PREDICTION OF THE EFFECT OF AMINO ACID MUTATIONS ON THE SECONDARY STRUCTURE OF PROTEINS

Miroslav Kadlec

Bachelor Degree Programme (3), FIT BUT

E-mail: xkadle27@stud.fit.vutbr.cz

Supervised by: Jaroslav Bendl

E-mail: ibendl@fit.vutbr.cz

Abstract: This paper is focused on amino acid substitution and its impact on protein secondary structure. The point is to prove, that although the amino acids are frequently mutated during evolution, the elements of protein secondary structure are more robust and compact. The application for simulation of in-silico evolution was created and the results are discussed.

Keywords: protein secondary structure, amino acid mutation, amino acid substitution

1. ÚVOD

Sekundární struktura proteinu je prostorové uspořádání jednotlivých regionů na řetězci aminokyselin. Jsou přitom rozpoznávány tyto prvky sekundární struktury: α-šroubovice, β-list a smyčka. Kombinace těchto prvků definuje tzv. proteinový fold. Struktury s podobným foldem náleží do stejné proteinové rodiny, což znamená, že nesou podobnou funkci.

Tato práce má za cíl potvrdit hypotézu [1], že ačkoliv během evoluce dochází k velkému počtu aminokyselinových mutací, sekundární struktura proteinu je vůči nim relativně imunní a zůstává prakticky nezměněná i po provedení relativně velkého počtu substitucí. Díky této robustnosti sekundární struktury se nemění celkový fold a protein si zachovává původní vlastnosti a funkci. Ačkoliv je tato hypotéza v odborné komunitě obecně přijímána, kvantitativní ohodnocení vlivu mutací dosud chybělo.

Za tímto účelem byla vytvořena aplikace, která umožňuje simulovat evoluci aminokyselinových sekvencí a vyhodnocovat změny v sekundární struktuře. S touto aplikací byl proveden experiment ověřující platnost výše zmíněné hypotézy.

2. MUTAČNÍ PROTOKOL

Mutační protokol určuje, jakým způsobem bude evoluce aminokyselinové sekvence probíhat. Skládá se z předem určeného počtu kroků. V každém kroku je provedena mutace aminokyselinové sekvence a vyhodnocena dílčí změna sekundární struktury. K vyhodnocení podobnosti původní sekundární struktury s tou aktuální se používá metrika Q3 a SOV [2]. Zatímco Q3 určuje poměr počtu aminokyselin se správně predikovanou sekundární strukturou vůči celkovému počtu aminokyselin, SOV bere v úvahu správný počet, typ a pořadí elementů.

2.1. DATASET RS126

Experimenty byly provedeny na datasetu RS126. Tento dataset obsahuje 126 proteinů, které jsou vybrány tak, aby byl co nejlépe pokryt prostor všech známých proteinových sekvencí. K těmto proteinům známe jak primární, tak sekundární strukturu.

2.2. MUTACE

V každém evolučním kroku je proveden definovaný počet mutací na náhodně určených pozicích. Výběr nové aminokyseliny pro každou mutovanou pozici probíhá na základě pravděpodobnosti aminokyselinových záměn definované substituční maticí PAM120. Tato matice je odvozena od matice PAM1 pomocí maticového násobení. PAM1 byla sestavena na základě pozorování velkého počtu mutací ve vzájemně podobných proteinech.

2.3. POČET MUTAČNÍCH KROKŮ

Počet mutačních kroků je určen příslušným parametrem v konfiguračním souboru aplikace. Tento parametr byl získán tak, že se spočítaly ideální počty kroků pro každou sekvenci v datasetu a tyto hodnoty se zprůměrovaly.

Při kalkulaci počtu mutačních kroků pro konkrétní protein v datasetu je nejprve ze souboru načtena jeho sekvence a pro ni se predikuje sekundární struktura. Poté je vygenerována náhodná sekvence aminokyselin o stejné délce a se stejným zastoupením jednotlivých aminokyselin, jako v sekvenci načtené.

Následně je proveden jeden mutační krok, při kterém je zmutován předem definovaný počet aminokyselin (při současných experimentech je mutováno 10% všech aminokyselin v sekvenci). Pro novou sekvenci je predikována sekundární struktura a jsou vypočítány podobnosti mezi predikovanou sekundární strukturou původní a mutované sekvence a náhodné a mutované sekvence. Mutační krok je opakován až do doby, kdy se sekundární struktura mutované sekvence začne více podobat náhodné sekvenci než původní sekvenci. Podobnost je určována metrikou Q3.

Po provedení experimentu podle výše uvedeného algoritmu je pro dataset RS126 za optimální počet považováno 38 mutačních kroků.

2.4. PREDIKČNÍ NÁSTROJ PSIPRED

Pro predikci sekundárních struktur při určování optimálního počtu kroků i při samotném experimentu byl použit program PSIPRED. Jde o predikční nástroj využívající neuronových sítí. Podle autorů dosahuje průměrné úspěšnosti predikce 76-78% [3].

3. VÝSLEDKY

Experiment byl proveden s využitím 102 aminokyselinových sekvencí z datasetu R126 (pro zbývajících 24 se ho nepodařilo úspěšně dokončit).

Z grafu na obrázku 1A je zřejmé, že podobnost sekundárních struktur původní a mutované sekvence (metrika Q3) se s klesající sekvenční identitou snižuje relativně pomalu. Tento fakt je důkazem toho, že sekundární struktura proteinů je do značné míry imunní vůči náhodným mutacím. Při poklesu sekvenční identity pod práh 30% již dochází k rapidnímu poklesu podobnosti. Toto pozorování odpovídá známé skutečnosti [4], že proteiny se sekvenční identitou nad 35% většinou patří do stejné proteinové rodiny (v rámci rodiny platí velmi podobné uspořádání elementů sekundárních struktur).

Graf na obrázku 2B znázorňuje vývoj hodnot Q3, SOV a sekvenční identity v průběhu simulované in-silico evoluce. I zde je zřetelný rozdíl mezi metrikami podobnosti sekundární struktury (SOV, Q3) a sekvenční identitou.

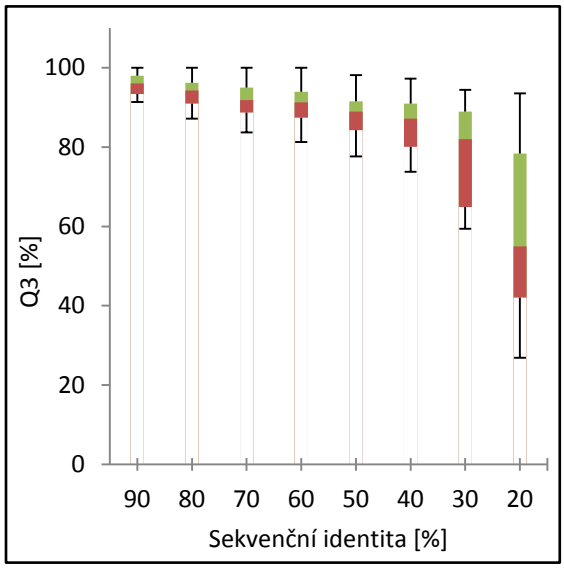
4. ZÁVĚR

Výsledky experimentu provedeného na datasetu RS126 potvrzují platnost hypotézy, že sekundární struktura proteinu je poměrně imunní vůči aminokyselinovým mutacím. Ze souhrnných výsledků je patrné, že po provedení mutací byla sekvenční identita mezi původní a mutovanou sekvencí nízká, naproti tomu hodnoty Q3 a SOV určující podobnost sekundárních struktur byly poměrně vysoké.

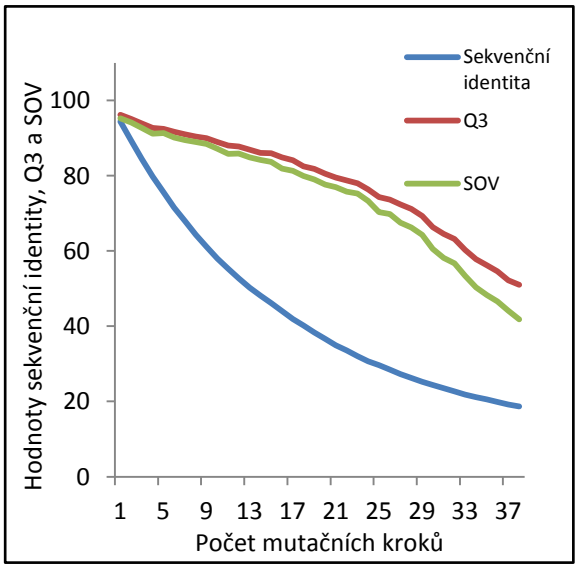
Ukázalo se tedy, že sekundární struktura proteinů je robustní ve smyslu odolnosti vůči náhodným mutacím.

Do budoucna hodlám v experimentech pokračovat s některými změněnými parametry. Bude možné změnit sílu mutace nebo použít jinou substituční matici. Také plánuji provést experiment na dalších datasetech například na CB513 obsahující již větší počet proteinů. Dalším plánovaným vylepšením je použití nástroje MAPP pro eliminaci škodlivých mutací, čímž by došlo k lepší simulaci selekčního tlaku, který v přírodě připouští jen minimum aminokyselinových záměn zhoršujících funkci.

A



B



Obrázek 1: Obrázek 1: Výsledky experimentů: (A) Graf závislosti hodnoty Q3 na sekvenční identitě, (B) graf závislosti sekvenční identity, Q3 a SOV na počtu provedených mutačních kroků během in-silico mutace.

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] SCHAEFER, Christian, SCHLESSINGER, Avner a ROST, Burkhard. Protein secondary structure appears to be robust under in silico evolution while protein disorder appears not to be. *Bioinformatics*, 2010, **26**(5), 625-631. ISSN 1367-4803.

[2] ZEMLA, Adam, VENCLOVAS, Česlovas, FIDELIS, Krzysztof a ROST, Burkhard. A Modified Definition of Sov, a Segment-Based Measure for Protein Secondary Structure Prediction Assessment. *Proteins: Structure, Function, and Bioinformatics*, 1999, **34**(2), 220-223. ISSN 0887-3585.

[3] MCGUFFIN, Liam J., BRYSON, Kevin a JONES, David T. The PSIPRED protein structure prediction server. *Bioinformatics*, 2000, **16**(4), 404-405. ISSN 1367-4803.

[4] ROST, Burkhard. Twilight zone of protein sequence alignments. *PEDS: protein engineering design & selection*, 1999, **12**(2), 85-94. ISSN 1741-0126.