PREDICTION OF PROTEIN STABILITY UPON MUTATIONS USING MACHINE LEARNING

František Malinka
Master Degree Programme (2), FIT BUT
E-mail: xmalin21@stud.fit.vutbr.cz

Supervised by: Jaroslav Bendl
E-mail: ibendl@fit.vutbr.cz

Abstract: This paper describes a new approach to the detection of protein stability change upon mutations. The main goal is to create a new meta-tool, which combines the outputs of eight well-established prediction methods. Such approach should result in improved prediction performance. To find the optimal strategy for combining the outputs of tools, various number of machine learning methods were used.

Keywords: stability prediction, protein stability, machine learning, protein mutation, protherm

1 ÚVOD

Proteiny jsou z chemického hlediska nejsložitější a funkčně nejdůmyslnější známé molekuly, proto není divu, že se velká část výzkumu v bioinformatice zabývá právě jimi. Mutace jednotlivých aminokyselin mohou mít významný vliv na výslednou strukturu i funkci proteinu. Otázkou ovšem zůstává, zda i po těchto mutacích zůstane protein ve své původní složené konformaci, či dojde k jeho rozložení (denaturaci). Schopností predikovat stabilitu proteinu získáváme možnost lepšího pochopení základních vztahů mezi proteinovou sekvencí, strukturou a funkcí. Taktéž se tato schopnost predikovat jeví jako klíčová v oborech zabývajících se genetickým inženýrstvím nebo v lékařství.

V průběhu posledního desetiletí bylo vyvinuto několik metod k určení vlivu aminokyselinových mutací na stabilitu proteinu. Většina z nich je primárně založena na výpočtu energetické funkce popisující interakce mezi jednotlivými residui. Určení této energetické funkce může být založeno na statistické analýze různých vlastností extrahovaných z datasetu proteinových struktur (*statistical potential approaches*), nebo se může jednat o kombinaci váhovaných fyzikálních a statistických energetických výrazů (*empirical potential approaches*). Dalším používaným přístupem je využití analýzy sil mezi jednotlivými atomy (*physical potential approaches*) [1]. Některé nástroje (např. AUTO-MUTE nebo I-Mutant3.0) určují svoji predikci na základně strojového učení (*machine learning approaches*), kde k natrénování modelu využívají databáze s experimentálně naměřenými hodnotami změn po provedení mutací. Některé přístupy taktéž mohou kombinovat výhody statistické analýzy a metod strojového učení, respektive neuronových sítí [1].

2 TVORBA TRÉNOVACÍHO DATASETU

Pro získání výchozích dat, která sloužila pro natrénování predikčního modelu, byla použita volně dostupná databáze ProTherm [2] obsahující experimentálně získaná termodynamická data proteinů a jejich mutací. Jednotlivé databázové záznamy byly pro jednodušší dotazování převedeny do databáze MySQL. Celkově sice databáze ProTherm obsahovala 22 491 záznamů, pro zpracování však bylo vybráno pouze 11 910 záznamů vyhovujících stanoveným kritériím (omezující byl například požadavek na existenci proteinové struktury). Zároveň došlo k rozpoznání jednobodových a vícebodových mutací a tyto mutace lze v databázi rozlišit skrze specifickou hodnotu odpovídajícího atributu. Při

převodu dat do relační databáze byl kladen důraz na korektnost atributů vztahujících se k mutacím a jejich příslušným pozicím. Opravnými algoritmy bylo tímto získáno 986 záznamů, které by jinak skončily neúspěšnou predikcí stability (došlo například k přepočtu pozice mutace).

Pro vytvoření datasetu byly brány v potaz pouze jednobodové mutace, touto selekcí tak byl daný prostor snížen na 9 662 záznamů. Na tyto záznamy byly aplikovány následující podmínky výběru. Záznamy nesměly obsahovat nevyplněnou ΔΔG (změna volné Gibbsovy energie). Pokud existuje mutace s více než jedním záznamem a jsou-li experimentální podmínky stejné, byl vložen do datasetu pouze jeden záznam se zprůměrovanou hodnotou ΔΔG. Pokud jsou experimentální podmínky odlišné, byl vložen do datasetu pouze záznam, který měl atribut *pH* nejbližší fyziologické hodnotě 7 a zároveň byl atribut *t* značící teplotu menší nebo roven hodnotě 50.

Po splnění podmínek výběru dataset obsahoval 1 596 záznamů, z toho u 179 případů došlo ke zprůměrování hodnoty ΔΔG a v důsledku rozdílných experimentálních podmínek bylo eliminováno 75 záznamů. Výsledný dataset byl vygenerován ve formátu ARFF, který je nativní pro platformu WEKA použitou k testování metod strojového učení.

3 TESTOVÁNÍ METOD STROJOVÉHO UČENÍ

Pro hledání optimálního rozdělení vah mezi vybranými nástroji (dosažení optimálního konsenzu) bylo využito technik strojového učení. Jednotlivé metody byly implementovány pomocí nástroje WEKA [3]. Celkově bylo testováno 28 různých technik umožňujících predikci spojité veličiny (ΔΔG). Pro SVM (*Support Vector Machine*) byla použita externí knihovna LibSVM. Pro co největší eliminaci problému přetrénování bylo použito 10-fold cross-validace.

V tabulce 1 lze nalézt metody strojového učení, které byly vybrány na základě korelace s experimentálně zjištěnými hodnotami ΔΔG. Dále je zde popsán počet stabilizujících (ΔΔG > 0) a destabilizujících mutací (ΔΔG < 0). Metoda *Majority* zde vyjadřuje průměrnou hodnotu spočítanou z výstupů jednotlivých nástrojů (prostý konsenzus).

	Majority	Gaussian Processes	LibSVM Linear kernel	KStar	M5Rules	M5P	Bagging (REPtree)	Random SubSpace
Stabilizující mutace	312	368	346	416	358	360	301	211
Destabilizující mutace	1283	1203	1250	1179	1238	1236	1295	1385
Korelační koef.	**0,475**	**0,642**	**0,579**	**0,713**	**0,656**	**0,678**	**0,678**	**0,663**

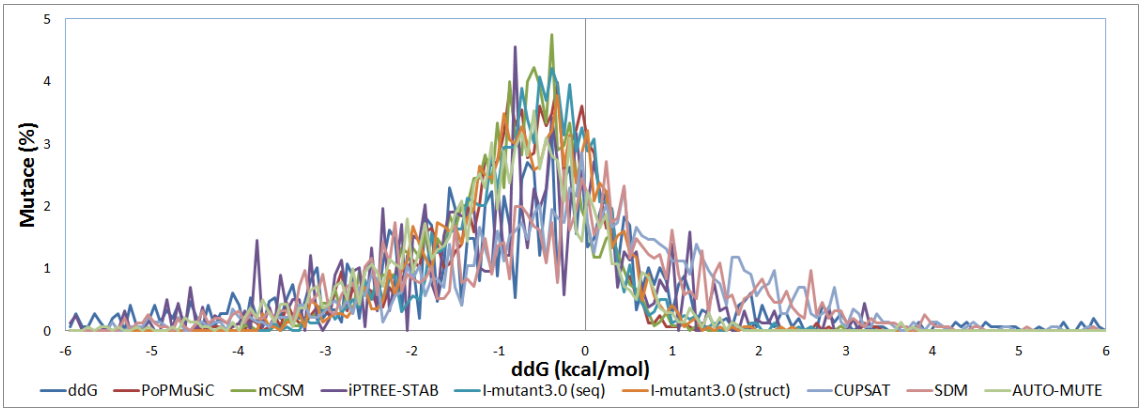
Tabulka 1: Korelační koeficienty pro vybrané metody strojového učení.

Tabulka 2 obsahuje korelační koeficienty a počty mutací pro testované nástroje. Na zvoleném datasetu vykazoval nástroj AUTO-MUTE nejlepších výsledků, jeho korelační koeficient se pohybuje okolo 0,583.

	AUTO-MUTE	SDM	CUPSAT	I-mutant3.0 (strukturní)	I-mutant3.0 (sekvenční)	iPTREE-STAB	mCSM	PoPMuSiC
Stabilizující mutace	218	627	690	273	277	378	159	235
Destabilizující mutace	1173	928	817	1157	1310	1216	1190	1341
Celkem	1393	1556	1510	1435	1594	1594	1349	1581
Korelační koef.	**0,583**	**0,362**	**0,177**	**0,529**	**0,464**	**0,504**	**0,488**	**0,462**

Tabulka 2: Korelační koeficienty pro testované predikční nástroje.

Obrázek 1 zobrazuje graf distribuce predikovaných a experimentálně naměřených ΔΔG hodnot, které jsou vyjádřeny normální distribuční křivkou. Z tohoto grafu lze vyčíst, že v použitém datasetu většina aminokyselinových mutací způsobuje destabilizaci proteinu, extrémní stavy stabilizace/destabilizace se vyskytují velmi zřídka.



Obrázek 1: Distribuce predikovaných a experimentálně zjištěných ΔΔG hodnot. Mezi testované nástroje patří: AUTO-MUTE, SDM, CUPSAT, I-Mutant3.0 (strukturní verze), I-Mutant3.0 (sekvenční verze), iPTREE-STAB, mCSM a PoPMuSiC.

4 ZÁVĚR

Na vytvořeném datasetu se pomocí nejúspěšnější metody strojového učení (KStar) podařilo dosáhnout korelačního koeficientu 0,713, kdežto nejlepší z testovaných nástrojů AUTO-MUTE dosáhl výsledku 0,583. KStar patří do kategorie *lazy learning* metod, které k rozhodnutí využívají hledání podobností atributů v trénovacím datasetu. Odlišný je však v tom, že při vyhodnocování podobností používá výpočet vzdálenosti založený na entropii. Nutné je ovšem podotknout, že i přes použitou 10-fold cross-validaci mohou být některé modely přetrénované a vykazovat tak nadhodnocené výsledky. Přesto však lze považovat dosud získané výsledky za velmi slibné. Při další práci na projektu bude pro eliminaci přetrénování využito nově vybudovaného nezávislého datasetu jednobodových mutací. Tento dataset bude uplatňovat rozdělení cross-validačních foldů na tři schémata. Mutace prvního schématu budou náhodně distribuovány mezi jednotlivé foldy. Druhé (třetí) schéma bude zaručovat, že všechny mutace na stejné pozici residua (stejném proteinu) existují pouze v jednom konkrétním foldu. Predikční model bude navíc rozšířen o schopnost predikovat efekt i u vícebodových mutací.

PODĚKOVÁNÍ

Rád bych zde poděkoval za možnost využít distribuovanou výpočetní infrastrukturu MetaCentra (projekt LM2010005) k ohodnocení datasetu proteinových mutací pomocí testovaných nástrojů.

REFERENCE

[1] KHAN, Sofia, VIHINEN, Mauno. Performance of protein stability predictors. Human mutation, 2010, 31.6: 675-684.

[2] KUMAR, MD Shaji, et al. ProTherm and ProNIT: thermodynamic databases for proteins and protein–nucleic acid interactions. Nucleic Acids Research, 2006, 34: D204-D206.

[3] HALL, Mark, et al. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 2009, 11.1: 10-18.