PREDICTING THE EFFECT OF AMINO ACID SUBSTITUTIONS ON PROTEIN FUNCTION USING MAPP METHOD

Ondřej Pelikán

Master Degree Programme (3), FIT BUT

E-mail: xpelik08@stud.fit.vutbr.cz

Supervised by: Jaroslav Bendl

E-mail: ibendl@fit.vutbr.cz

Abstract: There are many tools for prediction of the effect of amino acid substitutions on protein function. In this study, we are focusing on analysis of MAPP method. Although MAPP was developed as one of the first methods, it has never been frequently used for research purposes. It is mainly due to the difficulty of configuration when user must prepare own alignment and phylogenetic tree using third-party software. This work aims to develop comprehensive library for full automation of entire prediction process. Due to this reason, optimal parameters of underlying tools are being searched and consequently evaluated on two independent datasets.

Keywords: MAPP, SNP, protein mutations, protein predictions, protein nucleotide variants

1. ÚVOD

Díky moderním sekvenačním metodám je k dispozici tak velké množství dat o genetických variacích organismů, že je nelze všechna zkoumat experimentálně. Z toho důvodu se vyvíjí počítačové metody, které sice nejsou tak přesné jako laboratorní experimenty, ale jsou vhodné pro předzpracování velkého objemu dat.

Jedním z nástrojů pro předpověď vlivu mutace na funkci proteinu je program MAPP (Multivariate Analysis of Protein Polymorphism) [1]. V originálním textu nebyla metoda MAPP otestována na žádném známém datasetu. Cílem této práce je najít optimální kombinaci vstupních parametrů pro metodu MAPP, otestovat ji na datasetu, který se používá v proteinovém inženýrství a vytvořit knihovnu pro její snadné použití. Významný přínos práce spočívá také v implementaci dosud neexistujícího webového rozhraní.

2. METODA MAPP

Tato metoda zkoumá především konzervované pozice, což jsou místa, která se během procesu vývoje proteinu měnila jen málo a v těchto místech metoda předpokládá, že mutace bude mít dopad na funkci proteinu. Konzervovaná místa určuje z vícenásobného zarovnání a fylogenetického stromu. Tyto vstupy však uživatel musí sám vytvořit pomocí programů třetích stran. Při rozhodování bere MAPP v potaz také odlišnosti v šesti základních fyzikochemikálních vlastnostech mezi původní a substituovanou aminokyselinou [1].

3. VYHLEDÁNÍ PARAMETRŮ PRO METODU MAPP

Hledání optimální kombinace podpůrných programů pro metodu MAPP bylo prováděno na proteinu laktózového represoru lacI, na kterém metodu testovali sami autoři a uvádí přesnost předpovědi 69,2%. Záznamy tohoto proteinu obsahují 1 517 škodlivých a 2 260 neutrálních mutací.

3.1. POUŽITÉ PROGRAMY

Jedna varianta analýzy se skládala z kombinace trojice programů a jejich nastavení. První program z databáze sekvencí získá homologní sekvence, druhý program z nich vytvoří zarovnání a třetí program ze zarovnání vytvoří fylogenetický strom. U každé varianty byla vyhodnocena výsledná přesnost analýzy. Stavový prostor tvořený výběrem programů a jejich parametry je prakticky neomezený. K testování bylo vybráno 95 kombinací s cílem co nejlepšího pokrytí stavového prostoru.

K výběru sekvencí pro analýzu byly použity programy *HMMER, BLASTP, PSI-BLAST, CD-HIT*, pro konstrukci zarovnání *ProbCons, ClustalW, ClusalΩ, MAFFT* a pro tvorbu fylogenetického stromu *FastTree* a *RAxML*.

3.2. VÝSLEDKY

Největší vliv na výslednou přesnost metody MAPP má etapa výběru sekvencí pro analýzu. Jak ukazuje **tabulka 1**, přesnost roste s počtem sekvencí vybraných programem *BLASTP*. Program *BLASTP* implicitně sekvence řadí podle podobnosti. Při malém počtu velmi podobných sekvencí rozhodovací model metody MAPP detekuje prakticky každou pozici za vysoce konzervovanou, a tedy škodlivou z hlediska efektu mutace. Více sekvencí znamená také větší rozmanitost na jednotlivých pozicích a tedy i více informací o konzervovanosti pro metodu MAPP. Čas potřebný k provedení analýzy ale geometricky roste s počtem sekvencí. Např. čas potřebný pro 50 sekvencí je 14s a pro 1 500 sekvencí až 6h. I při 1 500 sekvencích však nebyl dostatečně pokryt prostor podobných sekvencí ohraničený hodnotou e-value.

Tento problém je možné řešit shlukováním sekvencí, ke kterému slouží program *CD-HIT*. Konkrétní postup výběru sekvencí byl takový, že se programem *BLASTP* nebo *HMMER* vybralo z databáze velké množství sekvencí (25 000) a na tomto výběru bylo provedeno shlukování. Navzájem podobné sekvence se seskupily do shluků a z každého shluku byla vybrána jedna reprezentativní sekvence. Tento přístup umožňuje dostatečné pokrytí prostoru nalezených homologních sekvencí při zachování malého počtu sekvencí, které vstupují do analýzy.

Nejlepšího výsledku (71,2%) bylo dosaženo s metodou výběru 25 000 sekvencí programem *HMMER*, následné shlukování programem *CD-HIT* a vybrání 50 reprezentativních sekvencí z vytvořených shluků. Tato varianta je označena jako *HMMER+CD-HIT* (viz **tabulka 2**). Pro konstrukci zarovnání byl vybrán program *MAFFT* a pro tvorbu fylogenetického stromu program *FastTree*, který je až o dva řády rychlejší než *RAxML* při zachování přesnosti (viz **tabulka 3)**.

4. OVĚŘENÍ PŘESNOSTI METODY A KOMBINACE PROGRAMŮ NA DATASETU

Přesnost metody byla ověřena na dvou datasetech. *Dataset 1* je podmnožinou *Protein Mutant Database (PMD)*. Obsahuje 2 247 škodlivých a 1 248 neutrálních mutací. *Dataset 2* se skládá z 12 silně promutovaných proteinů a celkem obsahuje 8 074 neutrálních a 4 387 škodlivých mutací. Oba datasety byly převzaty z [2]. Přesnost se měří jako podíl počtu mutací se správně predikovaným vlivem na funkci (dvoustavová hodnota škodlivý efekt / neutrální efekt) vůči celkovému počtu mutací. Z důvodu nestejného počtu škodlivých a neutrálních mutací v datasetech byla výsledná přesnost normalizována tak, aby nedocházelo ke zkreslování výsledku z nevyváženosti tříd.

Při testování *datasetu1* metoda podávala průměrné výsledky (viz **tabulka 4**) a při testování *datasetu 2* metoda MAPP předpovídala dopad mutací s vyšší přesností než ostatní uvedené nástroje (kromě *SNAP)*. Nejedná se však o zásadně významné navýšení přesnosti, takže i u tohoto datasetu se dá považovat nástroj MAPP za srovnatelný s ostatními metodami (viz **tabulka 5**).

Počet sekvencí	50	150	250	500	1000	1500
Přesnost	38,4%	38,1%	63,5%	66,5%	67%	67,1%

Tabulka 1: Závislost přesnosti predikce na počtu sekvencí.

Metoda výběru	BLASTP	HMMER	BLASTP + CD-HIT	HMMER + CD-HIT
Přesnost	67,1%	68,1%	70%	**71,2%**

Tabulka 2: Závislost přesnosti predikce na metodě výběru sekvencí.

		Konstrukce zarovnání			
		ProbCons	**ClustalW**	**ClustalΩ**	**MAFFT**
Tvorba fylogen. stromu	**FastTree**	69,9%	70,3%	71%	**71,2%**
	RAxML	70%	70,4%	70,9%	71,2%

Tabulka 3: Závislost přesnosti predikce na volbě programu pro konstrukci zarovnání a tvorby fylogenetického stromu.

PhD-SNP	SIFT	PolyPhen-2	SNAP	PON-P	MAPP
64%	62%	62%	63%	69%	**63%**

Tabulka 4: Přesnosti predikce nástrojů na *datasetu 1*. Hodnoty všech metod s výjimkou MAPP byly převzaty z [3], výsledek MAPP byl vypočítán s použitím nastavení popsaného v tomto článku.

PhD-SNP	SIFT	nsSNPAnalyzer	SNAP	PANTHER	MAPP
63%	65%	62%	71%	60%	**69%**

Tabulka 5: Přesnosti predikce nástrojů na *datasetu 2*. Hodnoty všech metod s výjimkou MAPP byly převzaty z [2], výsledek MAPP byl vypočítán s použitím nastavení popsaného v tomto článku.

5. ZÁVĚR

Pro nalezenou kombinaci parametrů a podpůrných programů bylo dosaženo na jednom proteinu o 2% větší přesnosti, než na tomto proteinu uváděli autoři. S tímto nastavením byly metodou ohodnoceny 2 nezávislé datasety. Z porovnání různých nástrojů pro předpověď vlivu mutací je patrné, že metoda MAPP podává srovnatelné výsledky s ostatními nástroji.

Největší vliv na přesnost metody MAPP má počáteční výběr homologních sekvencí. Vybrané sekvence musí mít dostatečnou rozmanitost a rozumnou velikost s ohledem na časovou náročnost zbytku analýzy. Při řešení tohoto problému pomohlo shlukování velkého počtu sekvencí.

Aby metoda mohla být snadno použita uživateli v odborné komunitě, byla v jazyce *Python* vyvinuta knihovna automatizující výběr sekvencí, konstrukci zarovnání a fylogenetického stromu. Rovněž je zveřejněno webové rozhraní na adrese `http://ll47.sci.muni.cz:6214/`.

REFERENCE

[1] STONE, Eric. SIDOW, Arend. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Research. 2005-06-17, vol. 15, issue 7, s. 978-986.

[2] BENDL, Jaroslav a kol. PredictSNP: Robust and Accurate Consensus Classifier for Prediction of Disease-Related Mutations. PLoS Comp Biology. 2014-1-16, vol. 10, issue 1.

[3] OLATUBOSUN, Ayodeji a kol. PON-P: Integrated predictor for pathogenicity of missense variants. Human Mutation. 2012, vol. 33, issue 8, s. 1166-1174.