligand interactions (9–11) or detecting binding pockets or access tunnels (12–14). Strategies for improving protein stability include rigidification of flexible sites, cavity-filling, tunnel engineering, consensus and ancestral mutation methods, or redesigning of surface charges (15–17). While hot spots for some of these strategies can be identified straightforwardly using a single computational tool (18), others require multi-step analyses or the use of molecular modelling methods (19). Having obtained a set of promising sites for manipulating the desired property, the next challenge is to draw up a list of allowed substitutions at individual positions. This can be done by considering the amino acid distribution at the corresponding positions in sequence homologs (20,21), by using reduced sets of amino acids with either specific desired physicochemical properties or a balanced set of these properties (22,23), or on the basis of the predicted effects of specific substitutions on the protein's properties (24,25). Finally, an appropriate degen-

*To whom correspondence should be addressed. Tel: +420 5 4949 3467; Fax: +420 5 4949 6302; Email: jiri@chemi.muni.cz
Correspondence may also be addressed to Jan Brezovsky. Email: brezovsky@mail.muni.cz
† These authors contributed equally to the work as first authors.

erate codon covering the specified set of amino acids must be selected for each targeted position. Ideally, these codons should exhibit minimal amino acid bias and minimize the frequency of premature stop codons (26). Several tools are available to facilitate this task and to calculate the size of the designed library (27).

Here, we present HotSpot Wizard 2.0, a web server for the automated identification of hot spots and design of smart libraries for engineering protein stability, enzymatic activity, substrate specificity and enantioselectivity. Compared to its predecessor (28), HotSpot Wizard 2.0 introduces several major improvements, extending the scope and quality of its analyses. It implements four different established protein engineering strategies, enabling the user to selectively target sites affecting the protein's stability and catalytic properties. Users can easily select suitable substitutions for individual hot spots based on predictions of tolerated amino acids or amino acid distributions in sequence homologs, and suitable degenerate codons for these substitutions can be designed automatically via the HotSpot Wizard interface. A new graphical user interface provides an intuitive and comprehensive overview of the results of the analysis, allowing users to think directly about the obtained designs. The resulting pipeline of twenty integrated tools and three databases represents a unique one-stop solution that makes library design accessible even to users with no prior knowledge of bioinformatics.

MATERIALS AND METHODS

The workflow of HotSpot Wizard is outlined in Figure 1. In order to explore the mutational landscape and find the most promising mutagenesis targets, a protein selected by the user is annotated using several prediction tools and databases (Phase 1). With this knowledge in hand, four protein engineering strategies are used to identify suitable hot spots for improving desired protein properties (Phase 2). Finally, suitable substitutions and appropriate degenerate codons are proposed for each selected hot spot, enabling the design of a smart library (Phase 3).

Phase 1: annotation of the protein

The first step in the workflow requires the user to specify the protein structure of interest, either by providing its PDB ID or by uploading a suitable PDB file. If possible, the biological assembly of the target protein is automatically generated by the MakeMultimer tool (http://watcut.uwaterloo.ca/tools/makemultimer), and information about 'essential residues' directly involved in catalysis or binding is obtained from the Catalytic Site Atlas (29) and UniProtKB/Swiss-Prot (30) databases. The DSSP algorithm (31) is then used to assign the protein's secondary structure, and its accessible surface area is computed using the Shrake and Rupley algorithm (32) with BioJava (33). The average B-factors are computed for the protein's amino acid residues (34). The raw B-factor values are accompanied by residue rankings ranging from 1–100%; rankings of 1–25%, 26–75% and 76–100% indicate high, moderate and low levels of relative structural flexibility, respectively. Protein pockets are then identified with Fpocket (35). For each chain, the pocket containing the greatest number of essential residues is identified as the catalytic pocket. If there are two or more pockets that satisfy this criterion, a decision is made according to the Fpocket score. Having identified the putative catalytic pockets, their centers of mass are determined and used as starting points to identify access tunnels with CAVER (36). Sequence homologs of the target protein are then obtained by performing a BLAST (37) search against the UniRef90 (38) database, using the target protein sequence as a query. All identified homologs are aligned with the query protein using USEARCH (39). By default, sequences whose identity with the query is below 30% or above 90% are excluded from the list of homologs. The remaining sequences are then clustered using UCLUST (39), with a 90% identity threshold to remove close homologs. The cluster representatives are sorted based on the BLAST query coverage and by default, the first 200 of them are used to create a sequence data set. A multiple sequence alignment of the resulting sequence data set is created with Clustal Omega (40) and used to (i) estimate the conservation of each position in the protein based on the Jensen–Shannon entropy (41); (ii) identify correlated positions using an ensemble of the MI (42), aMIc (43), OMES (44), SCA (45), DCA (46), McBASC (47) and ELSC (48) methods; (iii) predict the tolerated amino acids at each position in the protein sequence using RAPHYD (see Supplementary Data 1); and (iv) analyze amino acid frequencies at individual positions within the protein. The conservation scores are used to assign mutability values to individual residues. To facilitate interpretation, these values are divided into three groups: values of 1–3, 4–5 and 6–9 indicate low, moderate and high mutability, respectively.

Phase 2: identification of mutagenesis hot spots

Based on the comprehensive annotation of the target protein, four protein engineering strategies are used to identify different types of hot spots: (i) functional hot spots, (ii) stability hot spots based on structural flexibility, (iii) stability hot spots based on sequence consensus and (iv) correlated hot spots. Some examples illustrating the use of these strategies to engineer selected properties in 12 different proteins (34,49–62) are shown in Figure 2. Functional hot spots correspond to highly mutable residues located in the catalytic pockets or tunnels connecting these pockets with the bulk solvent. Residues located in close proximity to the active site have been identified as good mutagenesis targets for engineering activity, enantioselectivity and substrate specificity (52,63,64). To prevent mutagenesis at positions that are indispensable for protein function, all essential residues are designed immutable and thus excluded from the list of potential hot spots. Supplementary Data 2 shows that HotSpot Wizard provides a significantly greater proportion of viable mutants than random mutagenesis. Stability hot spots are identified by analyzing structural flexibility and sequence consensus. The former approach aims to rigidify flexible protein regions by mutating residues with high average B-factors (34). B-factor provides a metric for flexibility which is due in part to inherent flexibility of the macromolecule, but also includes stabilizing/destabilizing energy from packing in the crystal lattice. The rationale for targeting these flexible residues is that they have relatively

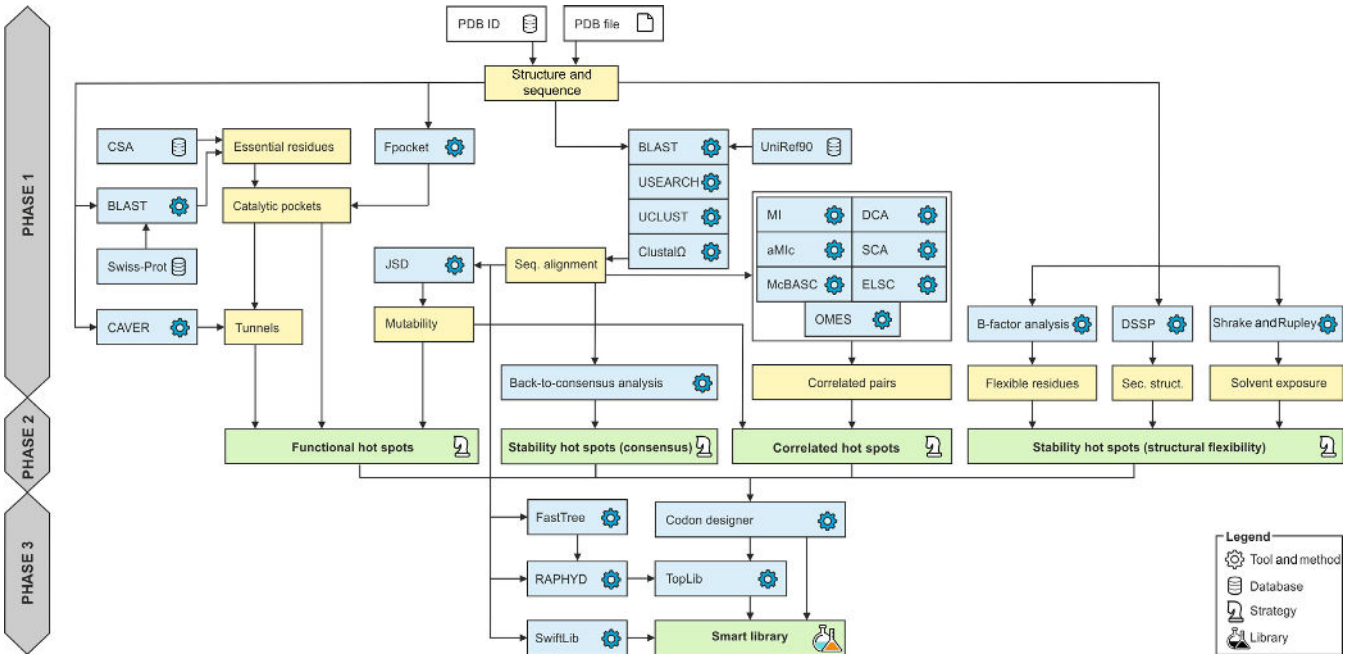


Figure 1. Workflow of HotSpot Wizard.

few contacts with neighbors, so their substitution can produce more interactions (34,54,55). In contrast, the sequence consensus protocol implements majority and frequency ratio approaches, both of which suggest mutations at positions where the wild-type amino acid differs from the most prevalent amino acid (i.e. the consensus residue) at a given position in the multiple sequence alignment. The assumption that the most common amino acid is likely to be stabilizing has proven to be very successful at creating more stable proteins (56–58,65). By default, if the consensus residue is present in at least 50% of all analyzed sequences, the corresponding position is identified as a hot spot in the majority approach. The frequency ratio approach has a less strict criterion for the consensus residue's frequency – the default value is 40%, but it must also be at least five times more frequent than the wild-type residue as a hot spot. The final strategy involves searching for coordinated changes of the amino acids at two separate positions within the protein. Such pairs of positions are referred to as correlated hot spots, and arise when one amino acid substitution has an unfavorable effect that is compensated for by a second mutation of a residue that is located in close structural proximity to the first. This second, correlated mutation typically helps to maintain protein function, stability or folding (66). Methods developed for identifying correlated pairs have revealed mutations responsible for modulating substrate specificity (67), enantioselectivity (68) and mutagenesis targets for stability engineering (69). The identification of correlated positions in HotSpot Wizard is based on an ensemble of seven prediction tools. Each tool generates a raw score for each pair of residues in the protein that measures the pair's degree of correlation. The mean and standard deviation of the degrees of correlation for all pairs of residues in the protein are then calculated and the raw scores are converted into Z-scores, which measure the number of standard deviations by which each pair's raw score deviates from the mean. Based on the work of Martin *et al*. (70), a pair is considered to be correlated if its average Z-score ≥ 3.5 and both of its positions have at least a moderate degree of mutability – by definition, highly conserved positions cannot co-evolve (71).

Phase 3: design of the smart library

The efficiency of directed evolution experiments can be improved by focusing mutagenesis on a limited number of hot spots, but also by restricting the number of allowed substitutions at individual positions using appropriate codons (20–25). For each protein engineering strategy, HotSpot Wizard provides a way to prioritize amino acids at the randomized positions (Table 1) and identifies degenerate codons encoding all desired amino acids with the minimum redundancy and the smallest possible ratio of stop codons. Alternatively, the SwiftLib tool (73) can be used to calculate optimal degenerate codons while keeping the library diversity within the specified limits (the default 10 000). Although the resulting library may not necessarily fully cover the desired set of amino acids, the probability of omitting the important amino acids is relatively low as their weights are set according to selected prioritization method (e.g. based on amino acid distributions in sequence homologs). For both approaches, the most common metrics, such as expected coverage or library size, are computed with TopLib (72).

DESCRIPTION OF THE WEB SERVER

Input

The only required input to the web server is a tertiary structure of the query protein, provided either as a PDB ID or a PDB file. The user can then choose a predefined biological

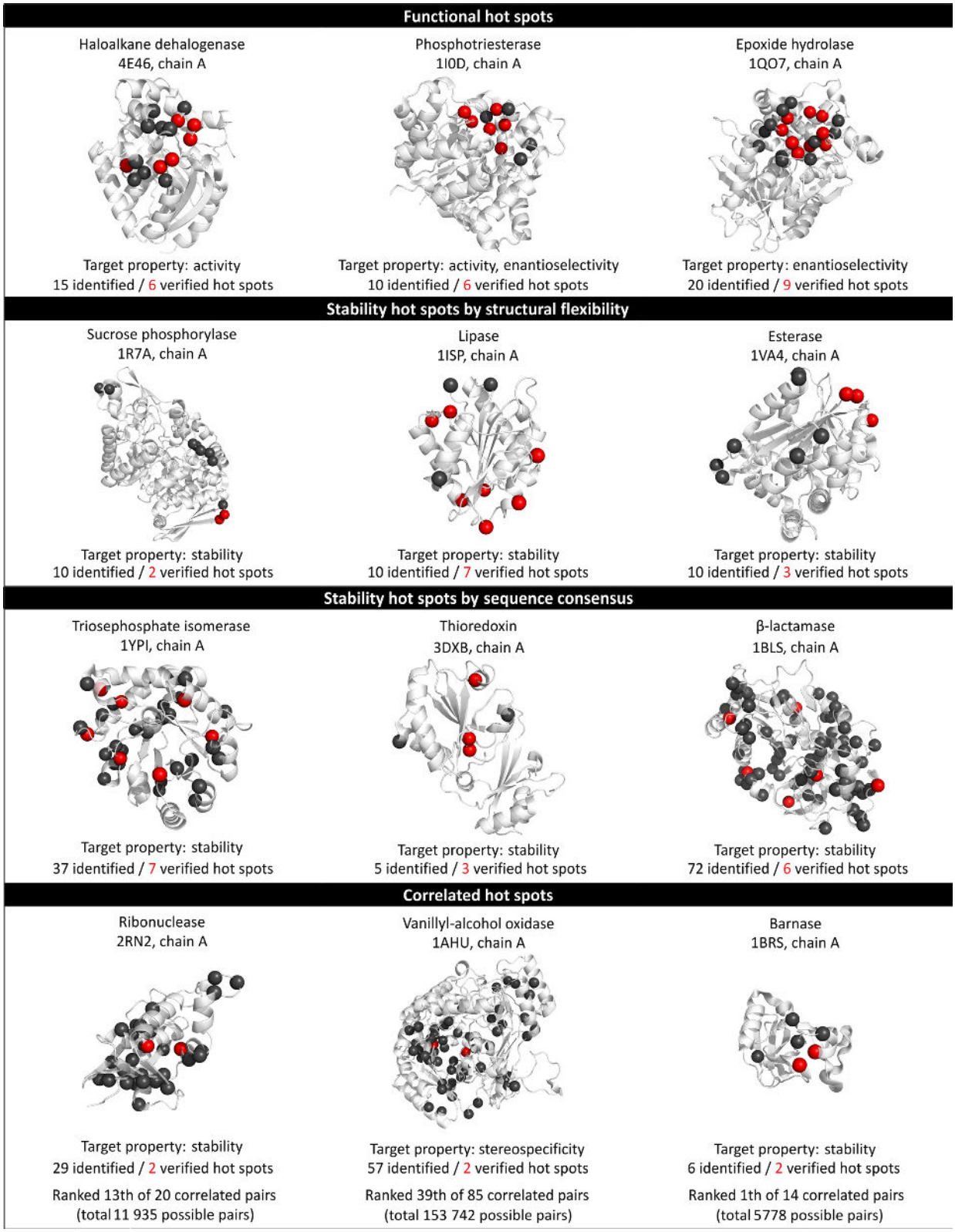


Figure 2. Some notable applications of the four protein engineering strategies implemented in the HotSpot Wizard web server.

Table 1. Methods for selecting substitutions at hot spot positions identified using the four different protein engineering strategies

Selection mode	Availability in strategies	Description
Amino acid frequency	FUNC, FLEX	suggests amino acid residues fulfilling the criterion of minimal frequency in the multiple sequence alignment
Mutational landscape	FUNC, FLEX	suggests amino acid residues fulfilling the criterion of minimal probability of preservation of protein function
Sequence consensus	CONS	suggests amino acid residues fulfilling the criteria of at least one of approaches implemented in sequence consensus strategy: (i) majority approach or (ii) frequency ratio approach
Correlated positions	CORREL	suggests amino acid residues fulfilling the criterion of minimal frequency of co-occurrence with some other specific residue from coupled position
Manual	ALL	manual selection of amino acid residues

FUNC – Analysis of functional hot spots; FLEX – Analysis of stability hot spots/structural flexibility approach; CONS – Analysis of stability hot spots / sequence consensus approach; CORREL – Analysis of correlated hot spots

unit generated by the MakeMultimer tool or manually select chains for which the calculation should be performed. The calculations can be configured in either basic or advanced mode. Basic mode directs the user's attention to the most important parameters, providing an overview of the identified essential residues and highlighting the main parameters involved in the identification of pockets and tunnels. The designation of essential residues is a key step in the functional strategy because these residues are excluded from the list of potential hot spots and are also used to detect catalytic pockets and access tunnels. The user should therefore inspect the automatically generated list of essential residues and correct it if necessary. If no essential residues are detected, the user should specify them manually. In basic mode, the user can specify three parameters: (i) the probe radius, which is used in pocket identification and defines the minimum radius of an alpha sphere in a pocket (default 2.8 Å); (ii) the minimum probe radius, which defines the minimum radius of a putative tunnel (default 1.4 Å); and (iii) the clustering threshold, which determines how the hierarchically clustered tunnels are cut and thus affects the number of tunnels that can be identified (default 3.5 Å). Advanced mode allows expert users to fine-tune parameters of individual calculations in the pipeline to achieve more specialized objectives.

Output

Upon submission, a unique identifier is assigned to each job to track the calculation. The 'Results browser' panel provides information on the status of individual steps in the computational pipeline (Figure 3A). Once the job is finished, the navigation panel provides links to the results obtained using each of the four different protein engineering strategies (Figure 3B). The result pages for each strategy are all organized in the same way, which is described below.

Residue features. The 'Residue features' panel lists all of the identified hot spots together with information relevant to the selected protein engineering strategy (Figure 3C). Several checkboxes can be found at the top of this panel, allowing users to reduce the list of hot spots by applying additional criteria such as excluding buried residues, correlated positions or residues forming a catalytic pocket. The 'Show all residues' button enables users to inspect any residue of the target protein and possibly select hot spots based on their own criteria. Importantly, a pop-up window containing detailed information about a given residue is displayed after clicking the 'book' icon in the last column of the table. Users can visualize individual residues within the protein structure by selecting the 'eye' icon in the first column, and can add residues to the list of mutagenesis hot spots by clicking the 'plus' icon in the second column. All selected mutagenesis hot spots listed in the 'Residues selected for mutagenesis' panel (Figure 3D) can be used for designing a smart library by clicking the 'Design library' button.

Residue details. The information in the 'Residue details' panel is organized into several tabs (Figure 3F): (i) 'Overview', which provides basic information on the residue's characteristics such as its mutability, average B-factor and secondary structure; (ii) 'Annotations', describing the residue's function (only available for essential residues); (iii) 'Tunnels and Pockets', which lists the pockets and/or tunnels of which the residue is a part; (iv) 'Sequence consensus', listing potential consensus mutations for a given position; (v) 'Amino acid frequencies', providing the distribution of amino acids in the corresponding column of the multiple sequence alignment; (vi) 'Mutational landscape', quantifying the probability of preservation of protein function for individual substitutions at a given site; and (vii) 'Correlated positions', listing all positions correlated with the site in question.

Design of smart library. The 'Library design' panel allows the user to select a set of substitutions and design degenerate codons for systematic mutagenesis of the selected positions (Figure 3G). An automatic method for prioritizing amino acids suitable for the chosen protein engineering strategy will be pre-selected. The panel contains two tabs, each corresponding to one library optimization mode. In the 'Standard mode', users can manually define their own set of required substitutions for individual positions if they so desire. After any change in the list of amino acids, HotSpot Wizard automatically identifies the most suitable codons covering all desired amino acids with the lowest possible redundancy, and the library size corresponding to the specified expected coverage. The parameters of the library can be modified interchangeably, allowing the user to adjust the final library based on its size or preferred degree of its coverage. In the 'SwiftLib mode', users specify the maximum acceptable library diversity and the method reports the op-

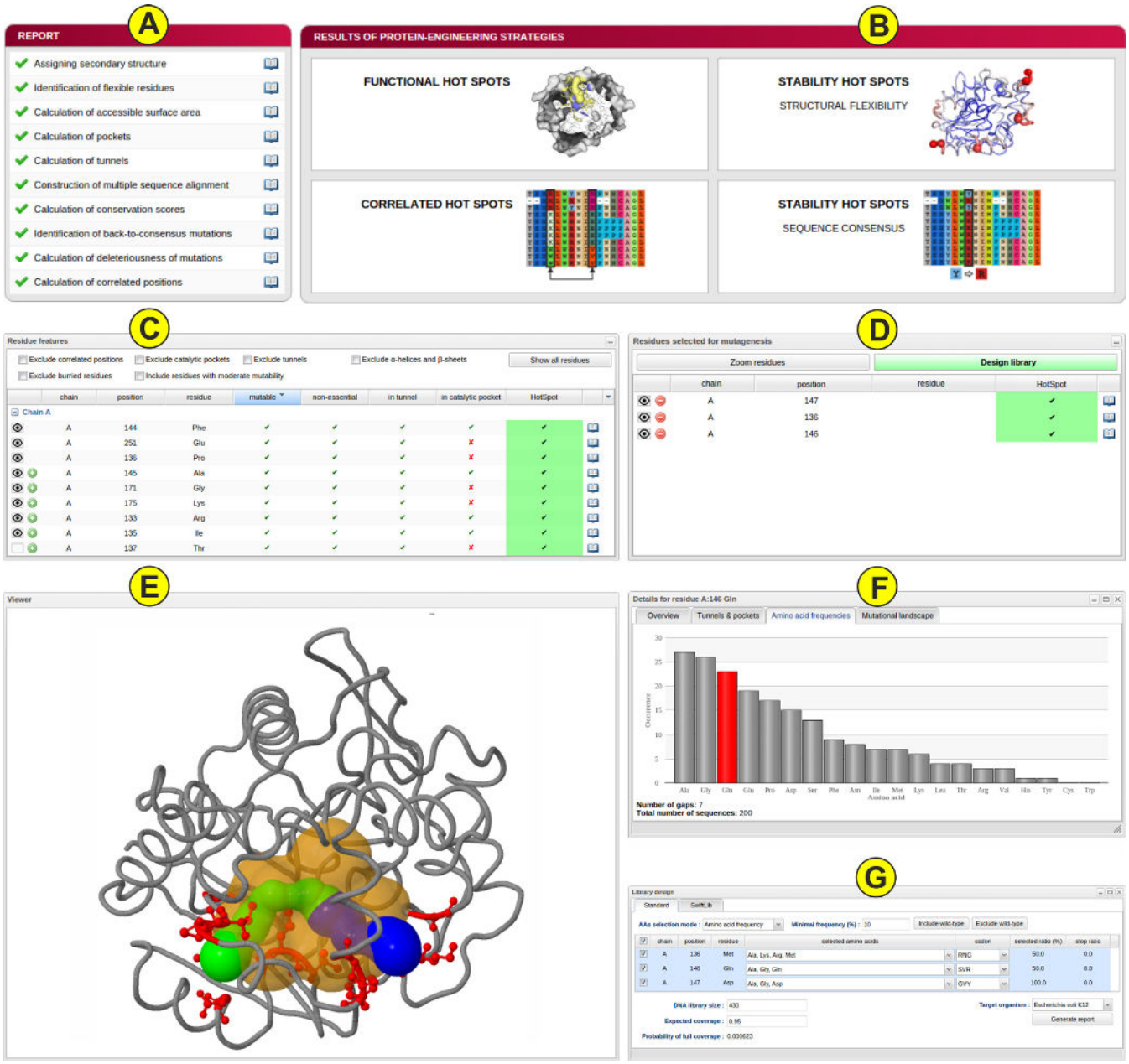


Figure 3. HotSpot Wizard's graphical user interface, showing results obtained for the haloalkane dehalogenase LinB (PDB ID: 1CV2). (**A**) The 'Report' panel shows the status of the calculations in the individual steps of the computational pipeline. (**B**) Results obtained using the four protein engineering strategies. (**C**) The 'Residue features' panel, which provides an overview of the identified hot spots. (**D**) The 'Residues selected for mutagenesis' panel, which presents a user-adjustable list of residues representing targets for mutagenesis. (**E**) The JSmol viewer allows interactive visualization of the protein and the identified tunnels and pockets. (**F**) The 'Residue details' pop-up window, which provides comprehensive information on the residue's annotations, organized under several tabs. (**G**) The 'Library design' panel, which shows the list of substitutions and appropriate codons for randomization of selected positions.

timal combination of codons with the minimal redundancy of amino acids. However, this efficiency is often achieved at the price of omitting some of desired amino acids with lower weights. The initial amino acid weights derived from the selected prioritization scheme can be changed by selecting the 'Edit amino acid weights'. Additionally, users can request multiple solutions and thus inspect also the solutions which are considered as less optimal by the method, but may better meet the users' needs. Finally, users can generate a nucleotide sequence from the designed amino acid sequence based on the codon usage of selected organism (default is *Escherichia coli*) with the European Molecular Biology Open Software Suite (EMBOSS) Backtranseq tool (74).

Protein visualization. The protein of interest is interactively visualized in the web browser using the JSmol applet (http://wiki.jmol.org/index.php/JSmol). Users can dis-

play individual amino acid residues as well as identified tunnels and pockets (Figure 3E). The hot spot residues are colored in red, residues in tunnels and pockets in yellow and all other residues in grey.

Structural features. The main characteristics of all pockets and access tunnels are presented in the 'Pockets' and 'Tunnels' panels, respectively. These panels allow users to visualize individual pockets and tunnels in the structure and to open a pop-up window showing a list of all the residues comprising the chosen structural feature.

CONCLUSIONS AND OUTLOOK

HotSpot Wizard 2.0 is a web server for the automatic identification of hot spots and the design of site-specific mutations and mutant libraries for engineering protein stability, catalytic activity, substrate specificity and enantioselectivity. The server provides a unified interface allowing users to apply four well-established protein engineering strategies that combine structural, functional and evolutionary information to identify suitable positions for mutagenesis. Moreover, HotSpot Wizard integrates several schemes for automatic prioritization of mutations and codon optimization for selected hot spot positions to facilitate the design of smart libraries. The automation of the multi-step procedure makes the process of library design accessible to users without expertise in bioinformatics because it eliminates the need to select, install and evaluate tools, optimize their parameters, perform conversions between different data formats, and interpret intermediate results.

In the future, we plan to implement a protocol for structure prediction based on homology modeling, extending the applicability of HotSpot Wizard to proteins for which no experimental structure is yet available. Additionally, we aim to assess other established protein engineering strategies and, if they prove suitable, to develop new modules so they can be added to the server's portfolio of methods.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to express many thanks to Dr Antonin Pavelka (Masaryk University, Brno, Czech Republic) for valuable discussions and Dr Yuval Nov (University of Haifa, Haifa, Israel) for kindly providing the source code of TopLib. Uwe Bornscheuer (University Greifswald), Marco Fraaije (Groningen University) and Moshe Goldsmith (Weizmann Institute of Science) are sincerely acknowledged for constructive comments on the tool. MetaCentrum and CERIT-SC are acknowledged for providing access to computing and storage facilities [LM2015085 and LM2015042].

FUNDING

Ministry of Education of the Czech Republic [LO1214, LQ1605, LM2015055 and LM2015047]; Grant Agency of the Czech Republic [GA16-06096S]; European Commission REGPOT [316345]; Horizon 2020 Research Infrastructure ELIXIR-EXCELERATE [676559]; Brno University of Technology [FIT-S-14-2299 to M.M.]. Funding for open access charge: Grants from Czech Ministry of Education [LO1214, LQ1605, LM2015055 and LM2015047].
Conflict of interest statement. None declared.

REFERENCES

1. Romero,P.A. and Arnold,F.H. (2009) Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.*, **10**, 866–876.
2. Currin,A., Swainston,N., Day,P.J. and Kell,D.B. (2015) Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. *Chem. Soc. Rev.*, **44**, 1172–1239.
3. Cheng,F., Zhu,L. and Schwaneberg,U. (2015) Directed evolution 2.0: improving and deciphering enzyme properties. *Chem. Commun. (Camb.)*, **51**, 9760–9772.
4. Lutz,S. (2010) Beyond directed evolution–semi-rational protein engineering and design. *Curr. Opin. Biotechnol.*, **21**, 734–743.
5. Acevedo-Rocha,C.G., Reetz,M.T. and Nov,Y. (2015) Economical analysis of saturation mutagenesis experiments. *Sci. Rep.*, **5**, 10654.
6. Lo Surdo,P., Walsh,M.A. and Sollazzo,M. (2004) A novel ADP- and zinc-binding fold from function-directed in vitro evolution. *Nat. Struct. Mol. Biol.*, **11**, 382–383.
7. Denard,C.A., Ren,H. and Zhao,H. (2015) Improving and repurposing biocatalysts via directed evolution. *Curr. Opin. Chem. Biol.*, **25**, 55–64.
8. Bornscheuer,U.T., Huisman,G.W., Kazlauskas,R.J., Lutz,S., Moore,J.C. and Robins,K. (2012) Engineering the third wave of biocatalysis. *Nature*, **485**, 185–194.
9. Xie,Z.-R. and Hwang,M.-J. (2015) Methods for predicting protein-ligand binding sites. *Methods Mol. Biol.*, **1215**, 383–398.
10. Yuan,Y., Pei,J. and Lai,L. (2013) Binding site detection and druggability prediction of protein targets for structure-based drug design. *Curr. Pharm. Des.*, **19**, 2326–2333.
11. Lavecchia,A. and Di Giovanni,C. (2013) Virtual screening strategies in drug discovery: a critical review. *Curr. Med. Chem.*, **20**, 2839–2860.
12. Sebestova,E., Bendl,J., Brezovsky,J. and Damborsky,J. (2014) Computational tools for designing smart libraries. *Methods Mol. Biol.*, **1179**, 291–314.
13. Brezovsky,J., Chovancova,E., Gora,A., Pavelka,A., Biedermannova,L. and Damborsky,J. (2013) Software tools for identification, visualization and analysis of protein tunnels and channels. *Biotechnol. Adv.*, **31**, 38–49.
14. Zhang,Z., Li,Y., Lin,B., Schroeder,M. and Huang,B. (2011) Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics*, **27**, 2083–2088.
15. Bommarius,A.S. and Paye,M.F. (2013) Stabilizing biocatalysts. *Chem. Soc. Rev.*, **42**, 6534–6565.
16. Wijma,H.J., Floor,R.J. and Janssen,D.B. (2013) Structure- and sequence-analysis inspired engineering of proteins for enhanced thermostability. *Curr. Opin. Struct. Biol.*, **23**, 588–594.
17. Yu,H. and Huang,H. (2014) Engineering proteins for thermostability through rigidifying flexible sites. *Biotechnol. Adv.*, **32**, 308–315.
18. Folkman,L., Stantic,B., Sattar,A. and Zhou,Y. (2016) EASE-MM: Sequence-based prediction of mutation-induced stability changes with feature-based multiple models. *J. Mol. Biol.*, **428**, 1394–1405.
19. Bednar,D., Beerens,K., Sebestova,E., Bendl,J., Khare,S., Chaloupkova,R., Prokop,Z., Brezovsky,J., Baker,D. and Damborsky,J. (2015) FireProt: Energy- and evolution-based computational design of thermostable multiple-point mutants. *PLoS Comput. Biol.*, **11**, e1004556.
20. Reetz,M.T. and Wu,S. (2008) Greatly reduced amino acid alphabets in directed evolution: making the right choice for saturation mutagenesis at homologous enzyme positions. *Chem. Commun. (Camb)*, **43**, 5499–5501.
21. Jochens,H. and Bornscheuer,U.T. (2010) Natural diversity to guide focused directed evolution. *Chembiochem*, **11**, 1861–1866.

22. Pines,G., Pines,A., Garst,A.D., Zeitoun,R.I., Lynch,S.A. and Gill,R.T. (2015) Codon compression algorithms for saturation mutagenesis. *ACS Synth. Biol.*, **4**, 604–614.
23. Reetz,M.T., Kahakeaw,D. and Lohmer,R. (2008) Addressing the numbers problem in directed evolution. *Chembiochem*, **9**, 1797–1804.
24. Goldsmith,M. and Tawfik,D.S. (2013) Enzyme engineering by targeted libraries. *Methods Enzymol.*, **523**, 257–283.
25. Chaparro-Riggers,J.F., Polizzi,K.M. and Bommarius,A.S. (2007) Better library design: data-driven protein engineering. *Biotechnol. J.*, **2**, 180–191.
26. Gaytán,P., Contreras-Zambrano,C., Ortiz-Alvarado,M., Morales-Pablos,A. and Yáñez,J. (2009) TrimerDimer: an oligonucleotide-based saturation mutagenesis approach that removes redundant and stop codons. *Nucleic Acids Res.*, **37**, e125.
27. Nov,Y. (2014) Probabilistic methods in directed evolution: library size, mutation rate, and diversity. *Methods Mol. Biol.*, **1179**, 261–278.
28. Pavelka,A., Chovancova,E. and Damborsky,J. (2009) HotSpot Wizard: a web server for identification of hot spots in protein engineering. *Nucleic Acids Res.*, **37**, W376–W383.
29. Furnham,N., Holliday,G.L., de Beer,T.A.P., Jacobsen,J.O.B., Pearson,W.R. and Thornton,J.M. (2014) The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res.*, **42**, D485–D489.
30. UniProt Consortium. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
31. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
32. Shrake,A. and Rupley,J.A. (1973) Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.*, **79**, 351–371.
33. Prlić,A., Yates,A., Bliven,S.E., Rose,P.W., Jacobsen,J., Troshin,P.V., Chapman,M., Gao,J., Koh,C.H., Foisy,S. *et al.* (2012) BioJava: an open-source framework for bioinformatics in 2012. *Bioinformatics*, **28**, 2693–2695.
34. Reetz,M.T., Carballeira,J.D. and Vogel,A. (2006) Iterative saturation mutagenesis on the basis of B factors as a strategy for increasing protein thermostability. *Angew. Chem. Int. Ed Engl.*, **45**, 7745–7751.
35. Le Guilloux,V., Schmidtke,P. and Tuffery,P. (2009) Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics*, **10**, 168.
36. Chovancova,E., Pavelka,A., Benes,P., Strnad,O., Brezovsky,J., Kozlikova,B., Gora,A., Sustr,V., Klvana,M., Medek,P. *et al.* (2012) CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. *PLoS Comput. Biol.*, **8**, e1002708.
37. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
38. Suzek,B.E., Wang,Y., Huang,H., McGarvey,P.B., Wu,C.H. and UniProt Consortium. (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, **31**, 926–932.
39. Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.
40. 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.
41. Capra,J.A. and Singh,M. (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics*, **23**, 1875–1882.
42. Korber,B.T., Farber,R.M., Wolpert,D.H. and Lapedes,A.S. (1993) Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proc. Natl. Acad. Sci. U.S.A.*, **90**, 7176–7180.
43. Lee,B.-C. and Kim,D. (2009) A new method for revealing correlated mutations under the structural and functional constraints in proteins. *Bioinformatics*, **25**, 2506–2513.
44. Kass,I. and Horovitz,A. (2002) Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins*, **48**, 611–617.
45. Lockless,S.W. and Ranganathan,R. (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, **286**, 295–299.
46. Weigt,M., White,R.A., Szurmant,H., Hoch,J.A. and Hwa,T. (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 67–72.
47. Olmea,O., Rost,B. and Valencia,A. (1999) Effective use of sequence correlation and conservation in fold recognition. *J. Mol. Biol.*, **293**, 1221–1239.
48. Dekker,J.P., Fodor,A., Aldrich,R.W. and Yellen,G. (2004) A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. *Bioinformatics*, **20**, 1565–1572.
49. Pavlova,M., Klvana,M., Prokop,Z., Chaloupkova,R., Banas,P., Otyepka,M., Wade,R.C., Tsuda,M., Nagata,Y. and Damborsky,J. (2009) Redesigning dehalogenase access tunnels as a strategy for degrading an anthropogenic substrate. *Nat. Chem. Biol.*, **5**, 727–733.
50. Gopal,S., Rastogi,V., Ashman,W. and Mulbry,W. (2000) Mutagenesis of organophosphorus hydrolase to enhance hydrolysis of the nerve agent VX. *Biochem. Biophys. Res. Commun.*, **279**, 516–519.
51. Watkins,L.M., Mahoney,H.J., McCulloch,J.K. and Raushel,F.M. (1997) Augmented hydrolysis of diisopropyl fluorophosphate in engineered mutants of phosphotriesterase. *J. Biol. Chem.*, **272**, 25596–25601.
52. Reetz,M.T., Wang,L.-W. and Bocola,M. (2006) Directed evolution of enantioselective enzymes: iterative cycles of CASTing for probing protein-sequence space. *Angew. Chem. Int. Ed Engl.*, **45**, 1236–1241.
53. Reetz,M.T., Torre,C., Eipper,A., Lohmer,R., Hermes,M., Brunner,B., Maichele,A., Bocola,M., Arand,M., Cronin,A. *et al.* (2004) Enhancing the enantioselectivity of an epoxide hydrolase by directed evolution. *Org. Lett.*, **6**, 177–180.
54. Cerdobbel,A., De Winter,K., Aerts,D., Kuipers,R., Joosten,H.-J., Soetaert,W. and Desmet,T. (2011) Increasing the thermostability of sucrose phosphorylase by a combination of sequence- and structure-based mutagenesis. *Protein Eng. Des. Sel.*, **24**, 829–834.
55. Jochens,H., Aerts,D. and Bornscheuer,U.T. (2010) Thermostabilization of an esterase by alignment-guided focussed directed evolution. *Protein Eng. Des. Sel.*, **23**, 903–909.
56. Sullivan,B.J., Nguyen,T., Durani,V., Mathur,D., Rojas,S., Thomas,M., Syu,T. and Magliery,T.J. (2012) Stabilizing proteins from sequence statistics: the interplay of conservation and correlation in triosephosphate isomerase stability. *J. Mol. Biol.*, **420**, 384–399.
57. Pey,A.L., Rodriguez-Larrea,D., Bomke,S., Dammers,S., Godoy-Ruiz,R., Garcia-Mira,M.M. and Sanchez-Ruiz,J.M. (2008) Engineering proteins with tunable thermodynamic and kinetic stabilities. *Proteins*, **71**, 165–174.
58. Amin,N., Liu,A.D., Ramer,S., Aehle,W., Meijer,D., Metin,M., Wong,S., Gualfetti,P. and Schellenberger,V. (2004) Construction of stabilized proteins by combinatorial consensus mutagenesis. *Protein Eng. Des. Sel.*, **17**, 787–793.
59. Akasako,A., Haruki,M., Oobatake,M. and Kanaya,S. (1997) Conformational stabilities of Escherichia coli RNase HI variants with a series of amino acid substitutions at a cavity within the hydrophobic core. *J. Biol. Chem.*, **272**, 18686–18693.
60. van den Heuvel,R.H.H., Fraaije,M.W., Ferrer,M., Mattevi,A. and van Berkel,W.J.H. (2000) Inversion of stereospecificity of vanillyl-alcohol oxidase. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 9455–9460.
61. Killick,T.R., Freund,S.M. and Fersht,A.R. (1998) Real-time NMR studies on folding of mutants of barnase and chymotrypsin inhibitor 2. *FEBS Lett.*, **423**, 110–112.
62. Encell,L.P., Friedman Ohana,R., Zimmerman,K., Otto,P., Vidugiris,G., Wood,M.G., Los,G.V., McDougall,M.G., Zimprich,C., Karassina,N. *et al.* (2012) Development of a dehalogenase-based protein fusion tag capable of rapid, selective and covalent attachment to customizable ligands. *Curr. Chem. Genomics*, **6**, 55–71.
63. Reetz,M.T., Bocola,M., Carballeira,J.D., Zha,D. and Vogel,A. (2005) Expanding the range of substrate acceptance of enzymes: combinatorial active-site saturation test. *Angew. Chem. Int. Ed Engl.*, **44**, 4192–4196.
64. Morley,K.L. and Kazlauskas,R.J. (2005) Improving enzyme properties: when are closer mutations better? *Trends Biotechnol.*, **23**, 231–237.
65. Lehmann,M., Loch,C., Middendorf,A., Studer,D., Lassen,S.F., Pasamontes,L., van Loon,A.P.G.M. and Wyss,M. (2002) The consensus concept for thermostability engineering of proteins: further proof of concept. *Protein Eng.*, **15**, 403–411.

66. de Juan,D., Pazos,F. and Valencia,A. (2013) Emerging methods in protein co-evolution. *Nat. Rev. Genet.*, **14**, 249–261.
67. Kuipers,R.K.P., Joosten,H.-J., Verwiel,E., Paans,S., Akerboom,J., van der Oost,J., Leferink,N.G.H., van Berkel,W.J.H., Vriend,G. and Schaap,P.J. (2009) Correlated mutation analyses on super-family alignments reveal functionally important residues. *Proteins*, **76**, 608–616.
68. Nobili,A., Tao,Y., Pavlidis,I.V., van den Bergh,T., Joosten,H.-J., Tan,T. and Bornscheuer,U.T. (2015) Simultaneous use of in silico design and a correlated mutation network as a tool to efficiently guide enzyme engineering. *Chembiochem*, **16**, 805–810.
69. Wang,C., Huang,R., He,B. and Du,Q. (2012) Improving the thermostability of alpha-amylase by combinatorial coevolving-site saturation mutagenesis. *BMC Bioinformatics*, **13**, 263.
70. Martin,L.C., Gloor,G.B., Dunn,S.D. and Wahl,L.M. (2005) Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, **21**, 4116–4124.
71. Fodor,A.A. and Aldrich,R.W. (2004) Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins*, **56**, 211–221.
72. Nov,Y. (2012) When second best is good enough: another probabilistic look at saturation mutagenesis. *Appl. Environ. Microbiol.*, **78**, 258–262.
73. Jacobs,T.M., Yumerefendi,H., Kuhlman,B. and Leaver-Fay,A. (2015) SwiftLib: rapid degenerate-codon-library optimization through dynamic programming. *Nucleic Acids Res.*, **43**, e34.
74. Li,W., Cowley,A., Uludag,M., Gur,T., McWilliam,H., Squizzato,S., Park,Y.M., Buso,N. and Lopez,R. (2015) The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res.*, **43**, W580–W584.