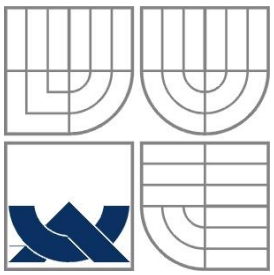
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

Fakulta informačních technologií
Faculty of Information Technology

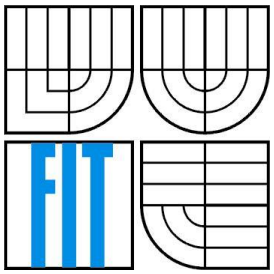
TECHNICKÁ SPRÁVA

TECHNICAL REPORT

Brno, 2017 Juraj Ondrej Dúbrava



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER SYSTEMS

PREDIKCIA VPLYVU AMINOKYSELINOVÝCH SUBSTITÚCIÍ NA STABILITU PROTEÍNU

PREDICTION THE EFFECT OF AMINO ACID SUBSTITUTIONS ON PROTEIN STABILITY

TECHNICKÁ SPRÁVA
TECHNICAL REPORT

AUTOR PRÁCE
AUTHOR

Juraj Ondrej Dúbrava

VEDÚCI PRÁCE
SUPERVISOR

Ing. Tomáš Martínek, Ph.D.

BRNO 2016

Abstrakt

V tejto práci sa zaoberáme témou predikcie vplyvu aminokyselinových substitúcií na stabilitu proteínu. Proteínové mutácie majú vplyv na stabilizáciu a rovnako na destabilizáciu proteínu. Na predikovanie ich vplyvu na stabilitu proteínu vzniklo v poslednom čase množstvo nástrojov založených na strojovom učení. Naším hlavným cieľom bolo navrhnúť a vytvoriť metaprediktor vplyvu mutácií na stabilitu proteínu založený na technikách strojového učenia. Jednotlivé časti opisujú problematiku stability, tvorbu trénovacieho datasetu, návrh a implementáciu metaprediktoru. Záver je venovaný plánom budúceho vývoja nástroja a zhrnutiu dosiahnutých výsledkov.

Abstract

In this work, we discuss the effects of amino acid substitutions on protein stability. These mutations can either stabilize or destabilize the protein structure. To predict the effects of mutations on protein stability, many tools based on the machine learning techniques have been recently developed.
Our main goal was to design and develop meta-predictor for the prediction the effects of mutations on protein stability using machine learning techniques. Individual chapters of this text will discribe the area of protein stability, the construction of training dataset and design and implementation of the developed meta-predictor. In the final chapter, our future plans in the developmet of the prediction tool and summarization of the current results will be presented.

Klúčová slova

Proteínová stabilita, aminokyseliny, konzervovanosť, predikcia, strojové učenie, dataset, metaprediktor

Keywords

Protein stability, amino acids, conservation, prediction, machine learning, dataset, metapredictor

Citácia

Juraj Ondrej Dúbrava: Predikcia vplyvu aminokyselinových substitúcií na stabilitu proteínu, technická správa, Brno, FIT VUT v Brně, 2017

PREDIKCIA VPLYVU AMINOKYSELINOVÝCH SUBSTITÚCIÍ NA STABILITU PROTEÍNU

Prehlásenie

Prehlasujem, že som túto technickú správu vypracoval samostatne pod vedením Ing. Tomáša Martínka, Ph.D.
Uviedol som všetky literárne pramene a publikácie, z ktorých som čerpal.

.....
Juraj Ondrej Dúbrava
22.1.2017

Poďakovanie

Rád by som sa poďakoval Ing. Tomášovi Martínkovi, Ph.D. za odborný dohľad nad touto prácou. Osobitne by som sa chcel poďakovať Ing. Milošovi Musilovi za odborné rady pri vývoji predikčného nástroja.

Obsah

1 Úvod

Táto práca sa zaoberá problematikou predikcie vplyvu aminokyselinových substitúcií na stabilitu proteínov. Reťazec proteínu je tvorený veľkým počtom aminokyselín, ktorých poradie v reťazci je veľmi dôležitým parametrom v otázke funkcie proteínu. Ak je nejaká aminokyselina v reťazci nahradená inou, jedná sa o jednobodovú mutáciu pôsobiacu na celý proteín. Výrazne sa môže zmeniť jeho funkcia a takisto jeho stabilita, ktorou sa v tejto práci zaoberáme. Mutácie môžu viesť k stabilizácií proteínu, čo môže napríklad viesť k väčšej odolnosti voči vyšším teplotám alebo nepriaznivému prostrediu. Na druhej strane môže mutácia spôsobiť destabilizáciu proteínu, čo nie je veľmi chcený účinok pôsobiacej mutácie.

Vzhľadom na skutočnosť, že stabilita proteínu je úzko prepojená s jeho funkciou, je téma predikcie vplyvu mutácií na stabilitu proteínu veľmi aktuálnou v oblasti bioinformatiky. Posledné roky boli z pohľadu vývoja rôznych predikčných nástrojov veľmi bohaté. Mnohé z vyvinutých nástrojov využívajú stále častejšie používané metódy strojového učenia, ktoré si v oblasti bioinformatiky nachádzajú široké uplatnenie. Nedá sa však povedať, že niektorý z týchto nástrojov pracuje s takou presnosťou predikcie, akú by sme si predstavovali. Je to spôsobené najmä obmedzeným množstvom relevantných a spoľahlivých dát, ktoré máme na zostavovanie trénovacích datasetov pre takéto metódy. Biologické databázy taktiež trpia mnohými nedostatkami a sme tak do značnej miery v tomto smere obmedzení. Zlepšenie predikcie stability pomocou nástrojov využívajúcich techniky strojového učenia aj s obmedzeným množstvom dát môže viesť k zlepšeniu v oblasti návrhu nových a stabilnejších proteínov, tvorbe nových účinnejších liečiv a takisto pri štúdií rozmanitých chorôb.

2 Proteínová stabilita

Proteíny môžeme charakterizovať ako základné stavebné prvky každého živého organizmu, ktoré vykonávajú množstvo rozličných funkcií. Do týchto funkcií patria napríklad replikácia DNA, katabolické alebo metabolické reakcie, transport molekúl z jedného miesta na iné. Pozostávajú z jedného alebo viacerých reťazcov, ktoré sú tvorené aminokyselinami. Poradie jednotlivých aminokyselín v proteínovom reťazci určuje funkciu a tiež štruktúru proteínu.

Stabilita proteínov je jednou z kľúčových vlastností, ktorá určuje aplikovateľnosť proteínu pod vplyvom drsných okolitých podmienok. Je veľmi úzko prepojená so štruktúrou proteínu. Zbalený stav proteínovej štruktúry je stabilizovaný rozličným typom interakcií ako napríklad elektrostatickými, hydrofóbnymi, van der Waalsovými a vodíkovými. Denaturovaný (unfolded) stav je ovplyvnený najmä entropickými a neentropickými voľnými energiami [7].

Väčší výskyt týchto interakcií vedie k stabilnejšiemu proteínu, ktorý je potom schopnejší zvládnuť napríklad extrémne teploty, kyslé alebo neutrálne pH či nepriaznivý vplyv organických rozpúšťadiel.

2.1 Meranie stability

Stabilita proteínu je možné merať ako zmenu tzv. Gibbsovej voľnej energie (ΔΔG), čo znamená rozdiel medzi voľnými energiami pri prechode zo stabilnej konformácie do denaturovaného stavu a naopak. V laboratóriách sa používa niekoľko techník na meranie proteínovej stability:

- Cirkulárny dichroizmus
- Diferenciálna skenovacia kalorimetria
- Absorpcia svetla
- Fluorescencia
- Jadrová magnetická rezonancia

Dôležitým faktom pre meranie stability proteínu a najmä jej zmeny je predikcia zmeny stability pod vplyvom aminokyselinových substitúcií. Ide o meranie a predikciu zmeny voľnej Gibssovej energie (ΔΔG) medzi pôvodným proteínom a mutantom. Väčšia snaha pri predikcií môže viesť k zlepšeniu návrhu nových odolnejších proteínov alebo pri štúdií rozličných chorôb.

2.2 Existujúce nástroje

V posledných rokoch bolo vyvinutých množstvo nástrojov na predikciu vplyvu mutácií na stabilitu proteínu. Tieto existujúce nástroje môžeme obecne rozdeliť na tie, ktoré využívajú techniky strojového učenia a metódy využívajúce energetické funkcie. Metódy využívajúce energetické funkcie môžu byť ďalej rozdelené na tie využívajúce tzv. fyzikálny potenciál, ktoré sa snažia simulovať základné sily medzi atómami. Ďalej sú to metódy využívajúce tzv. štatistický potenciál. Najväčšou nevýhodou metód, ktoré využívajú fyzikálny potenciál je ich vysoká výpočtová náročnosť.

Druhou veľkou skupinou metód sú tie, ktoré využívajú princípy predikcie založené na strojovom učení. Takéto nástroje musia byť najprv natrénované na sade proteínov a ich mutantov, pri ktorých bola experimentálne nameraná hodnota $\Delta\Delta G$. Takéto nástroje môžu dosahovať vysokú úspešnosť, ktorá nemusí byť skutočná, pretože nástroj bol trénovaný na obmedzenom množstve dát a mohlo dôjsť k pretrénovaniu.

Prehľad existujúcich nástrojov:

Strojové učenie

- AUTO-MUTE
- I-Mutant
- iPTREE-STAB
- EASE
- mCSM
- MAESTRO

Energetická funkcia – fyzikálny potenciál

- CC/PBSA
- ERIS
- Rosetta
- CUPSAT

Energetická funkcia – štatistický potenciál

- PopMuSiC
- DMutant
- FoldX

3 Dataset

Vytvorenie spoľahlivého, dostatočne veľkého a rozmanitého trénovacieho datasetu je rozhodujúcou vlastnosťou pri tvorbe každého predikčného nástroja. Dáta pre náš testovací dataset boli do veľkej miery získané z databázy ProTherm [8]. ProTherm je najrozsiahlejšou databázou, ktorá zahŕňa termodynamické parametre ako napríklad Gibssovu voľnú energiu, tepelnú kapacitu alebo entalpiu, ktoré sú počítané ako rozdiel medzi zmutovaným a pôvodným proteínom. ProTherm je naviac prepojená aj s inými databázami, napríklad SWISS-PROT, PDB, PIR.

Databáza ProTherm bola naposledy aktualizovaná vo februári v roku 2013 a aktuálne obsahuje približne 26 000 záznamov jedno a viacbodových mutantov navrhnutých nad viac ako 740 unikátnymi proteínmi, získaných rôznymi experimentálnymi technikami.

Pre tvorbu testovacích a trénovacích datasetov pre existujúce nástroje predikujúce vplyv aminokyselinových substitúcií na stabilitu proteínu je najpoužívanejším zdrojom dát práve databáza ProTherm. V súčasnom stave však trpí radou serióznych nedostatkov. Aby sme sa vysporiadali s problémami tejto databázy, vyextrahovali sme iba mutácie, u ktorých sú uvedené zmeny stability a overili všetky zdroje. Najväčšími problémami pri získavaní dát boli:

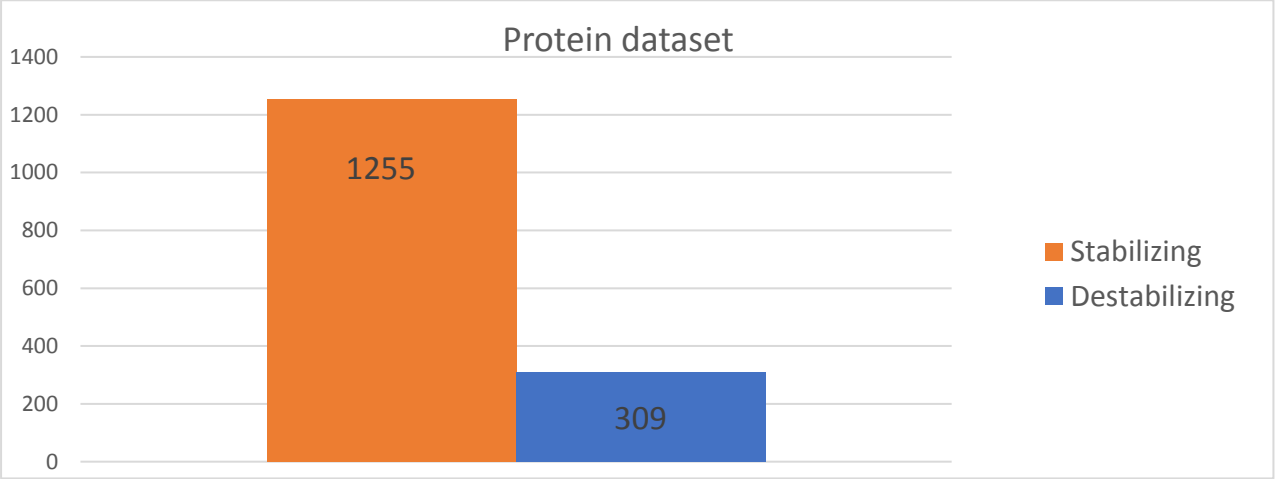
- Chýbajúca hodnota $\Delta\Delta G$
- Opačné znamienka $\Delta\Delta G$

Aby bol vytvorený dataset spoľahlivý a aby eliminoval možné experimentálne chyby v nameraných hodnotách zmeny Gibssovej voľnej energie ($\Delta\Delta G$), záznamy s hodnotami $\Delta\Delta G$ z intervalu (-0.5,0.5) boli odstránené. Záznamy s $\Delta\Delta G \geq 0.5$ kcal.mol^{-1} boli označené za destabilizujúce a záznamy s $\Delta\Delta G \leq 0.5$ kcal.mol^{-1} boli označené za stabilizujúce. Tento rozhodovací prah bol zvolený podľa tvrdenia, že experimentálna chyba merania $\Delta\Delta G$ je približne 0.48 kcal.mol^{-1} [6].

V prípade, že pre jednu mutáciu bolo uskutočnených viac meraní, bol ponechaný iba záznam merania, ktoré prebehlo pod experimentálnou hodnotou pH, ktorá bola blízko fyziologickej hodnote pH 7.

3.1 Parametre datasetu

Štruktúru výsledného datasetu je možné vidieť na priloženom grafe.



Obr.1 Štruktúra testovacieho datasetu

Z grafu vyplýva, že výsledný testovací / trénovací dataset obsahuje celkovo 1564 záznamov mutácií, z ktorých 1255 je stabilizujúcich a 309 destabilizujúcich jednobodových mutácií.
Každý záznam o mutácií obsahuje identifikátor databázy PDB pre daný proteín, reťazec, pozíciu mutácie v sekvencií a takisto pôvodnú aminokyselinu na danej pozícií a aminokyselinu, ktorá sa na danej pozícií nachádza po mutácií a rovnako údaj o zmene voľnej Gibssovej energie (ΔΔG). Tieto základné údaje sme rozšírili o štruktúrne a fyzikálno-chemické parametre určené na zlepšenie objektívnosti predikovaného výsledku.

Popis jednotlivých parametrov testovacieho datasetu:

1. Zmena polarity aminokyseliny

Tento údaj poskytuje informáciu o zmene v polarite aminokyseliny po mutácií oproti polarite pôvodnej aminokyseliny. Údaj sme dostali z tabuľkových hodnôt pre skúmané aminokyseliny. Uvažovali sme len polárne a nepolárne aminokyseliny.

2. Zmena náboja aminokyseliny

Poskytuje informáciu o zmene náboja pôvodnej aminokyseliny a aminokyseliny po mutácií. Výsledný údaj sme opäť získali z tabuľkových hodnôt pre dané aminokyseliny. Uvažovali sme negatívny, pozitívny a neutrálny náboj.

3. Zmena indexu hydrofobicity aminokyseliny

Údaj o výslednej zmene hydrofobicity aminokyseliny po mutácií a pôvodnej aminokyseliny. Číselný údaj značí rozdiel medzi tabuľkovými hodnotami týchto porovnávaných aminokyselín

4. Zmena veľkosti aminokyseliny

Údaj o zmene veľkosti aminokyseliny po mutácií proteínu a pôvodnej aminokyseliny. Aminokyseliny sme rozdelili do 3 intervalov na základe ich veľkosti. V jednom intervale sa tak nachádzajú len aminokyseliny, pre ktoré platí, že rozdiel ich veľkostí je menší, nanajvýš rovný 50. Následne sme zisťovali, do akého z patričných intervalov patrí pôvodná aminokyselina a aminokyselina po mutácií, z tohto údaju bolo možné určiť zmenu buď z veľkej na malú, opačnú zmenu alebo žiadnu zmenu, ak aminokyseliny patrili do jedného veľkostného intervalu.

5. Konzervovanosť

Jeden z najdôležitejších parametrov v zázname mutácie. Informácia o konzervovanosti pozície na ktorej došlo k mutácií je reprezentovaná percentuálnym vyjadrením zastúpenia pôvodnej aminokyseliny a aminokyseliny po výslednej mutácií na zadanej pozícií. Ak výsledné percentuálne zastúpenie pôvodnej aminokyseliny bolo vyššie ako 60%, skúmaná pozícia bola označená za konzervovanú. Pri prevyšujúcom počte iných aminokyselín bola pozícia, na ktorej došlo k mutácii označená ako nekonzervovaná. Percentuálnu hodnotu zastúpenia vstupných aminokyselín sme získali z viacnásobného zarovnania nášho testovaného proteínu s jemu homológnymi sekvenciami. Na získanie homológnych sekvencií daného proteínu sme použili standalone verziu nástroja BLAST-p [1] a následne bolo potrebné získať viacnásobné zarovnanie získaných sekvencií proteínov pomocou nástroja Clustal Omega [2]. Mutácie na konzervovanej pozícií sú z hľadiska skúmania veľmi zaujímavé, pretože konzervované aminokyseliny bývajú dôležité napríklad pre stabilitu, aktivitu alebo schopnosť proteínu vytvoriť terciárnu štruktúru.

6. Sekundárna štruktúra proteínu

Údaj o sekundárnej štruktúre proteínu sme získali použitím modulu DSSP prítomného v knižnici BioPython. Pre zjednodušenie sme každému proteínu priradili tento údaj len z 3 možností, ktorými sú helix, coil a sheet.

7. ASA (accessible surface area)

Jedná sa o údaj, ktorý udáva plochu aminokyseliny, ktorú môže dosiahnuť rozpúšťadlo. Na výpočet tohto parametru sme opäť použili modul DSSP prítomný v knižnici BioPython.

4 Vývoj

4.1 Testovanie datasetu

Na prvotné otestovanie nami zostaveného testovacieho datasetu sme sa rozhodli použiť nástroj WEKA (Waikato Environment for Knowledge Analysis) [3]. Jedná sa o sadu vizualizačných nástrojov a algoritmov na analýzu dát. V našom prípade nás zaujímali algoritmy strojového učenia, na ktorých sme testovali nami vytvorenú dátovú sadu a hľadali optimálnu metódu použiteľnú ako základ implementácie metaprediktoru. Pred zahájením testovania predikcie bolo potrebné zvoliť inú reprezentáciu hodnôt jednotlivých parametrov záznamu mutácie, pretože pre WEK-u by tento formát nebol zrozumiteľný. WEK-a používa na reprezentáciu datasetov formát súboru typu ARRF, ktorý sa skladá z atribútov, pri ktorých je nutné definovať množinu hodnôt, ktoré môžu nadobudnúť a následne nasleduje množina dát. Zo všetkých potrebných parametrov sme zostavili testovací súbor, ktorý obsahuje atribúty zodpovedajúce parametrom jedného proteínového záznamu.

Po príprave súboru s dátami pre WEK-u sme mohli prejsť k fáze testovania jednotlivých algoritmov strojového učenia. WEKA obsahuje rôzne typy klasifikátorov, ktoré je možné použiť. Pre prípad nášho datasetu však bolo potrebné mierne upraviť testovacie nastavenia, pretože destabilizujúce mutácie boli v značnom nepomere k stabilizujúcim. Nastavili sme tzv. cost sensitive matrix a zaistili tým, že zle predikované stabilizujúce mutácie sú viac postihované.

Po otestovaní predikcie na ponúknutých metódach strojového učenia WEK-y sme ponechali 7 možností s najlepšími výsledkami. Nasledujúca tabuľka obsahuje namerané výsledky testovania najlepších metód strojového učenia na našom datasete.

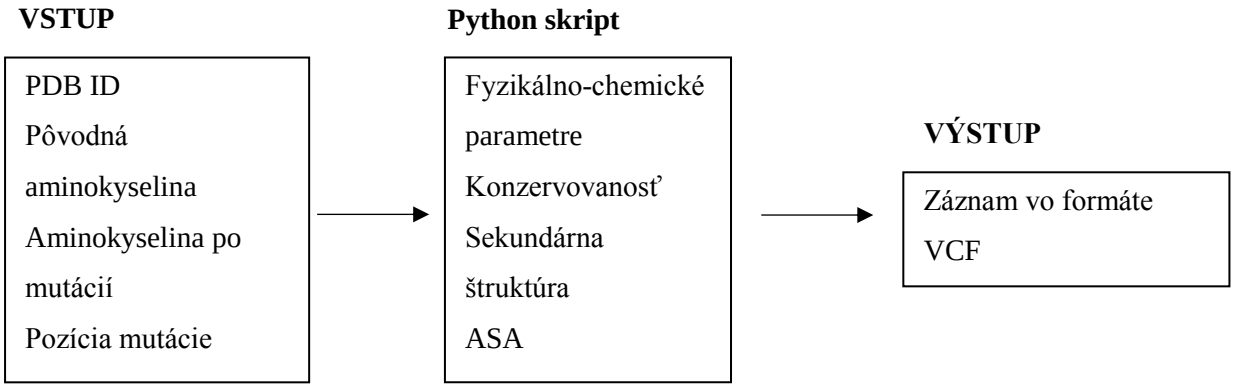
Metóda strojové učenia	TP rate	FP rate	Accuracy
Naive Bayes	0,784	0,719	0,765
LibSVM	0,774	0,774	0,6
SMO	0,774	0,774	0,6
DecisionTable	0,774	0,774	0,6
RandomForest	0,793	0,692	0,797
RandomTree	0,793	0,574	0,766
J48	0,774	0,774	0,6

Tab.1 Výsledky predikcie najlepších metód strojového učenia

Najdôležitejším parametrom je hodnota Accuracy udávajúca počet relevantných hodnôt. Po uskutočnení niekoľkých testovaní jednotlivých metód strojového učenia preukázala najlepšie výsledky metóda RandomForest [4].

4.2 Vstup a výstup

Po fáze otestovania datasetu sme pristúpili k fáze implementácie. Pre metaprediktor je dôležité pripraviť vstup, na základe ktorého vyhodnotí mutáciu ako stabilizujúcu alebo destabilizujúcu. Na získanie údajov o všetkých potrebných parametroch proteínovej mutácie sú použité webové alebo standalone verzie potrebných nástrojov. Získavanie a spracovanie údajov prebieha v skripte implementujúcom metaprediktor. Pripravený skript má 4 vstupné parametre. Prvým parametrom je PDB identifikátor proteínu, nasleduje pôvodná aminokyselina a aminokyselina po mutácií, posledným parametrom je pozícia na ktorej došlo k mutácií. Na nasledujúcej schéme je možné vidieť proces spracovania vstupných parametrov na výstupné.



Výstupný súbor vo formáte VCF ešte nie je konečným výsledkom, ale bude ďalej použitý ako vstup pre nami zvolenú metódu strojového učenia na predikovanie stability na základe vstupných informácií obsiahnutých vo VCF súbore.

4.3 Implementácia metaprediktoru

Vyvíjaný metaprediktor sme sa rozhodli implementovať vo forme skriptu v jazyku Python, pretože pre tento jazyk existuje rada knižníc implementujúcich rôzne bioinformatické nástroje. V tejto fáze vývoja v sebe skript zahŕňa výpočet a získavanie údajov o proteínovej mutácií. Na zjednodušenie spracovania a zisk potrebných dát sme sa rozhodli použiť knižnicu BioPython. Zahŕňa v sebe množstvo bioinformatických modulov a nástrojov výrazne zjednodušujúcich spracovanie dát z tejto oblasti.
Z programového hľadiska uľahčuje prístup k online databázam biologických informácií ako napríklad NCBI. Samostatné moduly rozširujú možnosti knižnice BioPython v oblasti viacnásobného zarovnania, proteínovej štruktúry alebo strojového učenia.

Skript v súčasnej podobe automatizuje získavanie potrebných údajov, ktoré budú tvoriť záznam mutácie v niekoľkých krokoch. Po získaní vstupných parametrov je možné okamžite vypočítať fyzikálno-chemické parametre proteínu. V nasledujúcom kroku začína skript v získavaní údajov z online databázy PDB. Podľa vstupného PDB identifikátora stiahne prislúchajúci FASTA súbor obsahujúci PDB identifikátor a sekvenciu proteínu. Po získaní FASTA súboru nasleduje stiahnutie PDB súboru obsahujúceho podrobnejšie informácie o proteíne. FASTA súbor je vstupným parametrom nástroja BLAST-p. Jedná sa o nástroj prehľadávajúci online databázy, v ktorých sa snaží nájsť sekvencie podobné vstupnej sekvencií využívajúc k tomu lokálne zarovnanie týchto sekvencií. BLAST-p je dostupný ako webový nástroj a k dispozícií je aj standalone verzia tohto nástroja, ktorú sme sa rozhodli pre naše účely použiť. Výstupom nástroja je XML súbor s homológnymi sekvenciami proteínov, ich počet sme zvolili na hodnotu 250. Sekvencie homológnych proteínov sme z výstupného XML súboru vyextrahovali a vytvorili z nich nový FASTA súbor obsahujúci názvy jednotlivých sekvencií a samotné sekvencie.

Pripravený FASTA súbor je v nasledujúcom kroku použitý ako vstup pre nástroj Clustal Omega, ktorý bol tiež použitý v standalone verzií. Jedná sa o nástroj produkujúci viacnásobné zarovnanie získaných homológnych sekvencií obsiahnutých vo vstupnom FASTA súbore. Výstupom nástroja Clustal Omega môže byť niekoľko formátov súborov. Zvolili sme Vienna formát, ktorý však ešte nie je možné použiť na získanie údaju o konzervovanosti pozície na ktorej došlo k mutácií. V poslednom kroku fázy prípravy sme z Vienna súboru vyextrahovali iba zarovnané sekvencie bez ich názvu a opäť z nich vytvorili samostatný súbor. Takto pripravený výsledný súbor je vhodný na zistenie konzervovanosti. Po získaní údaju o konzervovanosti pozície nasleduje získavanie informácie o sekundárnej štruktúre proteínu a rovnako výpočet hodnoty ASA. Informácie o obidvoch parametroch sme získali pomocou modulu DSSP prítomného v knižnici BioPython.
Ak všetky kroky získavania a výpočtu dát prebehli bez problémov, zo získaných údajov je vytvorený VCF súbor so všetkými parametrami proteínovej mutácie.

Skript prešiel fázou testovania na dátovej sade proteínových mutácií. Porovnávali sme výstupné hodnoty parametrov získaných pomocou skriptu s hodnotami parametrov v zostavenom datasete. Výstupy skriptu sa zhodovali s hodnotami v datasete, až na hodnoty ASA, ktoré sa líšili v rádoch stotín až tisícin oproti hodnotám prítomných v zostavenom datasete.

5 Plány do budúcna

V súčasnej dobe je metaprediktor stále vo fáze vývoja. Obsahuje len časť, ktorá zabezpečuje získavanie a spracovanie potrebných údajov do podoby VCF súboru. Po otestovaní datasetu na dostupných metódach strojového učenia v nástroji WEKA máme v pláne využiť pre potreby metaprediktoru metódu RandomForest, pretože v testoch preukázala najlepšie výsledky predikcie.

5.1 Scikit-learn

Pre implementovanie métody strojového učenia RandomForest máme v pláne využiť knižnicu Scikit-learn. Ide o knižnicu strojového učenia pre programovací jazyk Python. Zahŕňa v sebe rozličné klasifikátory, zhlukovacie algoritmy ako napríklad support vector machines, random forests a gradient boosting.

5.2 Ensemble systémy

Metódy strojové učenia si v posledných rokoch našli možnosti širokého uplatnenia v mnohých bioinformatických aplikáciach. Avšak, najmä kvôli nedostatočnému

množstvu dát v trénovacích datasetoch, rozličnosti v dátach alebo kvôli problému pretrénovania môžu tieto techniky viesť k prehnaným výsledkom nezodpovedajúcich realite. Sľubne črtajúcim sa riešením je skombinovanie výstupov, ktoré sa získajú použitím viacerých samostatných klasifikátorov. Takto získané výsledky môžu byť rôzne, niekedy až protichodné, ale zvyšujú robustnosť a správnosť predikovaných výsledkov. Takýto prístup je nazývaný ensemble.

V našom prípade nastáva problém nedostatočného počtu proteínových záznamov v datasete. Aby sme predišli možnosti pretrénovania na obmedzenom množstve dát a zaistili tak lepší a realistickejší výsledok predikcie, naše plány pre metaprediktor zahŕňajú aj využitie spomenutých ensemble systémov. Plánujeme rozdelenie dát na menšie celky, ktoré plánujeme otestovať na viacerých vybratých klasifikátoroch, aby sme zistili rozdiely použitia jedného klasifikátora oproti výsledkom z viacerých klasifikátorov. Takáto stratégia sa nazýva bagging (bootstrap aggregation) [5] a je jednou z najstarších, ľahkou na implementáciu a zároveň intuitívna. Diverzita je dosiahnutá náhodným rozdelením dát do menších celkov použitých ako trénovacie dáta pre viacero klasifikátorov. Bagging je silným mechanizmom v prípade, ak máme k dispozícií obmedzené množstvo spoľahlivých dát. Jedným z najpopulárnejších algoritmov založených na stratégií bagging je metóda Random forest, uplatňujúca sa v mnohých bioinformatických aplikáciach.

6 Záver

Problematika predikcie vplyvu aminokyselinových substitúcií je v dnešnej dobe pre oblasť bioinformatiky veľmi zaujímavá. V nedávnej dobe vzniklo niekoľko nástrojov slúžiacich na predikciu stability založených na strojovom učení. Našou snahou bolo vytvoriť metanástroj k predikovaniu stability, ktorý by bolo možné porovnať s existujúcimi nástrojmi. Náš nástroj bude k predikcií tiež využívať metódu strojového učenia. Dôležitou súčasťou nástrojov využívajúcich techniky strojového učenia je sada trénovacích dát. Vytvorili sme testovací dataset, ktorého nevýhodou je obmedzené množstvo dát. Overenie predikcie na testovacej sade sme uskutočnili pomocou nástroja WEKA. Cieľom testovania bolo zistenie výsledkov predikcie u rôznych metód strojového učenia. Z meraní vyšla najlepšie metóda RandomForest uplatňujúca sa v mnohých bioinformatických aplikáciach. Samotný nástroj bol navrhnutý ako skript v jazyku Python. V aktuálnej fáze vývoja zahŕňa iba výpočet potrebných parametrov mutácie, ktoré budú použité ako vstup pre metódu RandomForest. V ďalšom vývoji nástroja je plánované doimplementovanie metódy RandomForest pomocou knižnice Scikit-learn a rovnako aj využitie ensembled systémov vzhľadom na obmedzené množstvo dát.

Literatúra

[1] Altschul, S. F., Madden, T. L., Schäffer, A. A., et al.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research, 1997, 3389-3402.

[2] Sievers, F., Higgins, D. G.: Clustal Omega, accurate alignment of very large numbers of sequences. Multiple sequence alignment methods. Springer, 2014, 105-116.

[3] Frank, E., Hall, M., Trigg, L., et al.: Data mining in bioinformatics using Weka. Bioinformatics, 2004, 2479-2481.

[4] Qi, Y.: Random forest for bioinformatics. In Ensemble machine learning. Springer US, 2012, 307-323.

[5] Breiman, L.: Bagging predictors. Machine learning, Springer, 1996.

[6] Khatum, J., Khare, S. D., Dokholyan, N. V.: Can contact potentials reliably predict stability of proteins? J Mol Biol., 2004.

[7] Gromiha, M. M.: Protein Bioinformatics. Elsevier, 2010, ISBN 978-81-312-2297-3

[8] Bava, K. A., Gromiha, M. M., Uedaira, H., et al.: ProTherm, version 4.0: Thermodynamic database for proteins and mutants. Nucleic acids research, 2004.