



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV INFORMAČNÍCH SYSTÉMŮ

DEPARTMENT OF INFORMATION SYSTEMS

**PREDIKCE VLIVU MUTACE NA ROZPUSTNOST
PROTEINŮ**

PREDICTION OF THE EFFECT OF MUTATION ON PROTEIN SOLUBILITY

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

AUTOR PRÁCE

AUTHOR

JÚLIUS MARKO

VEDOUCÍ PRÁCE

SUPERVISOR

Ing. JIŘÍ HON

BRNO 2020

Zadání bakalářské práce



Student: **Marko Július**

Program: Informační technologie

Název: **Predikce vlivu mutace na rozpustnost proteinů**
Prediction of the Effect of Mutation on Protein Solubility

Kategorie: Bioinformatika

Zadání:

1. Seznamte se s problematikou rozpustnosti proteinů a existujícími predikčními nástroji
2. Vytvořte trénovací a testovací datovou sadu vlivu mutace na rozpustnost proteinů na základě dostupných dat.
3. Navrhněte nástroj pro predikci vlivu mutace na rozpustnost proteinů a proveďte jeho implementaci.
4. Činnost nástroje zhodnoťte na testovací datové sadě a srovnajte jej s existujícími nástroji.
5. Zhodnoťte dosažené výsledky a diskutujte možnosti dalšího pokračování projektu.

Literatura:

- KLESMITH, Justin R., John-Paul BACIK, Emily E. WRENBECK, Ryszard MICHALCZYK a Timothy A. WHITEHEAD. Trade-offs between Enzyme Fitness and Solubility Illuminated by Deep Mutational Scanning. *Proceedings of the National Academy of Sciences*. 2017, 114(9), 2265-2270
- WRENBECK, Emily E., Matthew A. BEDEWITZ, Justin R. KLESMITH, Syeda NOSHIN, Cornelius S. BARRY a Timothy A. WHITEHEAD. An Automated Data-Driven Pipeline for Improving Heterologous Enzyme Expression. *ACS Synthetic Biology*. 2019, 8(3), 474-481
- Dále dle pokynů vedoucího.

Pro udělení zápočtu za první semestr je požadováno:

- Splnění bodů 1 a 2 zadání.

Podrobné závazné pokyny pro vypracování práce viz <https://www.fit.vut.cz/study/theses/>

Vedoucí práce: **Hon Jiří, Ing.**

Vedoucí ústavu: Kolář Dušan, doc. Dr. Ing.

Datum zadání: 1. listopadu 2019

Datum odevzdání: 28. května 2020

Datum schválení: 22. října 2019

Abstrakt

Rozpustnosť proteínov je kľúčový problém, ktorý úzko súvisí s požiadavkou na výrobu funkčného proteínu. Prediktor vplyvu mutácie na rozpustnosť proteínov by mohol viesť k značnému šetreniu času a financií, pretože je schopný *in silico* predikcie mutácií, ktoré zlepšujú rozpustnosť ešte pred zahájením hlbokého mutačného testovania v laboratóriu. V tejto práci je prezentovaný nový prediktor vplyvu mutácie na rozpustnosť proteínov SoluProtMut, založený na metódach strojového učenia. Väčšina existujúcich nástrojov predikujú tento vplyv zo sekvencie aminokyselín. Okrem tejto sekvencie, prediktor prezentovaný v tejto práci využíva aj priestorovú štruktúru proteínu, čím sa môže výrazne zvýšiť jeho presnosť.

Abstract

Protein solubility is a key problem in production of functional proteins. Prediction of the effect of mutation on protein solubility could save a lot of time and money, as it would provide *in silico* prediction of solubility enhancing mutations before performing deep mutational scanning in laboratory. In this work, new predictor of the effect of mutation on protein solubility SoluProtMut is introduced that is based on machine learning methods. Most of the existing predictors predict the effect from the amino acid sequence. In addition to the sequence, the tool presented in this work also uses the spatial structure of the protein, which can significantly increase its accuracy.

Kľúčové slová

rozpustnosť, proteín, mutácia, predikcia, strojové učenie, SoluProtMut

Keywords

solubility, protein, mutation, prediction, machine learning, SoluProtMut

Citácia

MARKO, Július. *Predikce vlivu mutace na rozpustnost proteinů*. Brno, 2020. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Ing. Jiří Hon

Predikce vlivu mutace na rozpustnost proteinů

Prehlásenie

Prehlasujem, že som túto bakalársku prácu vypracoval samostatne pod vedením pána Ing. Jiřího Hona. Uviedol som všetky literárne pramene, publikácie a ďalšie zdroje, z ktorých som čerpal.

.....

Július Marko

28. mája 2020

Podakovanie

Rád by som sa poďakoval Ing. Jiřímu Honovi za odbornú pomoc a vedenie bakalárskej práce. Ďalej by som sa chcel poďakovať Loschmidtovým laboratóriám, ktoré mi poskytli odbornú literatúru a organizácii MetaCentrum za poskytnutie výpočtových zdrojov.

Obsah

1	Úvod	2
2	Predikcia vplyvu mutácie na rozpustnosť proteínov	3
2.1	Štruktúra proteínov	3
2.2	Rozpustnosť proteínov	5
2.3	Vplyv mutácie na rozpustnosť	6
2.4	Dátové sady	6
2.5	Existujúce nástroje	9
3	Tvorba dátových sád	11
3.1	Trénovacia sada	11
3.2	Testovacia sada	14
4	Návrh nástroja	15
4.1	Výber predikčných vlastností	15
4.2	Návrh nástroja	18
5	Implementácia nástroja	21
5.1	Použité frameworky a knižnice	21
5.2	Implementácia	21
6	Experimenty	25
6.1	Výber hyper-parametrov	25
6.2	Regularizácia	25
6.3	Náhodné lesy	27
6.4	Support Vector Machines	27
6.5	Vyhodnotenie	27
7	Záver	31
	Literatúra	33
A	Obsah priloženého DVD	38

Kapitola 1

Úvod

Proteíny majú zásadný význam pre biologické systémy a vykonávajú obrovské množstvo funkcií.

Rozpustnosť proteínov je oblasť, ktorá úzko súvisí s požiadavkou na výrobu funkčného proteínu. Funkčné proteíny sú potrebné vo viacerých oblastiach. Vo výskume, kde vedci potrebujú určité množstvo proteínu, aby mohli študovať jeho štruktúru, funkciu alebo interakciu s ostatnými proteínami; v priemysle, ak sa niektorý z proteínov ukáže byť zaujímavý a je potrebné vyrábať ho vo veľkých množstvách alebo vo farmaceutickom priemysle na zvýšenie rozpustnosti terapeutických proteínov, napríklad protilátok.

Pre praktické aplikácie je často potrebné vyrábať proteíny vo veľkých množstvách. Každý proteín má danú rozpustnosť – schopnosť rozpúšťať sa v rozpúšťadle. Výsledná zmes má určené maximálne množstvo rozpustnej látky – proteínu, pričom s ďalším navyšovaním koncentrácie začína agregovať – tvorba zhlukov. Takémuto proteínu hovoríme nerozpustný. Zmes, v ktorej je proteín agregovaný, je ešte možné zachrániť experimentálnymi metódami, avšak nie sú výhodné, pretože vyžadujú čas a ďalšie finančné prostriedky. Cieľom je teda vybrať taký proteín, ktorý má najvyššiu rozpustnosť už na začiatku, aby sa zvýšila šanca na rozpustnosť vo vyšších koncentráciách.

Jednou z možností, ako zvýšiť rozpustnosť proteínu je jeho úprava – mutácia proteínu, ktorá ale zachová jeho funkčnosť. V súčasnej dobe je možné vykonávať tisíce bodových mutácií v laboratóriách a zisťovať zmenu rozpustnosti [45]. Táto práca sa zaoberá možnosťou zameniť túto laboratórnu metódu za *in silico* predikciu.

Hlavným cieľom tejto práce je vytvoriť nástroj, ktorý predikuje vplyv mutácie na rozpustnosť proteínov pomocou metód strojového učenia.

Kapitola 2

Predikcia vplyvu mutácie na rozpustnosť proteínov

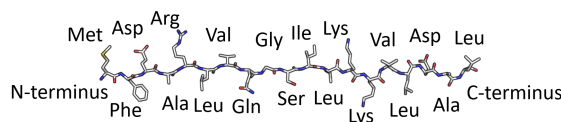
Kapitola sa zaoberá úvodom do problému predikcie vplyvu mutácie na rozpustnosť proteínov. Obsahuje základné biologické pojmy, problém rozpustnosti proteínov, vplyv mutácie na rozpustnosť, dostupné dátové sady a popis existujúcich nástrojov.

2.1 Štruktúra proteínov

Proteín alebo bielkovina je zlúčenina zložená zo sto a viac aminokyselín spojených navzájom peptidovou väzbou [11]. Z chemického hľadiska sú proteíny štrukturálne komplexné a funkčne sofistikované známe molekuly. Každý proteín má tvar a štruktúru, ktorá je daná postupnosťou aminokyselín v reťazci. Väčšinou sa proteín snaží zbaliť (protein folding) do tvaru, v ktorom je jeho voľná energia čo najmenšia [3]. Z hľadiska priestorového usporiadania sa delí štruktúra proteínov na nasledujúce úrovne výstavby:

Primárna štruktúra

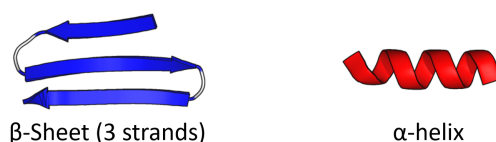
Primárna štruktúra zodpovedá poradiu jednotlivých aminokyselín v reťazci a jednoznačne determinuje vyššie úrovne štruktúry proteínov a ich biologickú aktivitu [11]. Zapisuje sa od N-konca po C-koniec typicky jednopísmenovými alebo trojpísmenovými skratkami aminokyselín. Aminokyseliny sú v reťazci spojené peptidovými väzbami, ktoré formujú polypeptidovú kostru, z ktorej vyčnievajú chemicky rôzne bočné reťazce. Bočný reťazec je časť aminokyseliny, ktorá nie je zapojená do formovania peptidových väzieb. Tento bočný reťazec dáva každej aminokyseline jeho unikátne vlastnosti: niektoré sú nepolárne a hydrofóbne (odpor voči vode), negatívne alebo pozitívne nabité, niektoré môžu byť chemicky reaktívne a pod.



Obr. 2.1: Primárna štruktúra s trojpísmenovými aminokyselinovými skratkami. N-terminus – N-koniec, C-terminus – C-koniec. Aminokyseliny sú spojené peptidovými väzbami. Charakter a sekvencia chemicky odlišných bočných reťazcov dáva každému proteínu jeho odlišné individuálne vlastnosti. Na obrázku je znázornený malý polypeptid tvorený iba 19 aminokyselinami – proteíny sú typicky tvorené reťazcami niekoľkých stoviek aminokyselín. [3]

Sekundárna štruktúra

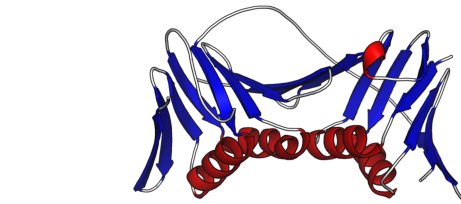
Sekundárna štruktúra udáva priestorové usporiadanie časti reťazca. Jediná forma väzby, ktorá sa uplatňuje pri vzniku a stabilizácii štruktúry sú vodíkové väzby. Tieto väzby stabilizujú usporiadanie reťazca buď vo forme α -skrutkovice, alebo vo forme β -skladaného listu. [11] α -skrutkovica má tvar pravotočivej skrutkovice, pri pohľade zhora tvorí skrutkovicu proteínová kostra, zatiaľ čo aminokyselinové zvyšky smerujú von. Medzi jednotlivými poschodiami skrutkovice vznikajú vodíkové mostíky. β -skladaný list (β -sheet) je tzv. plne extendovaný typ štruktúry, to znamená, že na rozdiel napr. od α -skrutkovice nie je príliš kompaktný, ale do všetkých strán pretiahnutý. Skladá sa z niekoľkých aminokyselinových reťazcov (peptidov), ktoré sú usporiadané vzájomne rovnobežne a dochádza medzi nimi k vzniku vodíkových väzieb (konkrétne medzi vodíkom amidovej skupiny a kyslíkom karbonylovej skupiny susedných reťazcov). Tým sa táto štruktúra stáva stabilnejšia. Susedné reťazce aminokyselín môžu byť paralelné alebo antiparalelné. Od tohto usporiadania sa odvíja presná štruktúra skladaného listu. [24]



Obr. 2.2: Sekundárna štruktúra. Polypeptidové reťazce sa často skladajú do jednej z dvoch usporiadaných opakujúcich sa foriem známych ako β -skladaný list (β -sheet) a α -skrutkovica (α -helix). [3]

Terciárna štruktúra

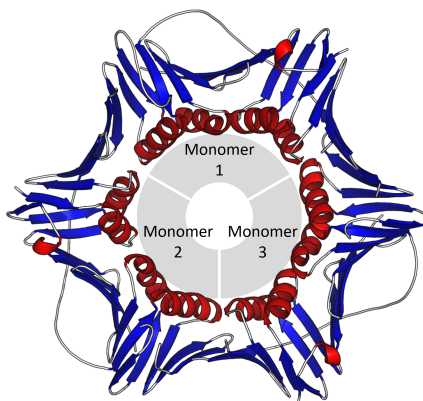
Terciárna štruktúra udáva celkové priestorové usporiadanie bielkoviny. Táto štruktúra je stabilizovaná interakciami medzi bočnými reťazcami jednotlivých nesusediacich aminokyselín [11] a zahŕňa omnoho dlhší úsek aminokyselín ako sekundárna štruktúra. Aj aminokyseliny, ktoré sú navzájom v sekvencii veľmi vzdialené, môžu v konečnom dôsledku interagovať v priestore a stabilizovať bielkovinu, prípadne tvoriť jej dôležité funkčné miesto [25]. Obsahuje jeden polypeptidový reťazec (kostra) s jednou alebo viacerými sekundárnymi štruktúrami – proteínovými doménami [1].



Obr. 2.3: Terciárna štruktúra znázorňuje formovanie polypeptidu do komplexného molekulárneho tvaru. Je to plná trojrozmerná konformácia vytvorená celým polypeptidovým reťazcom – vrátane α -skrutkovic, β -skladaných listov, náhodných cievok a akýchkoľvek ďalších slučiek a záhybov, ktoré sa tvoria medzi N- a C-koncami [3]. Je definovaná súradnicami atómov.

Kvartérna štruktúra

Veľká časť proteínov sú vlastne zostavy viacerých polypeptidových reťazcov. Ak je proteínová molekula vytvorená ako komplex viac ako jedného polypeptidového reťazca, potom je jej úplná štruktúra označená ako kvartérna štruktúra [3], ktorá udáva počet a kvalitu reťazcov – podjednotiek, ktoré tvoria príslušnú funkčnú bielkovinu. [11]. Zahŕňa organizácie od jednoduchých dimérov po veľké homooligoméry a proteínové komplexy [5].



Obr. 2.4: Kvartérna štruktúra ľudského svorkového proteínu PCNA (Proliferating cell nuclear antigen). Tento proteín slúži na upevňovanie DNA polymerázy k reťazcu DNA, čím zvyšuje procesivitu replikácie. Je to trimérny proteín – makromolekula zložená z troch menších podjednotiek, tzv. monomérov. (PDB: 1AXC)

2.2 Rozpustnosť proteínov

Rozpustnosť je základná biofyzikálna vlastnosť proteínov [17]. Ovplyvňuje ju viacero faktorov, ktoré môžeme rozdeliť na vonkajšie a vnútorné [18].

K vonkajším faktorom patrí pH, iónová sila, teplota a prítomnosť rôznych rozpúšťadlových prísad [30]. Tieto vonkajšie faktory môžu viesť k zvýšenej rozpustnosti [4, 34], avšak zmena podmienok roztoku nie je vždy dostatočná na zvýšenie rozpustnosti proteínu v požadovanom rozsahu [18].

Vnútorne faktory, ktoré ovplyvňujú rozpustnosť proteínov, sú definované predovšetkým aminokyselinami na povrchu proteínu, chýba však podrobné porozumenie toho, ako je možné zmeniť tieto vnútorné vlastnosti proteínu na zvýšenie jeho rozpustnosti [4, 34, 43].

K vnútorným vlastnostiam je taktiež možné zaradiť sklon proteínov k agregácii. Proteínová agregácia je jav, v ktorom dochádza k tvorbe nerozpustných proteínových zhlukov. Proteíny s náchylnosťou k agregácii majú teda negatívny dopad na rozpustnosť. Agregácia proteínov môže byť minimalizovaná pridaním vhodných chaperónov. Chaperóny predstavujú špeciálny typ proteínov, ktoré napomáhajú ostatným proteínom pri ich skladaní. Ďalším dôležitým faktorom je spôsob tvorby proteínov. Proteíny sú typicky produkované v organizmoch, v ktorých sa prirodzene nenachádzajú, ide o tzv. rekombinantnú tvorbu proteínov. Tvorbu rekombinantných proteínov umožňujú metódy genetickej modifikácie. Organizmus, resp. platforma, pomocou ktorej je proteín vytvorený, sa označuje pojmom expresný organizmus, resp. systém. Často používaným expresným organizmom je baktéria *E. coli*. So zmenou expresného systému však dochádza k zmene okolitých podmienok, v rámci ktorých sú proteíny tvorené. Preto sú prediktory a samotná predikcia rozpustnosti viazané na konkrétny expresný systém. Tento odsek bol prevzatý z práce [21].

2.3 Vplyv mutácie na rozpustnosť

Táto práca sa zaoberá vplyvom jednobodových mutácií na rozpustnosť proteínov. Jednobodová mutácia na úrovni proteínu je taká mutácia, kde sa na určitom mieste v aminokyselinovom reťazci proteínu zmení pôvodná aminokyselina na inú. Proteín, ktorý je prirodzene vyskytujúci sa v prírode (ďalej len ako „divoký typ“ alebo „wild-type“), je možné modifikovať (mutovať) – proteín sa potom označuje ako „mutant“. Mutácie však môžu ovplyvňovať skladanie proteínov a ich stabilitu, funkciu proteínov a interakcie proteín-proteín, ako aj expresiu proteínov a subcelulárnu lokalizáciu [29]. Pri návrhu nových proteínov je teda potrebné tieto vlastnosti brať do úvahy a vybrať také mutácie, ktoré zvyšujú rozpustnosť proteínu a zároveň ponechávajú jeho funkčnosť [17].

2.4 Dátové sady

Táto podkapitola obsahuje dostupné dátové sady, ktoré obsahujú dáta o proteínoch a vplyve mutácií na ich rozpustnosť.

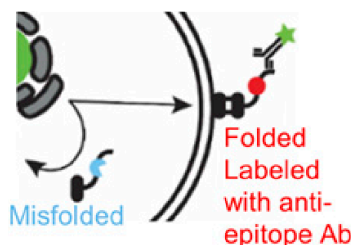
Klesmith et al. 2017

Táto dátová sada obsahuje dáta pre 2 proteíny – betalaktamáza (TEM-1.1) dlhá 263 aminokyselín, ktorá vznikla z betalaktamázy (TEM-1 BLA) vykonaním dvoch mutácií: na lokácii 70 v reťazci zo serínu na alanín, aby sa zrušila katalytická aktivita, pretože jedna z metód selekcie zahŕňa rast na beta-laktámových antibiotikách, druhá mutácia je destabilizačná a prebehla na lokácii 179 z kyseliny asparágovej na glycín, pretože TEM-1 BLA je stabilný pri selekčnej a skriningovej teplote 30-37 stupňov Celzia. Druhým proteínom je levoglukozankináza (LGK) dlhá 439 aminokyselín. Nad týmito proteínmi bolo vykonané hlboké mutačné skenovanie s použitím dvoch rôznych *in vivo* selekčných metód – zobrazenie na povrchu kvasiniek (YSD) a translokácia dvojarginínu (Tat). [17]

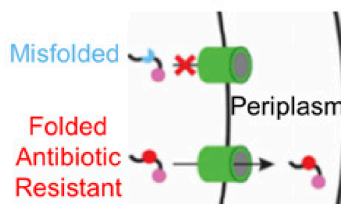
Expresný systém metódy YSD sú kvasinky. Hlavný princíp spočíva v tom, že pred úspešným zobrazením na povrch kvasiniek sú vznikajúce proteíny predmetom kontroly kvality. Proteíny, ktoré sa chybné zložili, sú označené na degradáciu a neskôr zničené pomocou

proteazómov. Väzba s fluorescenčnou protilátkou umožňuje rozlíšenie variantov, ktoré sa exprimujú na bunkovom povrchu od tých, ktoré nie. Obrázok 2.5 znázorňuje princíp tejto metódy. [17]

Tat, na rozdiel od YSD používa ako expresný systém baktériu *E. coli*. Táto metóda funguje na princípe, v ktorom sa proteín exportuje do bakteriálnej periplazmy iba vtedy, ak je správne poskladaný. Obrázok 2.6 znázorňuje princíp tejto metódy. [17]



Obr. 2.5: Zobrazenie na povrchu kvasiniek (YSD) je technika proteínového inžinierstva, ktorá využíva izoláciu rekombinantných proteínov zabudovaných do bunkovej steny kvasiniek na izoláciu a úpravu protilátok [13]. Proteín záujmu je zobrazený ako fúzia k proteínu Aga2p na povrchu kvasiniek. Na povrchu kvasinkového displeja sú proteíny fúzované s C-koncom epitopovej značky a N-koncom domény Aga2p, ktorá lokalizuje fúzny proteín na vonkajšom povrchu bunky [17]. Proteín Aga2p sa prirodzene používa v kvasinkách na sprostredkovanie kontaktov bunka-bunka počas párenia kvasiniek [13].



Obr. 2.6: Translokácia dvojarginínu (Tat). Proteín, ktorý je predmetom záujmu je fúzovaný s N-koncom peptidu ssTorA Tat (periplazmatický signálny exportér) a C-koncovým TEM-1 BLA s odstránenou sekvenciou Sec exportu. Baktéria *E. coli*, ktorá exprimuje fúzny proteín prežije v prítomnosti beta-laktámových antibiotík iba vtedy, ak je fúzny proteín prítomný v periplazme, pretože aktivita TEM-1 BLA závisí od tvorby disulfidových väzieb. Preto bunky produkujúce zložené rozpustné proteíny budú umožňovať rast na ampicínových platniach. [17]

Z článku [17] je známe, že mutácie, ktoré zvyšujú katalytickú aktivitu, sú v priemere destabilizujúce. Z 33 mutácií LGK, ktoré mali štatisticky signifikantné zvýšenie fitness, iba 7 malo skóre, ktoré ukazovalo zlepšenú rozpustnosť. Podobne tomu bolo u TEM-1.1. Z toho bola stanovená pravdepodobnosť, že daná mutácia zlepšuje stabilitu a zároveň fitness medzi 0,05 % a 0,15 %.

Wrenbeck et al. 2019

Okrem toho, že táto sada obsahuje dáta pre proteín LGK, takisto ako sada Klesmith et al. 2017, ďalším proteínom je pyrolidínketátsyntáza typ III (PKS) dlhá 392 aminokyselín. V čom sa táto dátová sada líši od predošlej, je spôsob selekcie, ktoré bolo vykonané hlbokým

mutačným testovaním s použitím metódy fúzie so zeleným fluorescenčným proteínom (GFP-fusion).

Metóda GFP-fusion používa ako expresný systém baktériu *E. coli*. Táto metóda funguje na princípe fluorescencie. Proteín, ktorý sa úspešne poskladá, povolí tvorbu chromoforov zeleného fluorescenčného proteínu, a teda je možné tento proteín odlíšiť pomocou fluorescenčnej intenzity. Ak sa proteín správne neposkladá, nie je fluorescenčný. [45]



Obr. 2.7: Fúzia k zelenému fluorescenčnému proteínu (GFP-fusion). Proteín, ktorý je predmetom záujmu, je geneticky kódovaný ako N-koncová fúzia ku zelenému fluorescenčnému proteínu, kde pri expresii môžu zložené proteíny umožniť skladanie a následné formovanie chromoforov zeleného fluorescenčného proteínu, zatiaľ čo nezložené proteíny budú nefluorescenčné. Počet poskladaných fúzných proteínov na bunku koreluje s rozpustnosťou/stabilitou proteínového variantu. [45]

OptSolMut dataset

OptSolMut dataset bol vytvorený ako trénovacia dátová sada zozbieraná z 15 rôznych štúdií určená pre trénovanie modelu strojového učenia. Obsahuje 137 jednobodových a viacbodových mutácií s údajmi o zmenách rozpustnosti. Snažili sa vybrať také mutácie, kde je mutant rozpustný pred mutáciou a aj po nej, a teda mení sa iba rozsah rozpustnosti.

Dátová sada obsahuje informácie o tom, či je mutant rozpustný pre 121 zo 137 záznamov (pre zvyšných 16 nebola informácia v literatúre dostupná). Z tých mutantov, ktoré mali túto informáciu dostupnú, iba 2 boli po mutácii uvedené ako nerozpustné, takže pre väčšinu mutantov to vedie buď ku zlepšeniu, alebo zhoršeniu rozpustnosti oproti divokému typu.

V článku uvádzajú informáciu, že pre 26 mutantov existujú všetky 4 prípady – zvýšenie alebo zníženie rozpustnosti a stability, takže zmeny v rozpustnosti a stabilite by mali byť v tejto dátovej sade nezávislé. [39]

PDB

Proteínová dátová banka (PDB) je globálny archív experimentálne stanovených trojdimenzionálnych štruktúrnych dát biologických makromolekúl. Dnes je všeobecne považovaný za hlavný zdroj údajov nevyhnutných na pochopenie funkčných úloh, ktoré makromolekuly zohrávajú v biológii a medicíne. Bola založená v roku 1971 ako prvý molekulárny dátový zdroj v biológii. Od roku 2003 ho riadi *Worldwide Protein Data Bank* (medzinárodné konzorcium), ktoré spoločne dohliada na ukladanie, validáciu a rozširovanie týchto otvorených dát. Hlavný archív obsahuje viac než 144 000 štruktúrnych modelov proteínov, DNA/RNA, ich komplexy s kovmi a malými molekulami a súvisiace experimentálne údaje a metaúdaje.

Každá PDB štruktúra obsahuje súradnice atómov, ktoré tvoria štruktúrally 3D model makromolekuly. Pozície týchto atómov sú špecifikované ako Karteziánske súradnice (x , y , z). Ďalšie informácie sú dostupné v článku [9].

2.5 Existujúce nástroje

V tejto podkapitole sa nachádza popis existujúcich nástrojov predikcie vplyvu mutácie na rozpustnosť proteínov.

Camsol

Nástroj Camsol sa skladá z viacerých častí. Jednou z nich je výpočet profilu rozpustnosti, ktorá dáva do vzťahu rozpustnosť a sklon k agregácii proteínu. V článku [36] sa uvádza, že napriek približnej zhode medzi rozpustnosťou a agregáciou, miera agregácie nepredstavuje celkové skóre rozpustnosti. Keďže rozpustnosť je termodynamický parameter, rôzne proteíny by mohli agregovať rôznymi rýchlosťami a pritom môžu mať stále rovnakú rozpustnosť.

Na výpočet profilu rozpustnosti sa na vstupe prijíma proteín v podobe sekvencie aminokyselín, ktorej na každú pozíciu reťazca priradí číslo, vyjadrujúce vplyv na celkovú rozpustnosť proteínu pomocou stratégie podobnej metóde Zygggregator, ktorá predikuje sklon k agregácii [38, 27]. Rozpustnosť pre pozíciu je priradená na základe lineárnej kombinácie fyzikálno-chemických vlastností aminokyselín, ako je hydrofóbnosť a elektrostatický náboj, vyhladenou kľavým oknom, ktoré berie do úvahy 7 susediacich aminokyselín. Skóre pre danú pozíciu menšie ako -1 indikuje, že má daná oblasť negatívny vplyv na rozpustnosť, naopak skóre väčšie ako 1 prispieva k lepšej rozpustnosti celého proteínu.

Úspešnosť celého nástroja bola testovaná na testovacej sade, ktorá obsahovala 56 proteínov rôznych druhov. Testovanie bolo obmedzené na zisťovanie, či mutácia zvyšuje alebo znižuje rozpustnosť. Nástroj správne predikoval zmenu rozpustnosti v 54 prípadoch z celkových 56, takže celková úspešnosť nástroja činí 0,96 %. Ďalšie informácie sú dostupné v článku [36].

ccSol Omics

Ďalším nástrojom je ccSol Omics, ktorý predikuje rozpustnosť na základe sekvenčných vlastností. Medzi tieto vlastnosti patrí skrútenie, hydrofóbnosť, hydrofilnosť a sklon k formovaniu α -skrutkovice a β -skladaného listu. Algoritmus prechádza sekvenciou od N-konca po jednej aminokyseline pomocou kľavého okna 21 aminokyselín až po C-koniec. Kvôli rôznym dĺžkam sekvencií bola použitá metóda založená na Fourierovej transformácii, ktorá umožňuje porovnávanie polypeptidových reťazcov rôznych dĺžok. Bolo použitých 100 Fourierových koeficientov, na ktorých bola natrénovaná neurónová sieť.

Trénovacia sada pochádzala z databázy TargetTrack (36990 záznamov) a pomocou 10-násobnej krížovej validácie dosiahol algoritmus presnosť 79 %. Na testovacej sade dosiahol tento algoritmus presnosť 74 % (31760 záznamov). Viac informácií o tomto nástroji v článku [2].

OptSolMut

Nástroj OptSolMut používa koncepty z výpočtovej geometrie pre definovanie skórovacej funkcie, ktorá predikuje zmenu rozpustnosti spôsobenou mutáciou na základe frekvencie tripletov aminokyselín, ktoré sú prevažne na povrchu proteínu. Táto skórovacia funkcia

zachytáva sekvenčné a aj štrukturálne informácie a je optimalizovaná pomocou metód lineárneho programovania (LP), ktorým spomedzi štandardných metód strojového učenia ako Support vector machines (SVM) a Lasso vyšli celkovo najlepšie výsledky. Viac informácií o tejto skorovacej funkcii v článku [40].

Na tréovanie a testovanie bola použitá dátová sada OptSolMut (2.4), ktorá obsahuje 137 jedno- a viacbodových mutácií. Úspešnosť bola vyhodnotená pomocou viacerých metód, jednou z nich je Leave-one-out cross-validation (LOOCV), ktorá rozdelila dátovú sadu na 137 častí (každá časť obsahuje 1 proteín) a v každej iterácii vynechá jednu časť na testovanie. Tento proces sa opakuje 137-krát a je výpočtovo náročnejší spomedzi ostatných metód krížovej validácie. Úspešnosť natrénovaného algoritmu na tejto dátovej sade stanovená pomocou LOOCV je 81 %. [40]

Soda

Posledným z prezentovaných nástrojov je Soda, ktorá určuje vplyv mutácie na rozpustnosť na základe piatich individuálnych komponentov. Je to sklon k agregácii (PASTA), skóre vnútornej poruchovosti proteínu (ESpritz) skombinované s hydrofóbnym profilom a sklon k formovaniu α -skrutkovice a β -skladaného listu (FESS). Nástroj okrem zadávania mutácií podporuje aj výpočet zmeny rozpustnosti pre vkladanie a mazanie časti sekvencie.

Tréovanie algoritmu prebehlo na dátovej sade PON-Sol [47]. Žiaľ, v článku sa nenachádza informácia o tréovaní algoritmu, sú známe iba výsledné váhy priradené k daným komponentom. Varianty v testovacej sade, ktoré nemali žiaden efekt na rozpustnosť spolu s nejednoznačnými, boli z tejto sady vyradené. Po odfiltrovaní obsahovala celkovo 142 proteínov. Na tejto sade dosiahol algoritmus pomocou 5-násobnej krížovej validácie úspešnosť 59 %. Vyhodnotenie úspešnosti bolo vykonané na testovacej dátovej sade Camsol [37], ktorá obsahuje 56 proteínov, kde všetky tieto proteíny majú párovú sekvenčnú identitu s tréovacou sadou menšiu ako 30 %, z čoho vyplýva, že si proteíny v sadách nie sú podobné. Na testovacej sade predikoval tento nástroj všetky položky úspešne – skóre 100 %. V článku sa ďalej uvádza, že tento nástroj dosahuje pozoruhodnú presnosť v predikcii zníženia rozpustnosti. [26]

Kapitola 3

Tvorba dátových sád

Táto kapitola obsahuje popis tvorby trénovacej a testovacej sady.

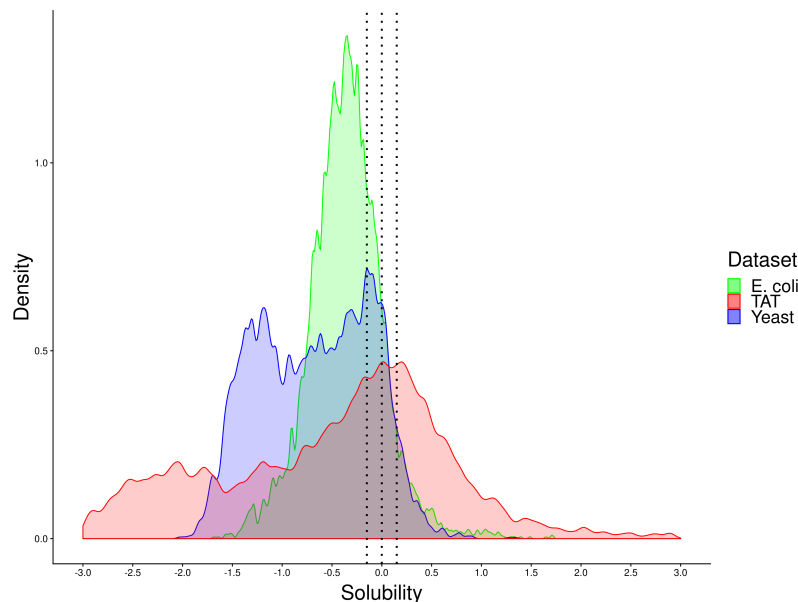
3.1 Trénovacia sada

V článkoch k dátovým sadám Klesmith et al. 2017 (2.4) a Wrenbeck et al. 2019 (2.4) je uvedené, že identifikovali prah rozpustnosti na hodnote 0,15 [45, 17], takže všetky mutácie, ktoré majú hodnotu rozpustnosti väčšiu ako tento prah, sú považované za zlepšujúce rozpustnosť. Tabuľka 3.1 sumarizuje tieto dátové sady po základnom predspracovaní.

Proteín	Dĺžka	Počet možných jednobodových mutácií.*	Bodové mutácie obsahujúce skóre rozpustnosti		
			YSD	Tat	GFP
TEM-1.1 (monomér) [17]	263 ak	4997	2467** (49%) 253 poz 2214 neg	3973 (80%) 770 poz 3203 neg	–
LGK (homodimér) [17, 45]	439 ak	8341	6298 (76%) 314 poz 5984 neg	6353 (76%) 1929 poz 4424 neg	6485 (78%) 424 poz 6061 neg
PKS (monomér) [45]	392 ak	7448	–	–	4840 (65%) 1420 poz 3420 neg

Tabuľka 3.1: Tabuľka proteínov, ktoré sú dostupné z mutačných dátových sád z článkov [17, 45]. *Bez mutácií na rované rezídium. **Obmedzené na Ambler pozície 61–215, aby sa predišlo YSD skresleniu [17]. Prahová hodnota na oddelenie pozitívnych a negatívnych vzoriek bola 0,15 [17, 45]. Za celkovým počtom vzoriek pre daný proteín a metódu expzie je zobrazený percentuálny podiel voči počtu možných jednobodových mutácií. Pod celkovým počtom sa nachádzajú počty vzoriek s detailom rozdelenia do dvoch skupín, ktoré je vyznačené za týmto počtom: poz – vzorka zlepšujúca rozpustnosť alebo neg – vzorka zhoršujúca rozpustnosť.

Tieto dátové sady, obsahujúce proteíny TEM-1.1, LGK a PKS, boli spojené na základe expresných systémov a vznikli nové sady – Yeast, Tat a E.coli, ktorých rozloženie rozpustnosti znázorňuje graf hustoty 3.1.



Obr. 3.1: Graf hustoty rozpustnosti pre systém kvasiniek (modrá), translokáciu dvojarginínu (červená) a *E. coli* (zelená). Vertikálne čiarkované čiary označujú hodnoty -0,15, 0 a 0,15 x-ovej osi.

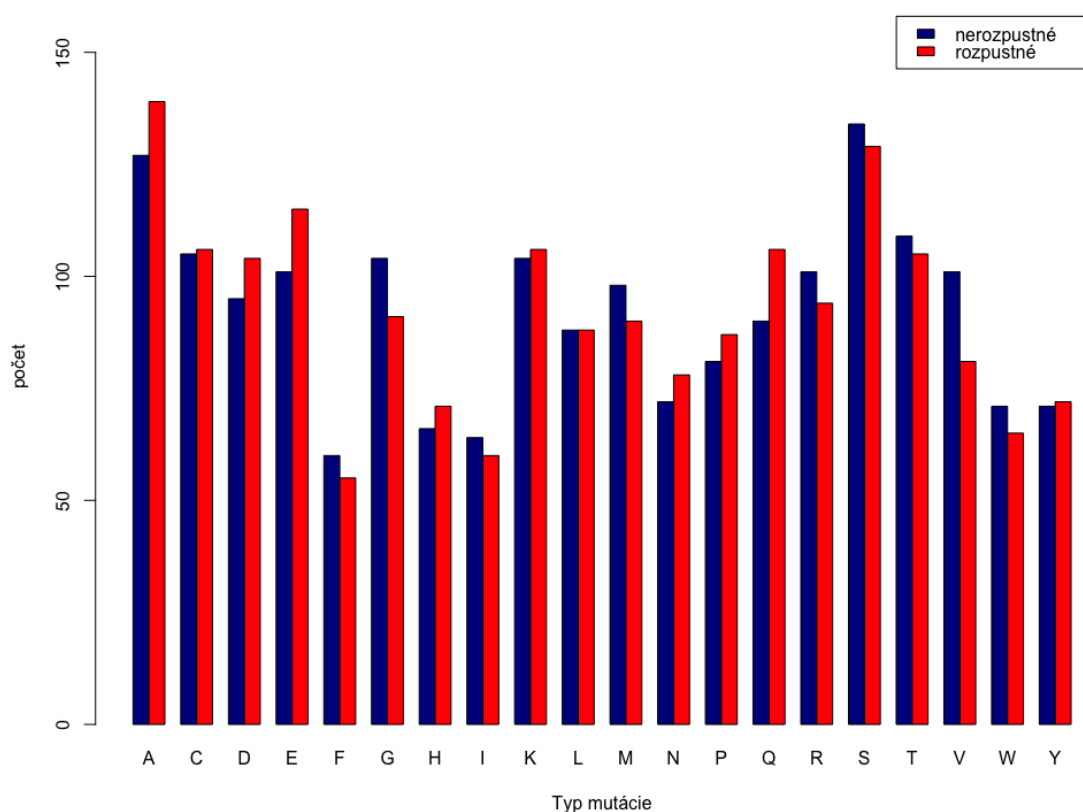
Tat sada obsahuje podozrivé množstvo mutácií, ktoré zvyšujú rozpustnosť, obzvlášť pre LGK proteín, navyše bola označená ako zastaraná od tvorcov tejto sady, takže sada nebude použitá pre účely tréningu. Dátové sady expresného systému Yeast majú najvyššiu koreláciu s vlastnosťou stability proteínu, predikovanou nástrojom FoldX – výberovým kritériom pravdepodobne dominuje vplyv tepelnej stability. Dátová sada Wrenbeck et al. 2019 (2.4) obsahuje najväčšie množstvo mutácií, keď spojíme sadu obsahujúcu proteín LGK a PKS, naviac vyzerá, že by nemala dominovať tepelná stabilita. Naviac, ostatné existujúce dátové sady dostupné z literatúry sú taktiež merané v expresnom systéme *E. coli*.

Vyváženie sady

Na dátovej sade Wrenbeck et al. 2019 (2.4) bude v tomto odseku vysvetlený problém nevyvážených sád. Táto sada obsahuje nevyvážený počet pozorovaných dát – 1844 pre mutácie zlepšujúce a 9481 pre zhoršujúce rozpustnosť. Takáto dátová sada je nevyvážená, pretože neobsahuje rovnaký počet dát pre všetky triedy. Metódy strojového učenia, ako napríklad Náhodné lesy (Random Forests), sú senzitívne na proporcie rôznych tried, a tak sa môže stať, že klasifikátor bude uprednostňovať triedu iba na základe väčšieho počtu pozorovaní triedy [33] – v tomto prípade mutácie zhoršujúce rozpustnosť. Je preto vhodné dátovú sadu vyvážiť, aby obsahovala rovnaký počet dát pre každú triedu, v tomto prípade dve triedy – mutácia zlepšuje rozpustnosť a mutácia zhoršuje rozpustnosť.

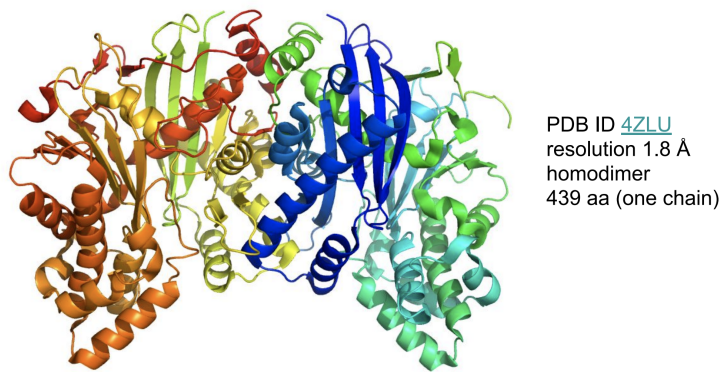
Ešte pred vyvažovaním sád boli vytvorené 4 podsady pre každú kombináciu proteínu a expresného systému (TEM-1.1 v YSD, LGK v YSD, LGK v *E. coli*, PKS v *E. coli*) spo-

jením sád. Keďže každá z nich má väčší počet nerozpustných mutácií, je z nerozpustných potrebné vybrať podmnožinu, aby bol počet dát v oboch triedach vyvážený. Tieto podmnožiny boli vyberané náhodne, berúc do úvahy distribúciu typu mutácie (typ mutácie predstavuje aminokyselinu, na ktorú sa mutuje). Najprv boli tieto sady vyvažované a nakoniec spojené na základe expresného systému (TEM-1.1 a LGK v YSD, LGK a PKS v *E. coli*). Z týchto dvoch sád bola pre účely tréningu vybraná Wrenbeck et al. 2019 (2.4), pretože *E. coli* je častejšie vyskytujúci sa expresný systém ako kvasinky. Túto sadu znázorňuje graf 3.2.

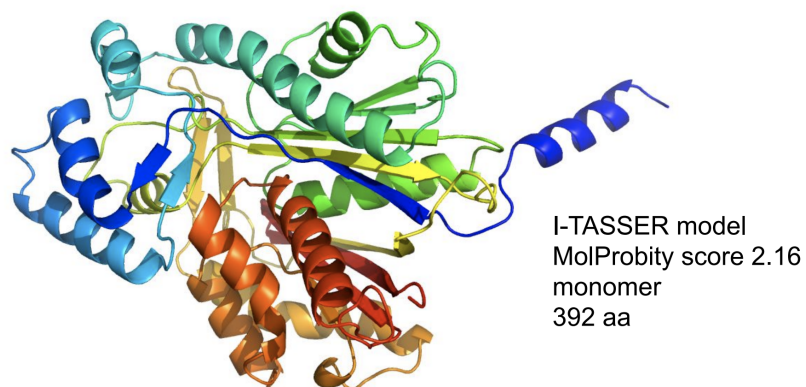


Obr. 3.2: Graf dátovej sady Wrenbeck et al. 2019 s počtom daných typov mutácií rozdelené do dvoch tried na základe prahu rozpustnosti 0,15 – rozpustné (zlepšená rozpustnosť) a zhoršená rozpustnosť (nerozpustné) oproti divokému typu.

Táto sada, ktorá bola vybraná pre účel tréningu modelu strojového učenia obsahuje spolu 3684 záznamov, ktoré sú pre proteíny LGK a PKS, ktorých 3D štruktúra je zobrazená na nasledujúcich obrázkoch.



Obr. 3.3: LGK konvertuje levoglukozán – produkt celulózovej biomasovej pyrolýzy. Používa sa pri výrobe biopalív odvodených od levoglukozánu. Štruktúra pre tento proteín bola prevzatá z PDB.



Obr. 3.4: PKS nachádza použitie v biosyntéze tropánových alkaloidov. Štruktúra tohto proteínu nie je dostupná v PDB, a preto bola predikovaná pomocou nástroja *I-TASSER* [46, 32, 49].

3.2 Testovacia sada

Testovacia sada vznikla spojením viacerých malých sád dostupných z literatúry. Patrí sem sada OptSolMut (2.4), Ponsol [47], Camsol [37] a A3d [48]. Z týchto sád boli vyradené proteíny, ktoré nemali príslušný polypeptidový reťazec v štruktúre PDB. Výsledná celková veľkosť dátovej sady je 181 mutácií, z toho 92 zlepšujúcich a 89 zhoršujúcich rozpustnosť.

Kapitola 4

Návrh nástroja

4.1 Výber predikčných vlastností

Táto podkapitola obsahuje predikčné vlastnosti, ktoré boli vybrané pre predikciu vplyvu mutácie na rozpustnosť proteínov. Hodnoty vypočítaných vlastností môžeme kategorizovať do dvoch skupín: hodnota vlastnosti pre divoký typ a hodnota spočítaná ako rozdiel medzi hodnotou mutanta a hodnotou divokého typu. Pri každej vlastnosti je táto informácia uvedená.

Fyzikálno-chemické

Do tejto skupiny patrí 13 fyzikálno-chemických vlastností vypočítaných na základe sekvencie aminokyselín. Výsledná hodnota je vždy rozdiel medzi hodnotou divokého typu a hodnotou mutanta.

Prvou vlastnosťou je *molekulová hmotnosť* proteínu. Ďalšia vlastnosť je K/R – pomer medzi počtom aminokyselín lyzínu a arginínu, ďalej D^*E , kde D je počet aminokyselín kyseliny asparágovej a E počet aminokyselín kyseliny glutámovej. Vzťah pre výpočet vlastnosti *Indexu nestability* proteínu pochádza z článku [14], ďalšia vlastnosť vychádzajúca z článku je *flexibilita* [42]. *GRAVY* (grand average of hydropathy) je definovaná súčtom hodnôt hydropatie pre všetky aminokyseliny v reťazci vydelená dĺžkou toho reťazca. *Izoelektrický bod* polypeptidov je daný pomocou metódy Bjellqista.

Všetky nasledujúce vlastnosti sú vypočítané ako súčet výskytov jednotlivých aminokyselín v reťazci. *Aromaticita* je daná súčtom výskytu aminokyselín tyrozínu, metionínu a fenylalanínu. *Frakcia nabitých aminokyselín* – počet aminokyselín, ktoré môžu byť pozitívne alebo negatívne nabité. Patrí sem kyselina asparágová, kyselina glutámová, lyzín a arginín. Pre *tepelnú stabilitu* to je izoleucín, valín, tyrozín, tryptofán, arginín, kyselina glutamová a leucín. *Frakcia skrutkovic* je daná súčtom výskytu aminokyselín valínu, izoleucínu, tyrozínu, fenylalanínu, tryptofánu a leucínu, ďalej pre *frakciu listov* to je kyselina glutámová, metionín, alanín a leucín. Poslednou vlastnosťou je *frakcia otočiek*, ktorá je daná súčtom výskytu asparagínínu, prolínu, glycínu a serínu.

AAindex1

Vlastnosti v skupine AAindex1 sa taktiež radia do skupiny fyzikálno-chemických. Je to databáza číselných indexov predstavujúcich rôzne fyzikálno-chemické vlastnosti aminokyselín. Obsahuje 566 vlastností, ktoré sú označené 10-miestnym kódom pozostávajúcim z písmen

a číslic. Každá z týchto vlastností obsahuje hodnoty pre 20 aminokyselín. Popis každej vlastnosti je možné nájsť na webovej stránke tejto databázy [15, 16]. Výsledná vlastnosť je hodnota rozdielu medzi hodnotou mutanta a divokého typu.

Disomine

Táto skupina vlastností sa spomedzi ostatných ako jediná počíta na vzdialenom webserveri, ktorý prijíma sekvenciu a z nej počíta štyri nasledujúce vlastnosti – disomine skóre (predikcia poruchy proteínu), dynamine (predikcia dynamiky kostry a postranných reťazcov proteínu) a efoldmine (predikcia zavčas zložených regiónov proteínu). [28]

Tango

Algoritmus Tango slúži pre predikciu oblastí náchylných na agregáciu z neposkladaného polypeptidového reťazca. Predpovedá β -krížovú agregáciu v peptidoch a denaturovaných proteínoch a pozostáva z fázového priestoru zahŕňajúceho náhodnú cievku a štyroch možných štruktúrnych stavov:

1. β -obrat
2. α -skrutkovica
3. agregácia β -listov
4. agregácia α -skrutkovic

Každý segment peptidu môže pozostávať z každého jedného z týchto stavov podľa Boltzmanovej distribúcie, t.j. frekvencia populácie každého štruktúrneho stavu pre daný segment bude relatívna k jeho energii. Z tohto dôvodu je β -krížová agregácia počítaná ako partičná funkcia fázového priestoru. [31, 12, 20]

Vážený index rozpustnosti (SWI)

SWI (Solubility-Weighted Index) je nový nástroj na predikciu proteínovej rozpustnosti. V článku [6] sa uvádza, že globálna štruktúrna flexibilita, ktorá môže byť modelovaná normalizovanými B-faktormi presne predikuje rozpustnosť 12216 rekombinantných proteínov v *E. coli*. Autori tohto nástroja optimalizovali B-faktory a odvodili hodnoty pre 20 aminokyselinových rezíduí, ktoré ďalej zlepšujú presnosť predikcie. Hodnota SWI pre danú mutáciu sa vyhodnotí ako súčet hodnôt, ktoré odvodili pre každé aminokyselinové rezíduum v sekvencii a vydeli celkovým počtom rezíduí v sekvencii. V prediktore SWI sa táto hodnota ďalej vloží do funkcie, ktorá vypočíta pravdepodobnosť rozpustnosti (logistická regresia). V nástroji uvedenom v tejto práci sa však ako výsledná hodnota používa rozdiel medzi výslednou hodnotou SWI divokého typu a mutanta – pravdepodobnosť sa už nepočíta.

Zmena stability proteínu ($\Delta\Delta G$)

Stabilita proteínu je definovaná ako rozdiel Gibbsovej voľnej energie (ΔG) medzi zbaleným a rozbaleným stavom proteínu [10]:

$$\Delta G = \Delta G_{\text{zbalený}} - \Delta G_{\text{rozbalený}} \quad (4.1)$$

kde G je Gibbsova voľná energia, ktorá je tvorená z nasledujúcich pojmov – entalpia (H), entropia (S) a teplota v Kelvinoch (T) [10], medzi ktorými je nasledovná relácia:

$$G = H - T \times S \quad (4.2)$$

Zmena v Gibbsovej voľnej energii rozbaľovania medzi variantou divokého typu a mutantom [23] je definovaná ako:

$$\Delta\Delta G = \Delta G_{divoký\ typ} - \Delta G_{mutant} \quad (4.3)$$

Táto hodnota sa počíta popri modelovaní štruktúry mutantu pomocou bioinformatického nástroja FoldX [35].

Povrchová plocha prístupná rozpúšťadlu (SASA)

Povrchová plocha prístupná rozpúšťadlu – SASA (Solvent accessible surface area) je ďalšou štrukturálnou vlastnosťou. V tomto prípade povrchová plocha proteínu prístupná rozpúšťadlu udáva mieru kontaktnej plochy medzi proteínom a rozpúšťadlom. Táto vlastnosť sa počíta pomocou nástroja *FreeSASA* [22] aproximačným algoritmom od autorov Lee a Richards [19]. Hodnoty týchto plôch sú v jednotkách $\text{\AA}^2$ ($\text{\AA}$ – Angström), čo je v prepočte $1.0 \times 10^{-20} \text{m}^2$. Pomocou parametrov tohto programu je vynútené, že na výstupe sú SASA hodnoty pre všetky úrovne štruktúry – vrátane atómov. Ďalej boli nastavené polomery atómov *ProtOr*, definované podľa *Tsai et al.* [41]. Počítajú sa absolútne a aj relatívne plochy. Relatívna prístupná plocha (RSA – relative accessible surface area) je vypočítaná nasledujúcim vzťahom:

$$RSA = \frac{ASA}{MaxASA} \quad (4.4)$$

kde ASA je vypočítaná plocha prístupná rozpúšťadlu a $MaxASA$ maximálna možná plocha prístupná rozpúšťadlu. Pre oba tieto typy plôch sú počítané vlastnosti pre každé rezíduum v proteíne: celková plocha, polárna plocha, nepolárna plocha, plocha hlavného reťazca a plocha bočného reťazca. Každá táto vlastnosť je počítaná pre rezíduum divokého typu na lokácií, kde sa má mutácia uskutočniť a to isté pre rezíduum mutantu po mutácii.

Sklon k priestorovej agregácii (SAP)

Ďalšou štrukturálnou vlastnosťou je sklon k priestorovej agregácii – SAP (spatial aggregation propensity), ktorá je tiež počítaná pomocou nástroja *FreeSASA*. Vzťah výpočtu hodnoty SAP pre atóm i :

$$SAP_{\text{atóm } i} = \sum_{\text{Priemer simulácie}} \left\{ \sum_{\substack{\text{Rezíduá s aspoň} \\ \text{jedným atómom} \\ \text{bočného reťazca} \\ \text{v okolí polomeru R} \\ \text{od atómu } i}} \left(\frac{\text{SAA atómov} \\ \text{bočného reťazca} \\ \text{v oblasti polomeru R}}{\text{SAA atómov} \\ \text{bočného reťazca} \\ \text{úplne exponovaného rezidua}} \times \text{Hydrofóbnosť rezidua} \right) \right\} \quad (4.5)$$

- SAA (solvent accesible area) je oblasť prístupná rozpúšťadlu

Táto skupina vlastností obsahuje celkovo štyri vlastnosti pre jednu mutáciu. Je to súčet hodnoty SAP pre všetky atómy v štruktúre divokého typu, rozdiel týchto súčtov medzi divokým typom a mutantom, hodnota SAP pre rezíduum divokého typu na lokácii, ktorá sa má mutovať a posledná je hodnota SAP pre rezíduum mutantu na lokácii, kde prebehla mutácia. Táto sada štyroch vlastností sa počíta pre dva polomery – 5 a 10. Polomer stanovuje, aké veľké okolie sa má prehľadávať pri hľadaní susedných atómov/rezíduí/reťazov/modelov alebo štruktúr. Ďalšie informácie dostupné o tejto metóde v článku [8].

Priestorový AAIndex1

V priestorovom AAIndex1 sa oproti sekvenčnému AAIndex1 (4.1) berie do úvahy okolie rezídua, pre ktoré sa hodnota počíta. Hodnota priestorového AAIndexu sa vypočíta nasledujúcim spôsobom: v okolí rezídua na lokácii danej mutácie sa nájdu susedné rezíduá, ktoré sa nachádzajú v okolí polomeru 7. Hodnota výslednej vlastnosti je spočítaná ako suma vážených hodnôt z databázy AAIndex1 rezíduí v okolí, kde váha v tejto váženej hodnote predstavuje počet rovnakých rezíduí v okolí vydelený celkovým počtom rezíduí v okolí, inak povedané to je zastúpenosť rezídua v okolí. Vzťah výpočtu hodnoty priestorového AAIndex1 (SAF) rezídua na lokácii mutácie je daná nasledujúcim vzťahom:

$$\text{SAF}_{\text{rezíduum } i} = \sum_{\substack{\text{Rezíduá v okolí} \\ \text{polomeru } R \\ \text{od rezídua } i}} \text{Zastúpenosť rezídua v okolí} \times \text{Hodnota AAIndex1 rezídua} \quad (4.6)$$

Oproti sekvenčnej vlastnosti AAIndex1 sa okrem hodnoty rozdielu medzi mutantom a divokým typom ukladá aj hodnota divokého typu, takže celkový počet hodnôt tejto skupiny vlastností oproti sekvenčnej AAIndex1 je dvakrát väčší, čo je vo výsledku 1132 hodnôt pre jednu mutáciu. Práve táto vlastnosť má najväčší počet hodnôt.

Kontaktné číslo

Kontaktné číslo (Contact number) je definované ako počet Cbeta atómov rezíduí v okruhu Cbeta atómu pre ktorý sa toto kontaktné číslo počíta. Táto vlastnosť sa počíta pre dva rôzne okruhy tohto Cbeta atómu: polomer 7 a 12. Celkovo táto skupina vlastností obsahuje iba dve hodnoty: hodnotu kontaktného čísla rezídua v proteíne divokého typu na lokácii, ktorá sa má mutovať a rozdiel hodnoty kontaktného čísla rezídua po mutácii na danej lokácii s touto hodnotou.

4.2 Návrh nástroja

Nástroj SoluProtMut slúži na predikciu zmeny v rozpustnosti proteínov po mutácií. Je založený na binárnom rozhodovaní, takže klasifikuje zmenu do dvoch tried – zlepšenie alebo zhoršenie rozpustnosti. Na vstup sa tomuto nástroju vkladá zoznam mutácií a štruktúra proteínu vo formáte PDB. Na výstupe sa objaví informácia o každej mutácii, či zlepšuje alebo zhoršuje rozpustnosť.

Výpočet vlastností

Nástroj počítá veľké množstvo vlastností. Nie všetky vlastnosti sa však používajú pre trénovanie modelu. Vyberajú sa iba tie najrelevantnejšie pomocou po sebe idúcich fáz (hodnoty t a n sú jednoducho nastaviteľné.):

1. regularizácia (Lasso)
2. odstránenie vysoko korelovaných vlastností, ktoré majú koreláciu vyššiu ako prah t
3. výber top n vlastností

Vybrané vlastnosti sa ukladajú do súboru, aby bolo možné po skončení trénovania vybrať také isté vlastnosti, ako sa použili pre trénovanie kvôli ďalším predikciám nových mutácií. V prediktore môže nastať situácia, že pre niektorú mutáciu nie je spočítaná každá vlastnosť – chýba hodnota. Tento problém je vyriešený pomocou ukladania priemernej hodnoty z trénovacích dát, takže všade, kde hodnota nejakej vlastnosti chýba, nastaví sa na priemernú hodnotu z trénovacej sady.

Súborová cache

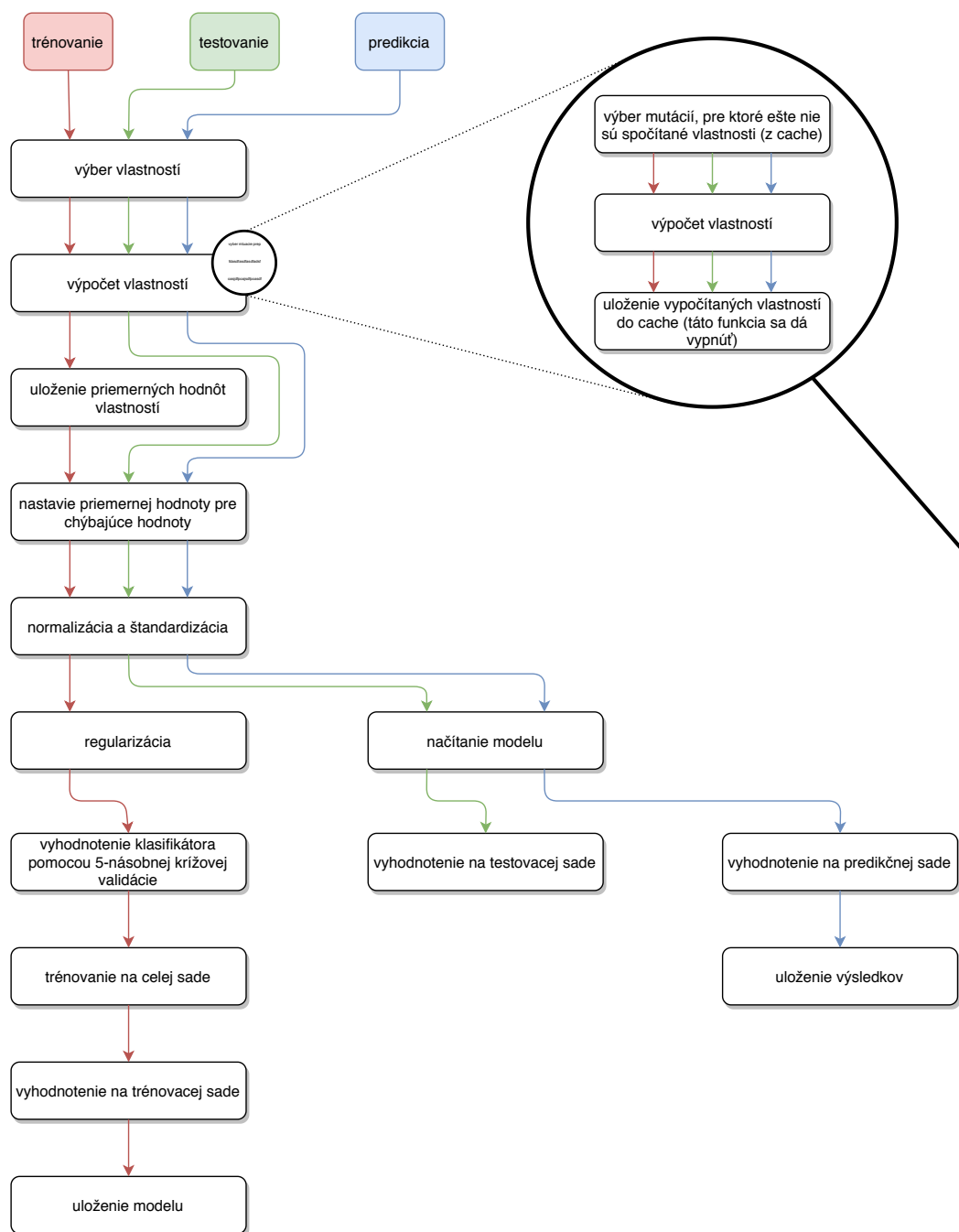
Prediktor obsahuje jednoducho implementovanú súborovú cache, pretože niektoré vlastnosti sú náročné na výpočet, pri každom novom spustení by ich bolo potrebné počítat, čo by trvalo značný čas. Prediktor tak vždy pri štarte dopočítava iba chýbajúce vlastnosti, ktoré sa nenachádzajú v cache. Vypočítané vlastnosti sú ukladané do súborov a v prípade neočakávanej chyby alebo vynúteného vypnutia programu sa už vypočítané vlastnosti nestratia, ale uložia. Táto funkcia cacheovania sa dá vypnúť, avšak iba pre zapisovanie do nej, čítanie je vždy zapnuté. Záznamy sú unikátne a riadia sa podľa kľúča formátu:

`{pdb}_{chain}_{mutation}_{location}`

kde *pdb* predstavuje id štruktúry, *chain* reťaz v tejto štruktúre, *location* pozícia, na ktorej má prebehnúť mutácia a *mutation* znamená, na akú aminokyselinu sa mutuje. Pre odstránenie spočítaných vlastností nachádzajúcich sa v cache stačí vymazať celý adresár tejto cache – prediktor pri ďalšom štarte vytvorí novú.

Postupnosť krokov nástroja

Postupnosť krokov nástroja pre trénovanie, testovanie a predikciu je zobrazený na nasledujúcom grafe:



Obr. 4.1: Diagram postupnosti krokov nástroja SoluProtMut pre tri rôzne módy – tréno-
vanie, testovanie a predikciu

Kapitola 5

Implementácia nástroja

V tejto kapitole je opísaná implementácia prediktoru vplyvu mutácie na rozpustnosť proteínov. Prediktor má názor SoluProtMut.

5.1 Použité frameworky a knižnice

Prediktor rozpustnosti bol implementovaný v jazyku Python. Pre algoritmy strojového učenia bola použitá knižnica **sklearn**. Manipulácia s dátami bola vykonaná pomocou knižnice **pandas** a **numpy**. Štatistické výpočty boli realizované pomocou knižnice **scipy**. Keďže prediktor pracuje s biologickými dátami, vo veľkom rozsahu sa používa knižnica **biopython**. V nástroji sa hlavne používajú funkcie na spracovanie sekvencií, operácie nad nimi a pre načítavanie štruktúr proteínov. Na vývoj bolo použité prostredie Anaconda s Pythonom 3 s doinštalovaním vyššie spomenutých knižníc, ktoré v tomto prostredí chýbali.

5.2 Implementácia

Vstupný bod programu predstavuje skript `predict.py`. Hlavná trieda prediktoru je **Predictor** a modeluje prediktor strojového učenia. Trieda **Model** je model strojového učenia, ktorý je parameter triedy **Predictor**. Predvoleným modelom sú *Náhodné lesy*, konkrétne implementácia **RandomForestClassifier** z knižnice **sklearn**. Každý natrénovaný model sa po natrénovaní ukladá pod názvom modelu, aby ho bolo možné v budúcnosti použiť. Operácie nad dátovými sadami sú vykonávané v triede **Dataset**. Medzi tieto operácie patrí výpočet vlastností nad týmto datasetom, normalizácia a štandardizácia dát a štatistické funkcie. Zdrojové kódy výpočtov vlastností, na základe ktorých sa predikuje rozpustnosť, sa nachádzajú v zložke **features**. Bázová trieda pre vlastnosti je **Feature**. Slúži hlavne pre zapúzdrenie výpočtov daných vlastností a ich následné ukladanie. Používa sa priamo pri vlastnostiach, ktoré používajú na výpočet iba sekvenciu. Vlastnosti, ktoré sa počítajú nejakým spôsobom zo štruktúry, používajú triedu **StructuralFeature**, ktorá rozširuje triedu **Feature** o prácu so štruktúrami. V skripte `config.py` sa nachádza konfigurácia prediktora. V tejto konfigurácii je možné nastaviť operačný systém, pracovný adresár prediktora, v ktorom si uchováva behové informácie, cesty k súborovej cache a ďalšie nastavenia.

Popis tried

V tejto podkapitole sa nachádza bližší popis tried (pre stručnosť budú metódy značené bez parametrov).

Predictor je ústredný bod celého nástroja. Hlavnými atribútmi tejto triedy sú dátové sady (objekty triedy **Dataset**), binárny rozhodovací prah a model strojového učenia (objekt triedy **Model**). Nachádza sa tu metóda, ktorá na začiatku predikcie vyberá požadované vlastnosti `init_features_for_dataset()`, ktorá vytvorí objekty pre vlastnosti explicitne vymenovaných v tejto funkcii, takže ak sa odstráni volanie metódy na vytvorenie objektov danej vlastnosti, táto vlastnosť sa nebude počítať a teda ani používať na predikciu. Táto trieda ďalej obsahuje metódy `train()`, `test()` a `predict()`, v ktorých sú definované kroky pre jednotlivé módy prediktoru – regularizácia spolu s tréновaním a krížovou validáciou modelu, testovanie na testovacej dátovej sade a predikcia na nových dátach. Výber módu je možný pomocou argumentov prediktoru. Trieda obsahuje metódu pre ladenie hyper-parametrov, ktorá sa však nepoužíva, no je vhodná pri experimentovaní s rôznymi algoritmami strojového učenia.

Model je abstrakcia nad modelom strojového učenia. Ako atribúty obsahuje názov modelu, mriežka parametrov a model strojového učenia, ktorý sa inštanciuje v konštruktore tejto triedy. Obsahuje metódy pre vykonanie jedného kroku krížovej validácie, akcie spojené s tréновaním, testovaním a predikciou, ako napríklad tréновanie, vyhodnotenie presnosti či ukladanie/načítanie modelov zo súboru.

Dataset zapúzdruje dátové sady. Načítava dátové sady a vykonáva všetky operácie nad nimi. Tieto sady sú reprezentované ako dátové rámce knižnice **pandas**, ktoré sú načítané zo súboru formátu *csv*. Trieda obsahuje nasledujúce atribúty: načítaná dátová sada, hodnoty vypočítaných vlastností, vybrané vlastnosti určené regularizáciou (proces regularizácie sa nachádza v triede **Predictor**). Trieda je multifunkčná, takže trénovacia, testovacia a predikčná sada je vytvorená pomocou rovnakej triedy, aj keď sú si sady rozdielne. Napríklad trénovacia sada obsahuje pozorovanú rozpustnosť, pričom predikčná sada rozpustnosť neobsahuje. Ďalej táto trieda obsahuje metódy pre štandardizáciu a normalizáciu dátových sád. Koeficienty zistené na trénovacej sade sa ukladajú do súboru, aby sa na testovacích a predikčných sádach dáta transformovali rovnakými koeficientami. Tak isto sa ukladá aj hodnota priemerných hodnôt vlastností na trénovacej sade, aby sa pri chýbajúcich hodnotách vlastností mohla táto priemerná hodnota použiť.

Feature je bazová trieda pre výpočet vlastností. Načítava vypočítané vlastnosti zo súborovej cache, ale ak sa tam vypočítaná hodnota vlastnosti pre danú mutáciu nenachádza, vlastnosť vypočíta. Obsahuje nasledovné atribúty: vypočítané vlastnosti načítané zo súboru, cestu k vypočítaným vlastnostiam v cache, cestu k štruktúram proteínov, informáciu o operačnom systéme (kvôli externým programom), informáciu o type štruktúry (originálna/opravená). Ak ešte súbor s hodnotami vypočítaných vlastností neexistuje, tak ho vytvorí (názov vlastností získa pomocou abstraktnej metódy `get_fnames()` definovanej v odvodených triedach). Táto trieda obsahuje abstraktnú metódu `set_wt()`, ktorú implementujú odvodené triedy (konkrétne vlastnosti) a väčšinou slúži na načítanie štruktúry proteínu divokého typu a následné získanie sekvencie. Druhou dôležitou abstraktnou metódou je `add()`, ktorá slúži na výpočet vlastností pre danú mutáciu. Táto metóda je volaná v cykle iterujúceho cez mutácie, ktoré nemajú spočítané vlastnosti. Medzi triedy vlastností, ktoré dedia

túto triedu a implementujú príslušný výpočet vlastností patria *Pc*, *PcBasic*, *PcAAindex1*, *Swi*, *Tango* a *DisoMine*.

StructuralFeature je trieda odvodená od triedy **Feature** a rozširuje túto triedu o zložitejšiu prácu so štruktúrami, ktorá je potrebná pre výpočet štrukturálnych vlastností. Obsahuje abstraktné triedy **set_wt_structural** a **add_structural**, ktoré sú štrukturálnymi variantami metód **set_wt** a **add()** v básovej triede. Táto odvodená trieda ďalej nahrádza metódu **add()**, v ktorej nepostačí vykonať mutáciu na nejakej lokácii v sekvencii, ako je to pri sekvenciálnych vlastnostiach, ale je potrebné pracovať so štruktúrou mutanta. Táto štruktúra je modelovaná nástrojom FoldX. Medzi triedy vlastností, ktoré dedia túto triedu a implementujú príslušný výpočet vlastností patria *AAindex1Structural*, *ContactNumber*, *FoldxDDG*, *SAP* a *SASA*.

Práca s proteínovými štruktúrami

Štrukturálne vlastnosti sa počítajú zo štruktúry PDB. Podľa manuálu nástroja FoldX je vhodné tieto štruktúry pred samotnou prácou s nimi opraviť. Tento nástroj umožňuje opravu štruktúr pomocou príkazu *RepairPDB*, ktorý identifikuje a opraví rezíduá, ktoré majú chybné krútiacie uhly, VanderWalsove strety alebo celkovú energiu. Táto oprava sa v prediktore uskutoční vždy, keď sa daná vlastnosť počíta zo štruktúry. Výslednú štruktúru uloží do rovnakého adresára ako je pôvodná štruktúra s pridaním prípony *_Repair* (napríklad *4zlu_Repair.pdb*). V adresári so štruktúrami existuje ešte ďalší adresár pre štruktúry mutantov, ktoré boli modelované z opravených štruktúr divokého typu. Toto modelovanie je vykonané pri počítaní vlastností zmeny stability proteínu 4.1, príkazom *BuildModel* nástroja FoldX. Výsledný model mutanta sa používa pri výpočtoch štrukturálnych vlastností. Viac informácií o nástroji FoldX v literatúre [35].

Výpočet vlastností externými nástrojmi

Niektoré vlastnosti sa počítajú pomocou samostatného programu. Medzi tieto programy patrí FoldX, FreeSASA a Tango. FoldX počíta vlastnosť ddG. FreeSASA sa používa pri výpočtoch SAP, a SASA. Posledný nástroj Tango slúži na výpočet vlastností s rovnakým názvom. Výpočty z týchto programov sa ukladajú do pracovného adresára prediktora do samostatných zložiek s názvom nástroja. Pomocou premennej *OS* v súbore *config.py* je možné nastaviť operačný systém, podľa ktorého sa vyberá program, ktorý sa má spustiť. Je možné nastaviť dve hodnoty: "*linux*" a "*mac*". Tieto programy boli skompilované na Ubuntu 18.04.2 LTS (Bionic Beaver) a macOS 10.14.6 (Mojave). Podporované sú 64-bitové verzie operačných systémov.

Príklad použitia

Pred spustením prediktora je potrebné umiestniť všetky štruktúry proteínov, ktoré sa mutujú do zložky *pdb/*. Štruktúra sa priradí príslušnej mutácii na základe názvu *pdb*, takže názov súboru štruktúry musí byť rovnaký ako v stĺpci vstupného súboru prediktora, napríklad *4zlu.pdb* pre nasledujúci súbor *vstup.csv*:

```
pdb,chain,location,mutation
4zlu,A,10,P
4zlu,A,12,F
3upc,A,26,D
```

Príklad spustenia prediktoru SoluProtMut v móde predikcie nových mutácií:

```
python predict.py --predict --predict_dataset_path vstup.csv
--o_csv vysledky.csv
```

Výstupný súbor `vysledky.csv` bude vyzerat nasledovne:

```
pdb,chain,location,mutation,score,effect
4zlu,A,10,P,0.49293,decrease
4zlu,A,12,F,0.45082,decrease
3upc,A,26,D,0.56882,increase
```

kde `score` je výsledné skóre, ktoré priradil prediktor danej mutácii a `effect` indikuje, či mutácia zlepšila alebo zhoršila rozpustnosť.

Pomocou argumentov tohto nástroja je možné špecifikovať nasledovné správanie programu (nápoveda dostupná prepínačom `--help`):

- spustiť tréovanie, testovanie a predikciu (všetko možné v jednom behu)
- nastaviť dátové sady pre tréovanie, testovanie a predikciu
- vypnúť zapisovanie do cache jednotlivu pre tréovanie, testovanie a predikciu
- názov výstupného súboru z predikcie
- vybrať model strojového učenia, ktorý sa má použiť

Kapitola 6

Experimenty

V tejto kapitole je prezentovaných niekoľko experimentov rôznych metód strojového učenia.

6.1 Výber hyper-parametrov

Pred každým experimentom bol vykonaný výber hyper-parametrov daného modelu pomocou hrubého mriežkového vyhľadávania parametrov estimátoru. Skórovacia funkcia tohto vyhľadávania bola $F1$, známa aj ako vyvážené F-skóre. Skóre je dané nasledujúcim výrazom:

$$F_1 = 2 * \frac{\text{presnosť} * \text{citlivosť}}{\text{presnosť} + \text{citlivosť}} \quad (6.1)$$

Skóre $F1$ môže byť interpretované ako vážený priemer presnosti a citlivosti, kde $F1$ skóre dosiahne najlepšiu hodnotu v jednotke a najhoršie skóre v nule. Relatívny príspevok presnosti a citlivosti ku skóre $F1$ je rovnaký. [7]

6.2 Regularizácia

Vo všetkých experimentoch bol rovnaký výber vlastností, a to pomocou regularizácie metódou *Lasso* (Least absolute shrinkage and selection operator). Lasso je lineárny model, ktorý odhaduje riedke koeficienty. Je užitočný pre výber vlastností, pretože má tendenciu uprednostňovať riešenia s menším počtom nenulových koeficientov, čím sa účinne znižuje počet vlastností, od ktorých je dané riešenie závislé. Matematicky pozostáva z lineárneho modelu s pridaným termínom regularizácie. Minimalizová je nasledovná objektívna funkcia:

$$\min_w \frac{1}{2n_{\text{vzoriek}}} \|Xw - y\|_2^2 + \alpha \|w\|_1 \quad (6.2)$$

Lasso odhad teda rieši minimalizáciu penalizácie najmenších štvorcov s pridaním $\alpha \|w\|_1$, kde α je konštanta a $\|w\|_1$ je ℓ_1 -metrika koeficientového vektora [7].

Po regularizácii boli ďalej odstránené vlastnosti, ktoré boli vysoko korelované – mali koreláciu vyššiu než 0,75. Na konci ostalo 74 vlastností, ktoré boli použité pre tréning modelov a sú uvedené v tabuľke 6.1.

vlastnosť	koeficient	vlastnosť	koeficient
flexibility	0.1912	aai1_7_KHAG800101_wt	0.0147
sap_5_residue_d	0.0744	aai1_7_WILM950104_wt	0.0142
foldx_ddg	0.0674	aai1_7_GEOR030103	0.0131
fracnumcharge	0.0635	aai1_7_WILM950104	0.0128
sap_5_global_wt	0.0603	sap_5_residue_wt	0.0122
aai1_7_OOBM850101_wt	0.0548	aai1_7_AURR980110	0.0119
contact_number_12_wt	0.0522	aai1_WILM950104	0.0116
aai1_LEWP710101	0.0511	aai1_NAKH900107	0.0108
tango_aggr_res_d	0.0481	aai1_7_AURR980118	0.0103
aai1_7_TANS770108	0.048	aai1_7_VASM830101_wt	0.0101
sap_10_residue_wt	0.0476	aai1_7_QIAN880137_wt	0.0096
aai1_7_QIAN880139_wt	0.0432	aai1_RICJ880116	0.0096
aai1_PONP800105	0.0422	aai1_7_CHAM830102_wt	0.0094
aai1_7_RACS820107_wt	0.0416	sasa_rel_mainchain_d	0.0093
disomine_res_wt	0.0415	sasa_rel_total_d	0.0087
aai1_CHOP780209	0.0369	aromaticity	0.0086
aai1_7_TANS770107_wt	0.0344	aai1_AVBF000107	0.0085
sasa_rel_mainchain_wt	0.0326	aai1_RACS820103	0.0081
kr_ratio	0.0314	aai1_7_SUEM840102	0.0075
isoelectric_point	0.0304	aai1_BHAR880101	0.0075
aai1_7_SUEM840102_wt	0.0298	contact_number_12_d	0.0056
aai1_7_ROSM880103	0.0293	aai1_OOBM850104	0.0053
aai1_SUYM030101	0.0259	aai1_7_DAYM780201	0.0047
aai1_7_DAYM780201_wt	0.0257	aai1_7_LAWE840101_wt	0.0031
aai1_7_RACS820103_wt	0.0249	aai1_VASM830101	0.0031
aai1_7_LEWP710101_wt	0.0242	aai1_PONP800106	0.003
aai1_7_ROBB790101	0.0225	disomine_sidechain_res_wt	0.0024
aai1_ROSM880104	0.0219	aai1_7_PALJ810107	0.0023
instability_index	0.0189	aai1_7_RADA880105	0.0019
aai1_7_AURR980116	0.018	aai1_7_SNEP660104	0.0013
contact_number_7_d	0.0167	aai1_7_NAKH920105	0.0013
aai1_7_WERD780103	0.0158	aai1_7_WERD780102_wt	0.0013
aai1_7_TANS770102_wt	0.0158	aai1_7_RICJ880116	0.0008
aai1_RACS820112	0.0157	aai1_7_TANS770102	0.0003
aai1_PRAM820101	0.0154	aai1_7_OOBM850102_wt	0.0002
sasa_abs_mainchain_d	0.0153	sasa_rel_apolar_d	0.0002
tango_aggr_res_wt	0.0149	aai1_7_OOBM850103	0.0001

Tabuľka 6.1: Vlastnosti vybrané pomocou Lasso a následne odstránené vlastnosti s vysokou koreláciou s inou vlastnosťou. V tabuľke sa nachádza názov vlastnosti a koeficient po Lasso regularizácii. Záznamy sú zoradené od najvyššej hodnoty koeficientu. aai – AAI-index, vlastnosť_R... – číslo R je polomer pre výpočet štruktúrálnej vlastnosti. Na konci názvu vlastnosti môžu byť prípony d alebo wt. d – delta (rozdiel hodnoty vlastnosti medzi mutantom a divokým typom), wt – wild-type (hodnota vlastnosti pre divoký typ).

6.3 Náhodné lesy

V tomto experimente bol použitý model Náhodných lesov, čo je kombinovaná učiaca sa metóda pre klasifikáciu a regresiu. V tomto prípade bola použitá klasifikácia, pretože má výstup iba dva stavy – zlepšená alebo zhoršená rozpustnosť. Klasifikátor bol použitý z knižnice **sklearn**. Pomocou vyhľadávania hyper-parametrov boli odhadnuté nasledujúce parametre modelu:

- **n_estimators** (počet stromov v lese): 500
- **criterion** (funkcia na meranie kvality rozdelenia): entropy
- **class_weight** (váhy spojené s triedami): balanced
- **bootstrap** (či sa pri vytváraní stromov používajú bootstrapové vzorky): áno
- **min_weight_fraction_leaf** (minimálna váhovaná frakcia celkovej sumy váh aby mohol byť uzol listovým): 0,01
- **min_impurity_decrease** (uzol sa rozdelí, ak toto rozdelenie spôsobí zníženie nečistoty väčšie alebo rovné ako nastavená hodnota): 0,0005

Klasifikátor bol na trénovacej sade vyhodnotený pomocou 5-násobnej krížovej validácie, kde dosiahol vyváženú presnosť 68,4 %. Dosiahnutá vyvážená presnosť modelu natrénovaného a vyhodnoteného na celej trénovacej sade bola 84,8 % a na testovacej 71,5 %. Tabuľka 6.2 ukazuje dôležitosť vlastností po natrénovaní modelu.

6.4 Support Vector Machines

Ďalším modelom, ktorý bol vyskúšaný je Support Vector Machines (SVM). Pomocou vyhľadávania hyper-parametrov boli odhadnuté nasledujúce parametre modelu:

- **kernel** (typ kernelu použitý v algoritme): radial basis function
- **C** (regularizačný parameter): 100
- **gamma** (kernelový koeficient): 0.001

Vyhodnotenie prebehlo pri nastavení, kde bolo vybratých top 30 vlastností spomedzi 74. Klasifikátor bol na trénovacej sade vyhodnotený pomocou 5-násobnej krížovej validácie, kde dosiahol vyváženú presnosť 65,6 %. Dosiahnutá vyvážená presnosť modelu natrénovaného a vyhodnoteného na celej trénovacej sade bola 78,4 % a na testovacej 64,5 %. V prípade, že pre predikciu bolo vybraných všetkých 74 vlastností, významne sa líšila vyvážená presnosť na testovacej sade, ktorá bola iba 46,96 %.

6.5 Vyhodnotenie

Vyhodnotenie nového prediktoru SoluProtMut prebehlo na testovacej sade vytvorenej v tejto práci. Ako model bol použitý ten s najlepšou presnosťou z experimentov - Náhodné lesy.

vlastnosť	dôležitosť	vlastnosť	dôležitosť
disomine_res_wt	0.046957	aail_7_NAKH920105	0.010719
foldx_ddg_ddg	0.034545	aail_RICJ880116	0.010513
aail_7_RACS820107_wt	0.031225	aail_PRAM820101	0.010384
aail_7_RACS820103_wt	0.028578	aail_ROSM880104	0.010379
sap_10_residue_wt	0.027358	aail_PONP800106	0.010306
tango_aggr_res_d	0.026982	aail_WILM950104	0.010221
tango_aggr_res_wt	0.025016	aail_SUYM030101	0.009903
aail_7_CHAM830102_wt	0.023795	aail_7_SNEP660104	0.009707
aail_7_QIAN880139_wt	0.022881	aail_7_WERD780103	0.009701
sap_5_residue_wt	0.022771	aail_7_DAYM780201	0.009654
aail_7_KHAG800101_wt	0.021204	aail_OOBM850104	0.009560
aail_7_OOBM850102_wt	0.021046	aail_PONP800105	0.009016
aail_7_WILM950104_wt	0.020531	aail_BHAR880101	0.009008
aail_7_TANS770107_wt	0.019697	aail_NAKH900107	0.008801
contact_number_12_wt	0.019422	aail_7_GEOR030103	0.008661
flexibility	0.018428	aail_7_AURR980116	0.008623
aail_7_VASM830101_wt	0.018219	aail_7_PALJ810107	0.008609
aail_7_DAYM780201_wt	0.017604	aail_RACS820112	0.008538
aail_7_OOBM850101_wt	0.017145	aail_RACS820103	0.008500
aail_7_SUEM840102_wt	0.016508	aail_7_OOBM850103	0.008453
aail_7_QIAN880137_wt	0.015934	aail_7_SUEM840102	0.008410
sap_5_residue_d	0.015186	aail_7_TANS770102	0.008309
aail_7_LEWP710101_wt	0.015066	aail_VASM830101	0.008211
aail_7_TANS770102_wt	0.015050	sasa_rel_apolar_d	0.008081
aail_CHOP780209	0.014097	aail_7_RICJ880116	0.007981
aail_7_WERD780102_wt	0.014090	aail_LEWP710101	0.007806
disomine_sidechain_res_wt	0.013620	aail_7_RADA880105	0.007655
sasa_rel_total_d	0.013612	aail_7_AURR980118	0.007618
aail_7_LAWE840101_wt	0.013129	fracnumcharge	0.007598
aail_7_ROBB790101	0.013069	instability_index	0.007408
aail_7_TANS770108	0.013021	sasa_rel_mainchain_d	0.006848
sap_5_global_wt	0.013003	isoelectric_point	0.006238
sasa_rel_mainchain_wt	0.012802	sasa_abs_mainchain_d	0.005630
aail_7_AURR980110	0.012491	aromaticity	0.002428
aail_7_ROSM880103	0.012114	contact_number_12_d	0.000697
aail_AVBF000107	0.011528	contact_number_7_d	0.000639
aail_7_WILM950104	0.010899	kr_ratio	0.000565

Tabuľka 6.2: Tabuľka dôležitosti vlastností zoradená od najdôležitejšej zostupne podľa určeného algoritmom Náhodných lesov po natrénovaní na tréningovej sade. aai – AAIndex, vlastnosť R... – číslo R je polomer pre výpočet štruktúrálnej vlastnosti. Na konci názvu vlastnosti môžu byť prípony d alebo wt. d – delta (rozdiel hodnoty vlastnosti medzi mutantom a divokým typom), wt – wild-type (hodnota vlastnosti pre divoký typ).

Nástroj	Presnosť	Vyvážená presnosť	TP	FP	TN	FN	Prah
OptSolMut	0.8729	0.8728	81	12	77	11	-3.8075
SoluProtMut	0.7569	0.7534	89	41	48	3	0.4963
Soda	0.7017	0.6985	82	44	45	10	-0.2338
Camsol	0.663	0.6606	74	43	46	18	-0.0001
ccSol Omics	0.5635	0.5703	15	2	87	77	0.365

Tabuľka 6.3: Tabuľka zámeny (Confusion matrix) nástrojov vyhodnotených na testovacej sade. TP – skutočne pozitívny (true positive), FP – falošne pozitívny (false positive), TN – skutočne negatívny (true negative), FN - falošne negatívny (false negative). Prah – najlepšia rozdeľovacia hodnota vzhľadom na presnosť nástroja.

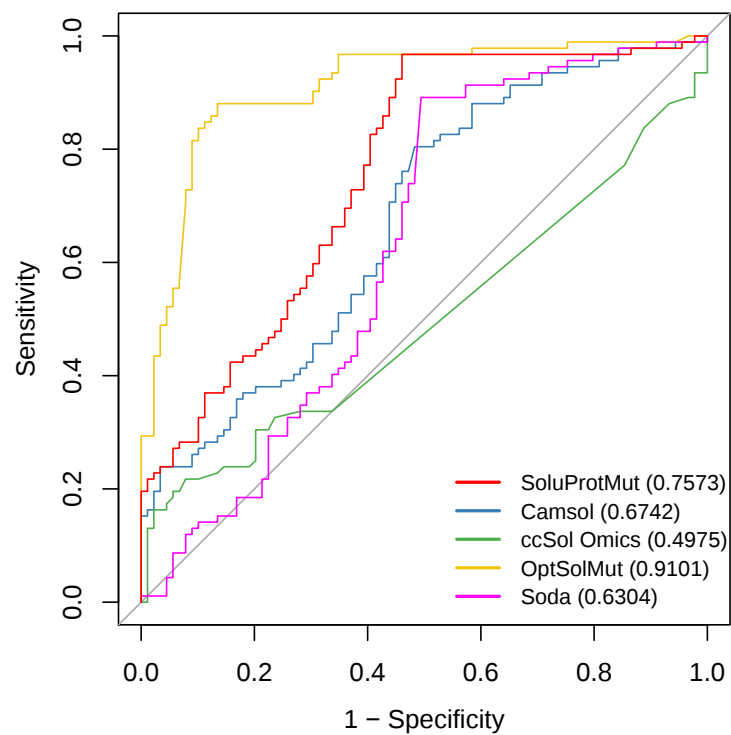
Spolu s týmto prediktorom boli vyhodnotené ďalšie prediktory: Camsol, ccSol Omics, OptSolMut a Soda. Všetky nástroje uvádzajú, na akých tréningových dátových sadách boli trénované, takže je možné určiť prekryv s testovacou sadou, na ktorej prebehlo vyhodnotenie. Matica zámeny (Confusion matrix) pre tieto nástroje je v tabuľke 6.3.

ROC (Receiver operating characteristic) krivka je nástroj pre hodnotenie a optimalizáciu binárneho klasifikačného systému, ktorý ukazuje vzťah medzi špecifitou a senzitivitou daného prediktora pre všetky prípustné hodnoty prahu. Senzitivita a špecifita je definovaná nasledovne:

$$\text{Senzitivita} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (6.3)$$

$$\text{Špecifita} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (6.4)$$

Na grafe sa bude najlepší prediktor vyznačovať ROC krivkou s najväčšou plochou pod touto krivkou (Area Under the Curve - AUC). Ak je plocha rovná hodnote 1, prediktor je ideálny. Ak má plocha pod krivkou hodnotu 0,5, potom prediktor nie je lepší než hádzanie mince [44]. Nasledujúce ROC krivky porovnávajú existujúce nástroje a nový prediktor SoluProtMut na testovacej sade. Najlepšie presnosť dosiahol nástroj OptSolMut s vyváženou presnosťou 87,28% – s veľkou pravdepodobnosťou je to kvôli prieniku dát so sadou, na ktorej bol natrénovaný (viac ako polovica testovacej sady, ktorá bola použitá na vyhodnotenie, bola použitá na tréning nástroja OptSolMut). Hneď druhým nástrojom je SoluProtMut prezentovaný v tejto práci s vyváženou presnosťou 75,34%, pričom treba podotknúť, že testovacia sada nemala žiaden prienik s tréningovou. Ako tretí skončil nástroj Soda s vyváženou presnosťou 69,85%, ktorý má prekryv v 34 mutáciách so sadou, na ktorej bol natrénovaný. Nečakane nízku presnosť vykázal nástroj ccSol Omics, ktorý mal vyváženú presnosť iba 57,03%. Autori tohto nástroja však uvádzajú presnosť 74% na ich testovacej sade.



Obr. 6.1: ROC krivky existujúcich predikčných nástrojov a prediktora SoluProtMut. (Legenda obsahuje názov nástroja a v zátvorke je plocha pod danou krivkou – AUC)

Kapitola 7

Záver

Úvod tejto práce bol zameraný na štruktúru, rozpustnosť a mutácie proteínov z pohľadu biológie, ktorá je nevyhnutná na pochopenie problematiky vplyvu mutácie na rozpustnosť proteínov. Po biologickom úvode nasledovala časť venovaná existujúcim dátovým sadám Klesmith et al. 2017, Wrenbeck et al. 2019, OptSolMut a Proteínová dátová banka PDB. Na konci kapitoly boli popísané existujúce nástroje predikcie vplyvu mutácie na rozpustnosť proteínov, spolu s vyhodnotením úspešnosti od ich tvorcov.

Ďalšia kapitola sa venuje tvorbe dátových sád. Ukázalo sa, že najvhodnejším kandidátom na trénovaciu sadu je sada z článku Wrenbeck et al. 2019, a to hneď z niekoľkých dôvodov: obsahuje najväčšie množstvo mutácií pri spojení dát pre proteíny LGK a PKS. Ďalej vyzerá, že by nemala dominovať tepelná stabilita a väčšina ostatných dostupných dátových sád zaoberajúcich sa vplyvom mutácií taktiež pochádza z expresného systému baktérie *E. coli*. Trénovaciu sadu bolo potrebné vyvážiť, pretože v nej prevažovali mutácie ktoré zhoršujú rozpustnosť, čo pri niektorých algoritmoch strojového učenia, ako napríklad Náhodné lesy, môže spôsobiť, že bude uprednostňovaná trieda na základe väčšieho počtu pozorovaní. Výsledná trénovacia sada obsahuje 3684 záznamov. Pre proteín PKS však nie je dostupná štruktúra v proteínovej dátovej banke. Keďže novovytvorený nástroj potrebuje pre výpočet vlastností štruktúru, pre proteín PKS bola štruktúra modelovaná pomocou nástroja *I-TASSER*. Testovacia sada vznikla spojením štyroch dostupných malých sád z literatúry: Optsolmut, Ponsol, Camsol a A3d. Obsahuje 181 záznamov, z čoho 92 zlepšuje a 89 zhoršuje rozpustnosť. Dalo by sa povedať, že je skoro vyvážená.

V ďalšej časti práce sú popísané predikčné vlastnosti. Delia sa na dve hlavné kategórie: sekvenčné a štrukturálne. Sekvenčné vlastnosti sa počítajú výhradne zo sekvencie aminokyselinových rezíduí, kým štrukturálne sú počítané z trojdimenzionálnej štruktúry proteínu. Medzi sekvenčné vlastnosti patria fyzikálno-chemické, ktorých výpočet prebieha priamo v prediktore a už predpočítané, ktoré pochádzajú z databázy AAIndex1. Ďalšia skupina sekvenčných vlastností nesie názov Disomine, čo je vzdialený server, ktorý počíta poruchovosť proteínu, dynamiku kostry a postranných reťazcov a zavčas zložené regióny. Ďalší je algoritmus Tango, ktorý predikuje oblasti náchylné na agregáciu zo sekvencie aminokyselinových rezíduí. Posledná zo sekvenčných vlastností je vážený index rozpustnosti. Prvá štrukturálna vlastnosť je zmena stability proteínu, ktorú predstavuje zmena v Gibbsovej voľnej energii rozbaľovania medzi variantom divokého typu a mutanta. Vlastnosti povrchová plocha prístupná rozpúšťadlu a sklon k priestorovej agregácii sú počítané pomocou nástroja FreeSASA. Poslednými vlastnosťami počítaných zo štruktúry je priestorový AAIndex1, ktorý oproti sekvenčnému indexu berie do úvahy okolie rezídua a kontaktné číslo, ktoré je definované ako počet C β atómov rezíduí v okolí.

Nástroj počíta veľké množstvo vlastností, avšak nie všetky sú potrebné pre predikciu. V časti návrhu nástroja je opísaný výber vlastností. Ďalej táto podkapitola obsahuje popis súborovej cache, ktorá slúži pre ukladanie vypočítaných vlastností, pretože niektoré sú náročné na výpočet. Na konci podkapitoly je zobrazený diagram postupnosti krokov nástroja pre tri rôzne módy: tréovanie, testovanie a predikcia. Kapitola Implementácia nástroja obsahuje detailný popis implementácie na úrovni zdrojového kódu. Ďalej obsahuje popis práce s proteínovými štruktúrami a príklad použitia tohto prediktora.

Ešte pred tréovaním bola vykonaná regularizácia pomocou Lasso. Ďalej boli odstránené vysoko korelované vlastnosti, a tak spomedzi všetkých ostalo na konci celkovo 74 vlastností. Na týchto vlastnostiach dosiahol najlepšiu úspešnosť algoritmus Náhodných lesov, kde pomocou 5-násobnej krížovej validácie na tréovacej sade dosiahol vyváženú presnosť 68,4% a na testovacej sade 71,5%. Tento model je vďaka jeho presnosti spomedzi ostatných aj predvoleným modelom tohto prediktora. Prediktor ďalej obsahuje natréovaný model pomocou SVM, ktorý však dosiahol horšie výsledky.

Vyhodnotenie nástroja prebehlo na testovacej sade, pričom okrem novovytvoreného nástroja v rámci tejto práce je porovnaný s ostatnými nástrojmi, ktoré sú opísané v úvodnej kapitole. Najlepšiu vyváženú presnosť (ďalej len ako presnosť) na testovacej sade dosiahol nástroj OptSolMut s hodnotou 87,28%, ale s veľkou pravdepodobnosťou je to kvôli prieniku dát so sadou, na ktorej bol natréovaný (viac ako polovica sady, na ktorej bol vyhodnotený, bola použitá na tréovanie tohto nástroja). Hneď druhý za ním sa umiestnil nástroj prezentovaný v tejto práci s presnosťou 75,34%, pričom treba podotknúť, že testovacia sada nemala žiaden prekryv s tréovacou, takže tento nástroj bol vyhodnotený na komplet nevidených dáta. Tretí sa umiestnil nástroj Soda s presnosťou 69,85%, u ktorého je taktiež prítomný prekryv s tréovacou sadou – konkrétne 34 mutácii. Nečakane nízku presnosť vykázal nástroj ccSol Omics, ktorý mal presnosť len 57,03%. Autori tohto nástroja však uvádzajú presnosť 74% na ich testovacej sade. V podkapitole Vyhodnotenie sa nachádza porovnanie týchto nástrojov – matica zámen (confusion matrix) a ROC krivky (Receiver operating characteristics) pre tieto nástroje.

Prediktor SoluProtMut oproti nástrojom Soda a ccSol Omics, ktoré predikujú zmenu rozpustnosti iba zo sekvencie aminokyselín, berie do úvahy aj priestorové usporiadanie proteínu. Vyzerá to, že vlastnosti spočítané z priestorovej štruktúry zlepšujú presnosť prediktora. Z najlepších 30 štruktúrálnych vlastností podľa dôležitosti algoritmu Náhodných lesov je iba päť počítaných zo samotnej sekvencie. Nástroj s lepšími výsledkami ako nástroj prezentovaný v tejto práci na testovacej sade skončil prediktor OptSolMut, avšak výsledok je značne skreslený, pretože viac ako polovica jeho tréovacej sady bola v prieniku so sadou na ktorej bol vyhodnotený. Nástroj OptSolMut by preto bolo vhodné vyhodnotiť na inej dátovej sade. Okrem tohto nástroja dosiahol prediktor SoluProtMut prezentovaný v tejto práci lepšiu presnosť ako ostatné nástroje. Celková presnosť však nie je úplne postačujúca a problém predikcie vplyvu mutácie na rozpustnosť proteínov tak ostáva nevyriešený. Nástroj však môže poskytnúť cenný odhad, ktoré mutácie uprednostniť pred ostatnými, a tým ušetriť čas a finančné náklady.

Literatúra

- [1] Tertiary structure. In: *IUPAC Compendium of Chemical Terminology*. IUPAC. DOI: 10.1351/goldbook.t06282. Dostupné z: <https://doi.org/10.1351/goldbook.t06282>.
- [2] AGOSTINI, F., CIRILLO, D., LIVI, C. M., PONTI, R. D. a TARTAGLIA, G. G. ccSOL omics: a webserver for solubility prediction of endogenous and heterologous expression in *Escherichia coli*. *Bioinformatics*. Oxford University Press (OUP). júl 2014, zv. 30, č. 20, s. 2975–2977. DOI: 10.1093/bioinformatics/btu420. Dostupné z: <https://doi.org/10.1093/bioinformatics/btu420>.
- [3] ALBERTS, B., BRAY, D., HOPKIN, K., JOHNSON, A., LEWIS, J. et al. *Essential cell biology, Fourth edition*. Garland Science, 2014. ISBN 978-0-8153-4454-4.
- [4] BAGBY, S., TONG, K., IKURA, M., THOMAS, V., JAMES, L. et al. *Methods in Enzymology*. Academic Press New York:, 2001.
- [5] BERG, J. M., TYMOCZKO, J. L. a STRYER, L. *Biochemistry (5th edition)*. New York : W.H. Freeman, 2002. ISBN 0-7167-3051-0.
- [6] BHANDARI, B. K., GARDNER, P. P. a LIM, C. S. Solubility-Weighted Index: fast and accurate prediction of protein solubility. Cold Spring Harbor Laboratory. február 2020. DOI: 10.1101/2020.02.15.951012. Dostupné z: <https://doi.org/10.1101/2020.02.15.951012>.
- [7] BUITINCK, L., LOUPPE, G., BLONDEL, M., PEDREGOSA, F., MUELLER, A. et al. API design for machine learning software: experiences from the scikit-learn project. In: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. 2013, s. 108–122.
- [8] CHENNAMSETTY, N., VOYNOV, V., KAYSER, V., HELK, B. a TROUT, B. L. Design of therapeutic proteins with enhanced stability. *Proceedings of the National Academy of Sciences*. Proceedings of the National Academy of Sciences. júl 2009, zv. 106, č. 29, s. 11937–11942. DOI: 10.1073/pnas.0904191106. Dostupné z: <https://doi.org/10.1073/pnas.0904191106>.
- [9] CONSORTIUM wwPDB. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Research*. Október 2018, zv. 47, D1, s. D520–D528, [cit. 2020-04-22]. DOI: 10.1093/nar/gky949. ISSN 0305-1048.
- [10] DAY, A. *Definition of Stability* [online]. [cit. 2020-05-02]. Dostupné z: <http://www.cryst.bbk.ac.uk/PPS2/projects/day/TDayDiss/StabilityDefined.html>.

- [11] DOBROTA, D. *Lekárska biochémia*. Vydavateľstvo Osveta, 2016. ISBN 978-80-8063-444-5.
- [12] FERNANDEZ ESCAMILLA, A.-M., ROUSSEAU, F., SCHYMKOWITZ, J. a SERRANO, L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature Biotechnology*. Springer Science and Business Media LLC. september 2004, zv. 22, č. 10, s. 1302–1306. DOI: 10.1038/nbt1012. Dostupné z: <https://doi.org/10.1038/nbt1012>.
- [13] GAI, S. A. a WITTRUP, K. D. Yeast surface display for protein engineering and characterization. *Current Opinion in Structural Biology*. Elsevier BV. august 2007, zv. 17, č. 4, s. 467–473. DOI: 10.1016/j.sbi.2007.08.012. Dostupné z: <https://doi.org/10.1016/j.sbi.2007.08.012>.
- [14] GURUPRASAD, K., REDDY, B. a PANDIT, M. W. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Engineering, Design and Selection*. Oxford University Press (OUP). 1990, zv. 4, č. 2, s. 155–161. DOI: 10.1093/protein/4.2.155. Dostupné z: <https://doi.org/10.1093/protein/4.2.155>.
- [15] KAWASHIMA, S., OGATA, H. a KANEHISA, M. AAindex: Amino Acid Index Database. *Nucleic Acids Research*. Oxford University Press (OUP). január 1999, zv. 27, č. 1, s. 368–369. DOI: 10.1093/nar/27.1.368. Dostupné z: <https://doi.org/10.1093/nar/27.1.368>.
- [16] KAWASHIMA, S., POKAROWSKI, P., POKAROWSKA, M., KOLINSKI, A., KATAYAMA, T. et al. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Research*. Oxford University Press (OUP). december 2007, zv. 36, Database, s. D202–D205. DOI: 10.1093/nar/gkm998. Dostupné z: <https://doi.org/10.1093/nar/gkm998>.
- [17] KLESMITH, J. R., BACIK, J.-P., WRENBECK, E. E., MICHALCZYK, R. a WHITEHEAD, T. A. Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. *Proceedings of the National Academy of Sciences*. National Academy of Sciences. 2017, zv. 114, č. 9, s. 2265–2270. DOI: 10.1073/pnas.1614437114. ISSN 0027-8424. Dostupné z: <https://www.pnas.org/content/114/9/2265>.
- [18] KRAMER, R. M., SHENDE, v. R., MOTL, N., PACE, C. N. a SHOLTZ, J. M. Toward a Molecular Understanding of Protein Solubility: Increased Negative Surface Charge Correlates with Increased Solubility. *Biophys J*. 2012, [cit. 2020-04-20]. DOI: 10.1016/j.bpj.2012.01.060. PMID: 22768947.
- [19] LEE, B. a RICHARDS, F. The interpretation of protein structures: Estimation of static accessibility. *Journal of Molecular Biology*. Elsevier BV. február 1971, zv. 55, č. 3, s. 379–IN4. DOI: 10.1016/0022-2836(71)90324-x. Dostupné z: [https://doi.org/10.1016/0022-2836\(71\)90324-x](https://doi.org/10.1016/0022-2836(71)90324-x).
- [20] LINDING, R., SCHYMKOWITZ, J., ROUSSEAU, F., DIELLA, F. a SERRANO, L. A Comparative Study of the Relationship Between Protein Structure and Beta-Aggregation in Globular and Intrinsically Disordered Proteins. *Journal of*

- Molecular Biology*. Elsevier BV. september 2004, zv. 342, č. 1, s. 345–353. DOI: 10.1016/j.jmb.2004.06.088. Dostupné z: <https://doi.org/10.1016/j.jmb.2004.06.088>.
- [21] MARUŠIAK, M. *Predikce rozpustnosti proteinů*. Brno, CZ, 2018. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Dostupné z: https://www.vutbr.cz/www_base/zav_prace_soubor_verejne.php?file_id=181116.
- [22] MITTERNACHT, S. FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Research*. F1000 Research Ltd. február 2016, zv. 5, s. 189. DOI: 10.12688/f1000research.7931.1. Dostupné z: <https://doi.org/10.12688/f1000research.7931.1>.
- [23] MONTANUCCI, L., CAPRIOTTI, E., FRANK, Y., BEN TAL, N. a FARISELLI, P. DDGun: an untrained method for the prediction of protein stability changes upon single and multiple point variations. *BMC Bioinformatics*. Springer Science and Business Media LLC. júl 2019, zv. 20, S14. DOI: 10.1186/s12859-019-2923-1. Dostupné z: <https://doi.org/10.1186/s12859-019-2923-1>.
- [24] MURRAY, R. K. *Harper's Biochemistry*. Lange Medical Books, 1993. ISBN 978-0838535622.
- [25] NELSON, D. L. *Lehninger Principles of Biochemistry (6th Edition)*. Macmillan, 2012. ISBN 978-1-4641-0962-1.
- [26] PALADIN, L., PIOVESAN, D. a TOSATTO, S. C. E. SODA: prediction of protein solubility from disorder and aggregation propensity. *Nucleic Acids Research*. Oxford University Press (OUP). máj 2017, zv. 45, W1, s. W236–W240. DOI: 10.1093/nar/gkx412. Dostupné z: <https://doi.org/10.1093/nar/gkx412>.
- [27] PAWAR, A. P., DUBAY, K. F., ZURDO, J., CHITI, F., VENDRUSCOLO, M. et al. Prediction of “Aggregation-prone” and “Aggregation-susceptible” Regions in Proteins Associated with Neurodegenerative Diseases. *Journal of Molecular Biology*. 2005, zv. 350, č. 2, s. 379 – 392. DOI: <https://doi.org/10.1016/j.jmb.2005.04.016>. ISSN 0022-2836. Dostupné z: <http://www.sciencedirect.com/science/article/pii/S0022283605004262>.
- [28] RAIMONDI, D. *DisoMine: a webserver for the prediction of disordered regions in proteins with recurrent Neural Networks and protein dynamics* [online]. [cit. 2020-05-02]. Dostupné z: <http://bio2byte.com/disomine/>.
- [29] REVA, B., ANTIPIN, Y. a SANDER, C. Predicting the Functional Impact of Protein Mutations: Application to Cancer Genomics. *Nucleic acids research*. Júl 2011, zv. 39, s. e118. DOI: 10.1093/nar/gkr407.
- [30] RIÈS KAUTT, M. a DUCRUIX, A. [3] Inferences drawn from physicochemical studies of crystallogenes and precrystalline state. In: *Macromolecular Crystallography Part A*. Academic Press, 1997, sv. 276, s. 23 – 59. Methods in Enzymology. DOI: [https://doi.org/10.1016/S0076-6879\(97\)76049-X](https://doi.org/10.1016/S0076-6879(97)76049-X). ISSN 0076-6879. Dostupné z: <http://www.sciencedirect.com/science/article/pii/S007668799776049X>.
- [31] ROUSSEAU, F., SCHYMKOWITZ, J. a SERRANO, L. Protein aggregation and amyloidosis: confusion of the kinds? *Current Opinion in Structural Biology*. Elsevier

- BV. február 2006, zv. 16, č. 1, s. 118–126. DOI: 10.1016/j.sbi.2006.01.011. Dostupné z: <https://doi.org/10.1016/j.sbi.2006.01.011>.
- [32] ROY, A., KUCUKURAL, A. a ZHANG, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols*. Springer Science and Business Media LLC. marec 2010, zv. 5, č. 4, s. 725–738. DOI: 10.1038/nprot.2010.5. Dostupné z: <https://doi.org/10.1038/nprot.2010.5>.
- [33] SANTACRUZ, A. *Why it is important to work with a balanced classification dataset* [online]. [cit. 2020-04-28]. Dostupné z: <http://amsantac.co/blog/en/2016/09/20/balanced-image-classification-r.html>.
- [34] SCHEIN, C. H. Solubility and secretability. *Current opinion in biotechnology*. Elsevier. 1993, zv. 4, č. 4, s. 456–461.
- [35] SCHYMKOWITZ, J., BORG, J., STRICHER, F., NYS, R., ROUSSEAU, F. et al. The FoldX web server: an online force field. *Nucleic Acids Research*. Oxford University Press (OUP). júl 2005, zv. 33, Web Server, s. W382–W388. DOI: 10.1093/nar/gki387. Dostupné z: <https://doi.org/10.1093/nar/gki387>.
- [36] SORMANNI, P., APRILE, F. A. a VENDRUSCOLO, M. The CamSol Method of Rational Design of Protein Mutants with Enhanced Solubility. *Journal of Molecular Biology*. 2015, zv. 427, č. 2, s. 478 – 490. DOI: <https://doi.org/10.1016/j.jmb.2014.09.026>. ISSN 0022-2836. Dostupné z: <http://www.sciencedirect.com/science/article/pii/S0022283614005312>.
- [37] SORMANNI, P., APRILE, F. A. a VENDRUSCOLO, M. The CamSol Method of Rational Design of Protein Mutants with Enhanced Solubility. *Journal of Molecular Biology*. Elsevier BV. január 2015, zv. 427, č. 2, s. 478–490. DOI: 10.1016/j.jmb.2014.09.026. Dostupné z: <https://doi.org/10.1016/j.jmb.2014.09.026>.
- [38] TARTAGLIA, G. G., PAWAR, A. P., CAMPIONI, S., DOBSON, C. M., CHITI, F. et al. Prediction of Aggregation-Prone Regions in Structured Proteins. *Journal of Molecular Biology*. 2008, zv. 380, č. 2, s. 425 – 436. DOI: <https://doi.org/10.1016/j.jmb.2008.05.013>. ISSN 0022-2836. Dostupné z: <http://www.sciencedirect.com/science/article/pii/S0022283608005676>.
- [39] TIAN, Y., DEUTSCH, C. a KRISHNAMOORTHY, B. Scoring function to predict solubility mutagenesis. *Algorithms for molecular biology : AMB*. Október 2010, zv. 5, s. 33. DOI: 10.1186/1748-7188-5-33.
- [40] TIAN, Y., DEUTSCH, C. a KRISHNAMOORTHY, B. Scoring function to predict solubility mutagenesis. *Algorithms for Molecular Biology*. Springer Science and Business Media LLC. 2010, zv. 5, č. 1, s. 33. DOI: 10.1186/1748-7188-5-33. Dostupné z: <https://doi.org/10.1186/1748-7188-5-33>.
- [41] TSAI, J., TAYLOR, R., CHOTHIA, C. a GERSTEIN, M. The packing density in proteins: standard radii and volumes 1 Edited by J. M. Thornton. *Journal of Molecular Biology*. Elsevier BV. júl 1999, zv. 290, č. 1, s. 253–266. DOI: 10.1006/jmbi.1999.2829. Dostupné z: <https://doi.org/10.1006/jmbi.1999.2829>.

- [42] VIHINEN, M., TORKKILA, E. a RIIKONEN, P. Accuracy of protein flexibility predictions. *Proteins: Structure, Function, and Genetics*. Wiley. jún 1994, zv. 19, č. 2, s. 141–149. DOI: 10.1002/prot.340190207. Dostupné z: <https://doi.org/10.1002/prot.340190207>.
- [43] VOLKIN, D. a MIDDAGH, C. Stability of protein pharmaceuticals. *Chemical and Physical Pathways of Protein Degradation*. Plenum Press, New York, NY, USA. 1992, s. 109–134.
- [44] WIKISKRIPTA. *ROC křivka* [online]. [cit. 2020-05-02]. Dostupné z: https://www.wikiskripta.eu/w/ROC_k%C5%99ivka.
- [45] WRENBECK, E. E., BEDEWITZ, M. A., KLESMITH, J. R., NOSHIN, S., BARRY, C. S. et al. An Automated Data-Driven Pipeline for Improving Heterologous Enzyme Expression. *ACS Synthetic Biology*. 2019, zv. 8, č. 3, s. 474–481. DOI: 10.1021/acssynbio.8b00486. PMID: 30721031. Dostupné z: <https://doi.org/10.1021/acssynbio.8b00486>.
- [46] YANG, J., YAN, R., ROY, A., XU, D., POISSON, J. et al. The I-TASSER Suite: protein structure and function prediction. *Nature Methods*. Springer Science and Business Media LLC. december 2014, zv. 12, č. 1, s. 7–8. DOI: 10.1038/nmeth.3213. Dostupné z: <https://doi.org/10.1038/nmeth.3213>.
- [47] YANG, Y., NIROULA, A., SHEN, B. a VIHINEN, M. PON-Sol: prediction of effects of amino acid substitutions on protein solubility. *Bioinformatics*. Oxford University Press (OUP). február 2016, zv. 32, č. 13, s. 2032–2034. DOI: 10.1093/bioinformatics/btw066. Dostupné z: <https://doi.org/10.1093/bioinformatics/btw066>.
- [48] ZAMBRANO, R., JAMROZ, M., SZCZASIUK, A., PUJOLS, J., KMIECIK, S. et al. AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures. *Nucleic Acids Research*. Oxford University Press (OUP). apríl 2015, zv. 43, W1, s. W306–W313. DOI: 10.1093/nar/gkv359. Dostupné z: <https://doi.org/10.1093/nar/gkv359>.
- [49] ZHANG, Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*. Springer Science and Business Media LLC. január 2008, zv. 9, č. 1. DOI: 10.1186/1471-2105-9-40. Dostupné z: <https://doi.org/10.1186/1471-2105-9-40>.

Príloha A

Obsah priloženého DVD

thesis/ text bakalárskej práce vo formáte PDF a zdrojový kód L^AT_EX

soluprotmut/

aaindex/ databáza AAIndex1

datasets/ trénovacia, testovacia a predikčná dátová sada

features/ zdrojové kódy výpočtov vlastností

pdbs/ štruktúry proteínov vo formáte PDB

tools/ externé nástroje FoldX, FreeSASA, Tango

utils/ sekvečné a štrukturálne utility

training_additional_data/ uložené modely strojového učenia

config.py konfigurácia prediktoru

feature.py базové triedy pre výpočet vlastností

predict.py prediktor SoluProtMut

README.md užívateľská príručka