PREDICTION OF PROTEIN STABILITY UPON MUTATIONS USING EVOLUTION STRATEGY

David Pavlík

Master Degree Programme (2), FIT BUT

E-mail: xpavli60@stud.fit.vutbr.cz

Supervised by: Jaroslav Bendl

E-mail: ibendl@fit.vutbr.cz

Abstract: Predicting the effect of amino acid mutation on protein stability is fundamental for studying diseases or designing and developing new industrial proteins. Many methods of prediction has been developed, employing different principles of decision making and showing different prediction performance. This work aims to use some of already existing prediction tools and combine their estimations of ΔΔG into one value. For finding the optimal weights of selected prediction tools, evolution strategy has been employed.

Keywords: evolution strategy, protein mutations, protein stability, protein predictions, amino acid mutation

1. ÚVOD

Stabilita proteinového řetězce úzce souvisí s uspořádáním aminokyselin v řetězci, a proto při mutaci jedné, či více aminokyselin v proteinu dochází ke změně jeho stability. Pro predikci těchto změn existuje řada nástrojů používajících vzájemně odlišné principy a vykazující různou míru úspěšnosti predikce. Výběr jediného nástroje pro predikci velikosti stabilitních změn proto není dobrou strategií. V rámci této práce jsem vybral tři známé predikční metody, otestoval jejich predikční přesnost a použil je k návrhu meta-klasifikátoru. Vybranými metodami jsou: FoldX, I-Mutant2.0 / strukturní verze a I-Mutant2.0 / sekvenční verze.

Základem pro otestování těchto nástrojů bylo vytvoření trénovací sady jednobodových mutací s již známou hodnotou změny volné energie (udávající změnu stability). Bylo využito volně dostupné databáze Protherm obsahující experimentálně zjištěná data k aminokyselinovým mutacím proteinů.

Po získání predikovaných hodnot změny stability k jednotlivým mutacím trénovací sady bylo využito evolučně inspirované techniky (konkrétně evoluční strategie) pro nalezení vektoru vah k jednotlivým nástrojům. Váhy tvořící vektor jsou posléze využity jako násobící koeficienty při kalkulaci konsenzuálního výsledku tvořeného kombinací výstupů jednotlivých nástrojů. Takto získaný výsledek by měl být bližší reálné velikosti stabilitní změny než predikce jednotlivých nástrojů.

2. STABILITA PROTEINU

Stabilita proteinu vychází z jeho tzv. teploty tání Tm. Při této teplotě dochází k přechodu proteinu do nativní (stabilní) konformace (tzv. proces *renaturace*) nebo do denaturovaného (rozbaleného) stavu (tzv. proces *denaturace*).

Stabilitu lze měřit jako změnu tzv. Gibbsovy (volné) energie (ΔG) v jednotkách kcal/mol, což udává množství změny energie v 1 molu látky při přechodu proteinu ze stabilní konformace do denaturovaného stavu, či naopak. Při predikci se tedy měří změna ΔG (ΔΔG) mezi původním proteinem a jeho mutantem.

2.1. PREDIKČNÍ NÁSTROJE

Porozumění mechanismu, kterým mutace ovlivňují stabilitu proteinů, je důležité především v otázce vztahu struktury a funkce proteinů, návrhu nových proteinů, charakterizace mechanismů chorob a vývojových dynamik organismů [1]. Na základě tohoto bylo vyvinuto několik nástrojů využívající různé metody pro predikci změn volných energií rozkladu proteinů (ΔΔG). Přehled jejich kategorií s několika zástupci je vidět v tabulce 1.

Metodika přístupu	Další dělení	Příklady zástupců nástrojů
založené na energetických funkcích	fyzikální potenciál	EGAD, CC/PBSA
	statistický potenciál	Hunter, PoPMuSiC, Dmutant, MultiMutate, SDM
	empirický potenciál	FoldX, CUPSTAT, PEATSA, ERIS
využívající metody strojového učení	SVM, Neuronové sítě, rozhodovací stromy	I-Mutant2.0, I-Mutant3.0, AUTO-MUTE, MUpro, iPTREE-STAB

Tabulka 1: Přehled metod predikce změny stability proteinu a zástupců nástrojů [1].

3. REALIZACE METAKLASIFIKÁTORU

První krokem bylo získání dostatečného množství dat pro vytvoření trénovací sady mutací s experimentálně naměřenými hodnotami ΔΔG z volně dostupné databáze ProTherm. Tato byla převedena do formátu vlastní relační MySQL databáze *Stability* (pomocí skriptu v jazyce Perl) pro možnost tvorby vlastních dotazů nad těmito daty.

3.1. DÁVKOVÉ VÝPOČTY

Druhým hlavním krokem bylo vytvoření sady skriptů pro řízení dávkových výpočtů na ohodnocení datasetu mutací jednotlivými predikčními nástroji. Každý skript pracuje podle principiálně stejného schématu, kdy postupně prochází datasetem a pro jednotlivé mutace je spuštěn výpočet na daném nástroji. Po dokončení každého výpočtu jsou výsledky zpracovány a uloženy do databáze Stability.

Jelikož nástroje využité pro tuto práci byly vybrány takové, které mají možnost lokální instalace, výpočty byly prováděny vzdáleně na výpočetních zdrojích MetaCentra. Pro účely této práce bylo spotřebováno zhruba 26 dnů procesorového času.

3.2. TRÉNOVACÍ DATASET

Při konstrukci trénovacího datasetu byly vybrány takové mutace, které (1) bylo možné ohodnotit všemi třemi integrovanými nástroji, (2) mají v databázi ProTherm definovanou hodnotu ΔΔG, (3) byly experimentálně změřeny v rozsahu ph=⟨3,9⟩ a teplotě pod 50°C. Navíc platí, že existuje-li více záznamů stejné mutace změřené při rozdílné hodnotě pH, pak se použije jen jediný záznam nejbližší fyziologickému pH=7. Mají-li však mutace totožné hodnoty pH, tak bude do datasetu vložen záznam se zprůměrovanými hodnotami ΔΔG. Po aplikaci těchto pravidel dataset obsahoval 892 záznamů. Dataset pokrývá poměrně velkou část stavového prostoru mutací, jak ukazuje distribuce mutací v tabulce 2. Určité míře přetrénování se při použité metodice vyhnout nelze, ovšem reálné výkonnostní metriky lze zjistit z připravovaného testu na patentech.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0	0,11	0,22	0,22	0,11	1,35	0,22	0,11	0,22	0,34	0,11	0,11	0,45	0,22	0,11	0,56	0,45	0,78	0,11	0,11
C	0,56	0	0	0	0	0,11	0	0,11	0	0,11	0,11	0	0	0	0	0,78	0,11	0,45	0	0
D	2,58	0,11	0	0,67	0,11	0,67	0,67	0,11	0,67	0,11	0,11	1,23	0,11	0,11	0,11	0,56	0,11	0,11	0	0
E	1,91	0,22	0,56	0	0,22	0,45	0,11	0,11	1,12	0,22	0,22	0,45	0,11	1,12	0	0,34	0,34	0,11	0,11	0,11
F	1,12	0	0	0	0	0,22	0	0	0,11	0,45	0	0	0	0	0	0,11	0,11	0,34	0,45	0,22
G	1,12	0,11	0,22	0,11	0	0	0,11	0	0,11	0,11	0	0,11	0,11	0	0,22	0,56	0,11	0,34	0,11	0
H	0,56	0,11	0,11	0,11	0	0,22	0	0	0,11	0,22	0	0,11	0,11	0,34	0,11	0,22	0,11	0	0,11	0,56
I	2,24	0,22	0	0	0,22	0,78	0	0	0	0,45	0,22	0,11	0	0	0	0	0,56	2,8	0,11	0
K	1,79	0	0,11	0,9	0,22	0,67	0,11	0	0	0	0,45	0,45	0,11	0,45	0,67	0	0	0,22	0	0,11

L	2,24	0,34	0	0,11	0,22	0,56	0,22	0,34	0	0	0	0,11	0,22	0,11	0,34	0,11	0,56	1,46	0,11	0,11
M	0,34	0	0	0	0,11	0	0	0,11	0,22	0,45	0	0	0	0	0	0	0,11	0,34	0	0
N	1,68	0	0,67	0	0	0	0,11	0,11	0	0,11	0,11	0	0	0,11	0	0,22	0,11	0,11	0	0
P	1,35	0	0	0	0	0,11	0	0	0	0,11	0	0	0	0	0	0,11	0	0,11	0	0
Q	1,23	0	0	0	0	0,56	0	0,11	0,22	0,11	0	0	0,11	0	0,11	0,11	0	0	0	0
R	0,9	0,11	0	0,11	0	0,11	0,22	0	0,22	0	0,11	0	0	0,22	0	0,11	0	0	0	0
S	1,68	0,11	0,11	0,11	0	0,34	0	0	0	0,11	0	0,11	0	0,11	0	0	0,34	0,22	0	0
T	1,57	0,22	0,34	0,22	0,11	0,78	0,22	0,34	0	0,11	0	0,22	0,34	0,34	0,11	0,78	0	1,79	0	0,11
V	4,04	1,01	0	0,11	0,22	1,46	0,34	1,91	0,11	1,12	0,11	0,22	0,34	0	0,11	0,22	2,35	0	0	0,22
W	0	0	0	0	0,9	0	0,22	0	0	0,22	0	0	0	0	0	0	0	0	0	0,56
Y	0,9	0,22	0,22	0	3,25	0,34	0,11	0	0	0,11	0	0,22	0,11	0,22	0,11	0,22	0	0,11	0,34	0

Tabulka 2: Procentuální zastoupení jednotlivých mutací v trénovacím datasetu.

3.3. VYUŽITÍ EVOLUČNÍ STRATEGIE

Pro účely této práce byla zvolena evoluční strategie (ES) typu 1+1, kde je populace tvořena pouze jedním jedincem, který je reprezentován vektorem hledaných parametrů, jež jsou váhy přiřazené jednotlivým nástrojům dle úspěšnosti jejich predikce. Na jednotlivé parametry je aplikována mutace z Normálního rozdělení $N(0,\sigma)$, kde parametr σ je modifikován tak, aby cca 1/5 potomků byla lepší, než rodiče (pravidlo 1/5) pro daný počet vytvořených potomků.

4. DOSAŽENÉ VÝSLEDKY

Jako hlavní metriku pro dosažené výsledky jsem zvolil Pearsonův korelační koeficient, který udává míru podobnosti mezi dvěma sadami dat a je v rozsahu od 1 (zcela přímá závislost) po -1 (zcela nepřímá závislost). Na zvoleném datasetu se mi pomocí ES podařilo přiblížit predikované hodnoty reálným o necelé 1% oproti nejlepšímu nástroji a o necelých 5% oproti prostému konsenzu (tab. 3).

STATISTIKA	PREDIKČNÍ NÁSTROJ			KONSENZUÁLNÍ METODA	
	FoldX	I-Mutant2.0 (seq)	I-Mutant2.0 (struct)	Prostý konsenzus (neváhovaný)	Evoluční strategie
Ohodnocených mutací (destabilizujících)	277	172	160		
Ohodnocených mutací (stabilizujících)	613	718	732		
Celkem mutací	890	890	892		
Korelační koeficient	0,2821	0,5649	0,4947	0,5226	**0,5716**

Tabulka 3: Statistika dosažených výsledků na trénovacím datasetu.

5. ZÁVĚR

Prozatím dosažené výsledky splnily očekávání. S využitím ES a zkombinováním trojice nástrojů je možné najít lepší výsledky predikce změny stability. Tato metoda bude dále v diplomové práci zdokonalena o využití autoevoluce parametrů evoluční strategie a dále bude testována na datové sadě anotovaných aminokyselinových mutací, které již byly vydolovány z dostupných patentů. Po přidání *state of the art* nástrojů Rosetta a ERIS lze očekávat zlepšení celkových výsledků.

PODĚKOVÁNÍ

Rád bych zde poděkoval za možnost využít distribuovanou výpočetní infrastrukturu MetaCentra (projekt LM201005) k ohodnocení datasetu proteinových mutací pomocí testovaných nástrojů.

REFERENCE

[1] Khan, S.; Vihinen, M.: Performance of protein stability predictors. Human Mutation, ročník 31, 2010: s. 675-684, doi:10.1002/humu.21242.