ANALYSIS OF THE TOOLS FOR DETECTING SIMILARITIES BETWEEN TERTIARY PROTEIN STRUCTURES

Jiří Trlica

Bachelor Degree Programme (4), FIT BUT

E-mail: xtrlic00@stud.fit.vutbr.cz

Supervised by: Jaroslav Bendl

E-mail: ibendl@fit.vutbr.cz

Abstract: Alignment of the three-dimensional structures of proteins is an essential task in bioinformatics. Because there are many tools offering this functionality, only a limited subset of them was chosen for comparison (DALI, LOCK 2, SPALIGN, MUSTANG and CLICK). These tools vary in the principle of calculation. Their performance was measured on three proteins, which represent main protein classes (all-α, all-β, α/β). These proteins were tested against a subset of PMD database containing 2 357 records. The results were visualized by ROC curves and the tools were compared by their area under ROC curve (AUC metric). According to this metric, the best results were obtained for SPALIGN.

Keywords: protein structure alignment, protein fold comparison, structural alignment software, LOCK 2, DALI, MUSTANG, SPALIGN, CLICK

1. ÚVOD

Jeden z největších problémů bioinformatiky spočívá v určování podobnosti proteinů. Proteiny jsou molekuly nezbytné pro životní pochody všech organizmů, neboť plní řadu důležitých funkcí (např. stavební nebo pohybová funkce) [1]. Určování podobnosti proteinových struktur nachází uplatnění zejména při snaze o zjištění přibližné funkce neznámých proteinů. Při řešení takové úlohy se neznámý protein porovnává s množinou proteinů se známou funkcí a poté je zařazen do stejné rodiny a foldu (pravidelné uspořádání sekundárních struktur, které lze pozorovat v rámci proteinové struktury), jaký náleží nejpodobnějšímu z nich. Ke vzájemnému zarovnání dvojic proteinových struktur lze využít řadu nástrojů. Poslední srovnání jejich výkonnosti je však již více než deset let staré [2]. Od té doby vzniklo několik nových nástrojů uplatňujících nové techniky zarovnání. Cílem tohoto příspěvku je porovnat sadu vybraných nástrojů a určit, který z nich pracuje nejpřesněji.

2. STRUKTURNÍ ZAROVNÁNÍ

Techniky pro zarovnání proteinových struktur mohou být využity pro zjištění proteinové rodiny a foldu. Pokud tedy víme, jakou funkci daný protein má, dá se s vysokou pravděpodobností předpokládat, že sekvenčně podobný protein v jiném organismu tuto funkci bude plnit také. Většina běžně užívaných zarovnávacích metod zvládne docela dobře rozpoznat společné znaky v proteinech - například odpovídající si úseky sekundárních struktur. Nicméně úplné zarovnání, tedy konkrétní mapování atomů z jedné struktury na druhou, je velice obtížné a jeho přesnost může záviset na několika různých faktorech. Velkou roli hraje i konkrétní nastavení použitého nástroje.

2.1. METRIKY PODOBNOSTI

Běžně používanou metrikou k ohodnocení podobnosti dvou zarovnaných proteinových struktur na atomární úrovni je tzv. RMSD (root-mean-square deviation). Tato číselná charakteristika, jejíž jednotkou je angström, je určitou obdobou standardní odchylky – udává průměrnou vzdálenost odpovídajících si dvojic atomů v porovnávaných proteinech. Jinou metrikou může být zarovnání

s využitím prvků sekundárních struktur. Zatímco v průběhu evoluce v proteinech dochází k mnoha mutacím měnících aminokyselinové složení, úseky sekundárních struktur zůstávají zachovány mnohem lépe.

3. VÝBĚR NÁSTROJŮ K ANALÝZE

Při výběru bioinformatických nástrojů byl kladen největší důraz na jejich odlišný přístup k zarovnání proteinových struktur. Byl analyzován nástroj, který využívá podobnost elementů sekundární struktury (nástroj LOCK 2) nebo zarovnává dvě struktury tím, že optimalizuje vyhodnocovací funkci, která měří strukturní podobnosti na atomární úrovni (nástroj SPALIGN). Jiný nástroj využívá k zarovnání pouze C_α na proteinové páteři bez jakékoliv optimalizace (nástroj DALI). Další využívá prostorovou informaci z C_α k vytvoření sekvenčního zarovnání (nástroj MUSTANG). Jako poslední zkoumaný nástroj srovnává shluky bodů (nástroj CLICK). Všechny tyto uvedené nástroje pojí jediné - a to výsledná RMSD hodnota každého zarovnání.

4. KONSTRUKCE TESTOVACÍHO DATASETU

Před procesem vzájemného srovnání nástrojů je nutné korektně zkonstruovat testovací dataset proteinových struktur. K tomuto účelu se použil nástroj PDBselect spuštěný nad databází PDB s parametrem maximální sekvenční identity nastaveným na 25%, čímž bylo zajištěno, že v datasetu nebudou příliš podobné proteiny ve více instancích. Vygenerovaný dataset má textový formát a obsahuje 2 357 záznamů, kde každý záznam se skládá z jednoznačného identifikátoru proteinové struktury a názvu přiřazené třídy a foldu. Foldy byly společně s třídami přiřazeny z databáze strukturní klasifikace proteinů SCOP.

4.1. VÝBĚR REPREZENTATIVNÍCH STRUKTUR

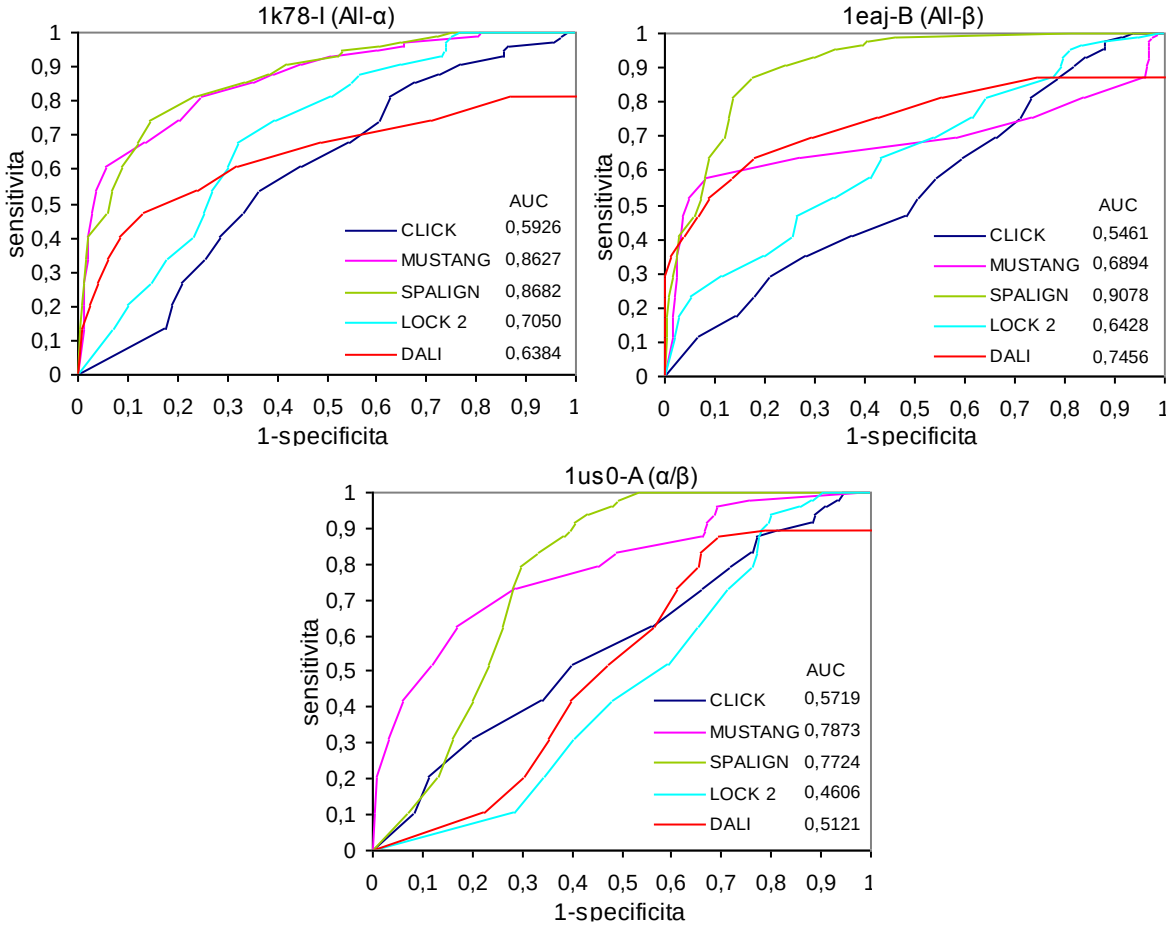
Z datasetu byly vybrány 3 struktury, které reprezentují odlišné třídy strukturní topologie proteinů:

- PDB ID: 1k78, řetězec I, třída all-α (v datasetu je 74 dalších struktur se stejným foldem),
- PDB ID: 1eaj, řetězec B, třída all-β (v datasetu je 86 dalších struktur se stejným foldem),
- PDB ID: 1us0, řetězec A, třída α/β (v datasetu je 48 dalších struktur se stejným foldem).

5. VÝSLEDKY EXPERIMENTŮ

Experiment probíhal tím způsobem, že každá reprezentativní struktura byla zarovnána ke všem ostatním strukturám v testovacím datasetu (to bylo provedeno pro každý nástroj zvlášť). Při každém zarovnání se uložila hodnota RMSD vypočítaná daným nástrojem. Pro každou reprezentativní strukturu byla vytvořena tabulka se sloupci: třída, fold a naměřené RMSD hodnoty od všech nástrojů. Poté byly v této tabulce vzestupně seřazeny všechny záznamy podle velikosti RMSD hodnoty pro aktuálně analyzovaný nástroj a napočítány tzv. false positive. Záznam je označený jako false positive v tom případě, když má jiné zařazení do proteinové třídy a foldu než má reprezentativní struktura a zároveň nižší RMSD hodnotu než záznam s nejvyšším RMSD se zařazením stejným jako reprezentativní struktura. Spolu s false positives byly vypočítány další dvě metriky – sensitivity (udává závislost růstu počtu správně zařazených struktur na současně rostoucím počtu struktur, které jsou nesprávně označeny jako strukturně podobné) a true positive (udává počet správných podobností jako má reprezentativní struktura) a všechny výsledné hodnoty uloženy do tabulky. K tomu, aby byla vygenerována ROC křivka, bylo zapotřebí vypočítat ještě další 3 metriky – false negative (udává počet ve skutečnosti podobných struktur, které nebyly nástrojem rozpoznány jako podobné), true negative (udává počet ve skutečnosti odlišných struktur, které nebyly nástrojem rozpoznány jako podobné) a specificity (udává závislost růstu počtu nesprávně zařazených struktur na současně rostoucím počtu struktur, které jsou správně označeny jako strukturně podobné). Z výše uvedených metrik se následně vypočítají aktuální hodnoty pro sensitivity (true_positive / (true_positive + false_negative)) a specificity (false_positive / (false_positive + true_negative)).

Z těchto hodnot se poté vygeneruje ROC křivka a spočítá se hodnota AUC (area under curve). Tato hodnota nejlépe popisuje chování zkoumaného nástroje, přičemž za nejlepší je považován program s nejvyšší hodnotou této metriky.



Obrázek 1: ROC pro všechny reprezentativní struktury.

6. ZÁVĚR

Úspěšnost nástrojů pro zarovnání proteinových struktur lze nejobjektivněji ohodnotit podle vypočítaných ploch pod odpovídajícími ROC křivkami. Podle této metriky je nejlepší nástroj SPALIGN (průměrné AUC pro všechny tři reprezentativní struktury je 0.85), dále pak MUSTANG (avg AUC = 0.78), DALI (avg AUC = 0.63), LOCK 2 (avg AUC = 0.6) a CLICK (avg AUC = 0.57). Toto pořadí víceméně odpovídá i datu vzniku jednotlivých nástrojů.

PODĚKOVÁNÍ

Tento příspěvek vznikl za podpory grantu FIT-S-11-2 a výzkumného záměru MSM 0021630528. Pro provádění experimentů byla využita distribuovaná výpočetní infrastruktura MetaCentra (projekt LM2010005).

REFERENCE

[1] Alberts, B. a kol. Základy buněčné biologie: Úvod do molekulární biologie buňky. 2 vyd. Espero Publishing, 2005. ISBN 80-902906-2-0.

[2] Singh, A. P., Douglas, B. L. Protein Structure Alignment: A Comparison of Methods. Bioinformatics, 2000. ISSN 1367-4803.