VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV INFORMAČNÍCH SYSTÉMŮ
DEPARTMENT OF INFORMATION SYSTEMS

PREDIKCE ROZPUSTNOSTI PROTEINŮ

PREDICTION OF PROTEIN SOLUBILITY

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE **MARTIN MARUŠIAK**
AUTHOR

VEDOUCÍ PRÁCE **Ing. JIŘÍ HON**
SUPERVISOR

BRNO 2018

Zadání bakalářské práce/20632/2017/xmarus07

Vysoké učení technické v Brně - Fakulta informačních technologií

Ústav informačních systémů Akademický rok 2017/2018

Zadání bakalářské práce

Řešitel: **Marušiak Martin**
Obor: Informační technologie
Téma: **Predikce rozpustnosti proteinů**
Prediction of Protein Solubility
Kategorie: Bioinformatika

Pokyny:

1. Seznamte se s problematikou rozpustnosti proteinů a existujícími predikčními nástroji.
2. Vytvořte trénovací a testovací datovou sadu rozpustných a nerozpustných proteinů na základě dostupných dat.
3. Navrhněte nástroj pro predikci rozpustnosti proteinů a proveďte jeho implementaci.
4. Činnost nástroje zhodnoťte na testovací datové sadě, kterou použijte rovněž pro ohodnocení dostupných existujících nástrojů.
5. Zhodnoťte dosažené výsledky a diskutujte možnosti dalšího pokračování projektu.

Literatura:

- TRAINOR, Kyle, Aron BROOM a Elizabeth M MEIERING. Exploring the relationships between protein sequence, structure and solubility. *Current Opinion in Structural Biology*. 2017, **42**, 136-146
- Dále dle pokynů vedoucího.

Pro udělení zápočtu za první semestr je požadováno:

- Splnění bodů 1 a 2 zadání.

Podrobné závazné pokyny pro vypracování bakalářské práce naleznete na adrese http://www.fit.vutbr.cz/info/szz/

Technická zpráva bakalářské práce musí obsahovat formulaci cíle, charakteristiku současného stavu, teoretická a odborná východiska řešených problémů a specifikaci etap (20 až 30% celkového rozsahu technické zprávy).

Student odevzdá v jednom výtisku technickou zprávu a v elektronické podobě zdrojový text technické zprávy, úplnou programovou dokumentaci a zdrojové texty programů. Informace v elektronické podobě budou uloženy na standardním nepřepisovatelném paměťovém médiu (CD-R, DVD-R, apod.), které bude vloženo do písemné zprávy tak, aby nemohlo dojít k jeho ztrátě při běžné manipulaci.

Vedoucí: **Hon Jiří, Ing.**, UIFS FIT VUT
Datum zadání: 1. listopadu 2017
Datum odevzdání: 16. května 2018

L.S.

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
Fakulta informačních technologií
Ústav informačních systémů
612 66 Brno, Božetěchova 2

doc. Dr. Ing. Dušan Kolář
vedoucí ústavu

Abstrakt

Proteínová rozpustnosť je úzko spojená s použiteľnosťou proteínov pre účely priemyselného využitia a vo výskume. Predikcia rozpustnosti by preto viedla k značnému ušetreniu finančných prostriedkov. V tejto práci je prezentovaný nový prediktor rozpustnosti Solpex založený na strojovom učení, ktorý na nezávislej testovacej sade dosiahol vyššiu presnosť ako porovnateľné existujúce prediktory. Realizácii prediktoru predchádzalo oboznámenie s biologickou podstatou rozpustnosti, preskúmanie existujúcich prístupov k predikcii, tvorba dátových sád, uskutočnenie experimentov a výber vlastností pre prediktor. Najpodstatnejšou z týchto častí je pravdepodobne tvorba dátových sád, ktoré sú kľúčové pre vytvorenie kvalitného prediktoru. V súvislosti s dátovými sadami je v tejto práci podrobne popísané spracovanie hlavného zdroja ich dát – databázy TargetTrack.

Abstract

Protein solubility is closely related to the usability of proteins in industrial use and research. The successful prediction of solubility would therefore lead to a significant saving of financial resources. This work presents new solubility predictor Solpex based on machine learning that achieved better performance on independent test set than any comparable solubility prediction tool. The predictor implementation was preceded by a study of the biological nature of solubility, evaluation of existing solubility prediction approaches, datasets building, many experiments with novel features and selection of the best features for the predictor. As the most important step in machine learning is the datasets building, this work mainly benefits from own rigorous processing of the main source of solubility data – the TargetTrack database.

Kľúčové slová

rozpustnosť, proteín, predikcia, strojové učenie, TargetTrack, Solpex

Keywords

solubility, protein, prediction, machine learning, TargetTrack, Solpex

Citácia

MARUŠIAK, Martin. *Predikce rozpustnosti proteinů*. Brno, 2018. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Ing. Jiří Hon

Predikce rozpustnosti proteinů

Prehlásenie

Prehlasujem, že som túto bakalársku prácu vypracoval samostatne pod vedením pána Ing. Jiřího Hona. Ďalšie informácie mi poskytli Loschmidtove laboratória. Uviedol som všetky literárne pramene a publikácie, z ktorých som čerpal.

.....
Martin Marušiak
13. mája 2018

Poďakovanie

Rád by som sa poďakoval Ing. Jiřímu Honovi za odbornú pomoc a vedenie bakalárskej práce. Ďalej sa chcem poďakovať Ing. Tomášovi Martínkovi, Ph.D., za uvedenie do problematiky predikcie rozpustnosti proteínov a vedenie v predmete projektová prax. Ďakujem Loschmidtovým laboratóriám za poskytnutie odbornej pomoci v oblasti proteínovej rozpustnosti, menovite sa chcem poďakovať prof. Mgr. Jiřímu Damborskému, Dr., za umožnenie tejto spolupráce a experimentátorom Mgr. Antonínovi Kunkovi a Mgr. Davidovi Kovářovi, Ph.D., za pomoc pri určení expresného systému protokolov databázy TargetTrack. Na záver sa chcem poďakovať organizácii MetaCentrum za poskytnutie výpočtových zdrojov v rámci projektu CERIT Scientific Cloud (LM2015085) a CESNET (LM2015042) financovaných z programu MŠMT Projekty veľkých infraštruktúr pre VaVaI.

Obsah

1 Úvod **2**

2 Predikcia rozpustnosti proteínov **3**
2.1 Biologický úvod 3
2.2 Databázy a dátové sady 5
2.3 Nástroje 8

3 Tvorba dátových sád **14**
3.1 Schéma databázy TargetTrack 14
3.2 Predspracovanie databázy TargetTrack 16
3.3 Určenie expresného systému 17
3.4 Určenie živočíšnych domén 19
3.5 Tvorba trénovacej a testovacej sady 20

4 Experimenty **25**
4.1 K-mery 25
4.2 Vlastnosti založené na symboloch 27
4.3 Sekvenčné vzory 28
4.4 Terciárna štruktúra 32

5 Prediktor rozpustnosti **35**
5.1 Výber vlastností 35
5.2 Implementácia 37
5.3 Vyhodnotenie 37

6 Záver **42**

Literatúra **44**

A Obsah priloženého DVD **50**

Kapitola 1

Úvod

Proteíny sú prítomné vo všetkých živých organizmoch a podieľajú sa na mnohých biologických procesoch potrebných pre život. Medzi funkcie, ktoré v organizme zabezpečujú patrí napríklad stavebná, katalytická, transportná či obranná vo forme rôznych protilátok. Práve kvôli tomu sú proteíny objektom výskumu a priemyselného využitia.

Na tvorbu proteínov sú vynakladané pomerne značné finančné prostriedky. Tvorba samotných proteínov však nestačí, mnohé aplikácie vyžadujú, aby proteíny boli rozpustné. Jedná sa napríklad o proteíny používané vo farmaceutickom priemysle v podobe rôznych terapeutických proteínov. Preto sú vyvíjané nové technologické postupy na zvýšenie rozpustnosti proteínov. Okrem nich vzniká snaha rozpustnosť proteínu predikovať pomocou metód strojového učenia. Predikcia nemá vplyv na rozpustnosť proteínu, je ale nástrojom pre výber lepšie rozpustných proteínov. Problematikou rozpustnosti proteínov sa zaoberá aj táto práca, ktorej hlavným cieľom je vytvorenie prediktoru rozpustnosti.

Tvorbe samotného prediktoru rozpustnosti sa venuje kapitola 5. Obsahuje popis výberu vlastností, implementáciu, vyhodnotenie a porovnanie prediktoru s existujúcimi riešeniami. Výsledný prediktor dostal názov Solpex.

S tvorbou prediktoru úzko súvisia zdroje dát. Zatiaľ však neexistujú dostatočne veľké databázy vytvorené špeciálne za účelom skúmania rozpustnosti. Pre vytvorenie rozmanitej dátovej sady je preto potrebné vychádzať z databáz, ktoré rozpustnosť priamo neuvádzajú. S týmito databázami je spojená dodatočná réžia vyplývajúca z nutnosti databázu vopred spracovať a zvoliť vhodný spôsob odvodenia rozpustnosti. Príkladom takejto databázy je TargetTrack [9]. Databáza TargetTrack predstavuje základný zdroj dát mnohých nástrojov a vychádzajú z nej aj dátové sady prediktoru Solpex. Keďže je databáza TargetTrack hlavným zdrojom dát, bola pre jej spracovanie a následnú tvorbou dátových sád prediktoru Solpex vyhradená osobitná kapitola 3. Všeobecný popis databázy TargetTrack, ako aj iných databáz a dátových sád sa nachádza v podkapitole 2.2.

V súčasnosti existuje niekoľko prediktorov rozpustnosti proteínov. V tejto práci bolo preskúmaných deväť z nich. Princípmi vybraných nástrojov sa venuje podkapitola 2.3. V podkapitole nástrojov sú taktiež uvedené rôzne zistenia a znalosti, ktoré autori pri tvorbe prediktorov nadobudli.

Aj napriek existencii niekoľkých prediktorov rozpustnosti je táto oblasť stále pomerne nepreskúmaná. Neoddeliteľnou súčasťou tvorby prediktorov sú preto experimenty. Podobne predchádzalo aj tvorbe prediktoru Solpex vykonanie niekoľkých experimentov, pre ktoré bola vyhradená osobitná kapitola 4.

Kapitola 2

Predikcia rozpustnosti proteínov

Táto kapitola sa zaoberá úvodom do problému predikcie rozpustnosti. Obsahuje základne biologické poznatky z oblasti rozpustnosti proteínov, popis dostupných zdrojov dát a existujúcich prístupov k predikcii rozpustnosti.

2.1 Biologický úvod

Podkapitola biologický úvod obsahuje základné biologické znalosti z oblasti rozpustnosti proteínov. V súvislosti s rozpustnosťou sú v tejto časti ďalej popísané jednotlivé úrovne proteínovej štruktúry.

2.1.1 Rozpustnosť

Samotná tvorba proteínov nie je postačujúca a pre mnohé aplikácie sú potrebné rozpustné proteíny. Príkladom je použitie terapeutických proteínov vo farmaceutickom priemysle [7] či získanie štruktúry proteínov [54]. Nejedná sa však len o problém pri tvorbe proteínov, nízka rozpustnosť proteínov v tele môže spôsobiť vznik rôznych chorôb [65].

Rozpustnosť proteínu ovplyvňuje niekoľko faktorov, ktoré je možné rozdeliť na vnútorné a vonkajšie [32]. Vonkajšie faktory sa týkajú okolitého prostredia proteínu. Medzi ne patrí hodnota pH, iónová sila, teplota a prítomnosť rôznych dodatočných látok v roztoku [52]. Vnútorné faktory sú dané aminokyselinovou sekvenciou proteínu a to najmä aminokyselinami nachádzajúcimi sa na povrchu proteínu, ktoré sú vystavené okolitému prostrediu. K vnútorným vlastnostiam je taktiež možné zaradiť sklon proteínov k agregácii. Proteínová agregácia je jav, v ktorom dochádza k tvorbe nerozpustných proteínových zhlukov [62, 65]. Proteíny s náchylnosťou k agregácii majú teda negatívny dopad na rozpustnosť. Agregácia proteínov môže byť minimalizovaná pridaním vhodných chaperónov [42]. Chaperóny predstavujú špeciálny typ proteínov, ktoré napomáhajú ostatným proteínom pri ich skladaní [20]. Ďalším dôležitým faktorom je spôsob tvorby proteínov. Proteíny sú typicky produkované v organizmoch, v ktorých sa prirodzene nenachádzajú, jedná sa o tzv. rekombinantnú tvorbu proteínov. Tvorbu rekombinantných proteínov umožňujú metódy genetickej modifikácie. Organizmus resp. platforma, pomocou, ktorej je proteín vytvorený sa označuje pojmom expresný orgazmus, resp. systém. Často používaným expresným organizmom je baktéria *E. coli* [53]. 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.

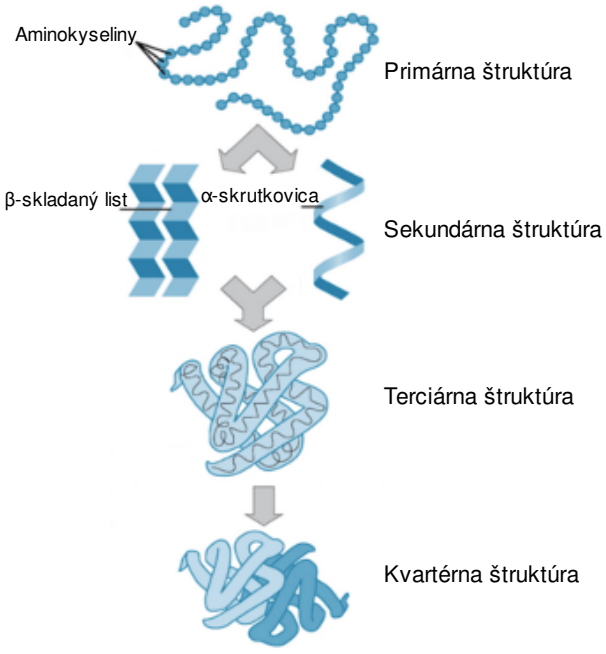
2.1.2 Štruktúra proteínov

Proteíny sú biologické makromolekuly zložené zo sekvencie aminokyselín. Prirodzene sa v proteínoch vyskytuje 20 rôznych druhov aminokyselín.

Proteíny majú niekoľko úrovní štruktúry, prvá – primárna úroveň, predstavuje samotnú aminokyselinovú sekvenciu. Teda definuje to, ako jednotlivé aminokyseliny sa sebou nasledujú. Ďalšie úrovne postupne špecifikujú vzťahy medzi jednotlivými aminokyselinami sekvencie, napr. ich priestorové usporiadanie. Sekundárnu štruktúru tvoria lokálne zložené časti sekvencie, patria sem dva typy útvarov: α-skrutkovice a β-skladané listy. Niektoré časti sekvencie sú neusporiadané a nepatria ani k jednému z uvedených typov. Ďalšou úrovňou je terciárna, ktorá predstavuje trojdimenzionálne usporiadanie proteínu. Poslednú úroveň tvorí kvartérna štruktúra, túto štruktúru majú len proteíny zložené z niekoľkých aminokyselinových sekvencií [28]. Jednotlivé úrovne štruktúry sú znázornené na obr. 2.1.

V prípade predikcie rozpustnosti je typicky k dispozícii len primárna štruktúra proteínu, vyššie úrovne často nie sú známe. Aj napriek tomu sa vyššie úrovne štruktúry pri predikcii používajú. Príkladom je použitie sekundárnej štruktúry v nástrojoch ESPRESSO [24] a DeepSol [29]. Pre jej získanie je však nutné použiť dodatočné nástroje na predikciu sekundárnej štruktúry.

S priestorným usporiadaným a terciárnou štruktúrou proteínu súvisia domény a vinutia (foldy). Doména predstavuje časť proteínu, ktorá je z pohľadu skladania nezávislá. Skladanie je proces, pri ktorom proteín postupne prechádza z primárnej do terciárnej štruktúry. Proteíny môžu obsahovať niekoľko domén. K jednotlivým doménam sa často viaže určitá špecifická funkcia, ktorú v rámci proteínu vykonávajú [16]. Jednotlivé typy domén je možné klasifikovať do vinutí, rodín či super rodín vinutí. Medzi najznámejšie klasifikačné databázy patrí SCOP [39] a CATH [45]. Podľa [43] môže spôsob skladania proteínu a jeho vinutia taktiež zohrávať rolu v rozpustnosti proteínov.



Obr. 2.1: Zobrazenie jednotlivých úrovní štruktúry proteínu. Poradie je uvedené tak, aby zodpovedalo procesu skladania proteínu z primárnej do kvartérnej štruktúry [28].

2.2 Databázy a dátové sady

Problém predikcie rozpustnosti je do veľkej miery spojený s problémom vhodnej voľby dát. V oblasti rozpustnosti proteínov existuje len málo databáz, v ktorých je rozpustnosť uvedená priamo a sú zároveň dostatočne veľké. Táto podkapitola popisuje niekoľko zdrojov dát, ich výhody a nevýhody.

2.2.1 eSOL

Autori databázy eSOL [43] vyhodnotili rozpustnosť kompletnej sady proteínov baktérie *E. coli*. Uvedené proteíny boli vytvorené pomocou bezbunkového systému expresie PURE [56]. Z celkového počtu 4 132 proteínov *E. coli* uvedených v knižnici ASKA [30] sa v dostatočnom množstve podarilo vytvoriť 3 173 proteínov. Rozpustnosť týchto proteínov je ohodnotená percentuálne a bola získaná s využitím centrifúgy. Autori vytvorili histogram z hodnôt rozpustnosti jednotlivých proteínov a zistili, že má bimodálny charakter, na základe ktorého je možné klasifikovať proteíny do dvoch kategórií: rozpustné a náchylné k agregácii.

V práci [43] okrem vyhodnotenia rozpustnosti taktiež skúmali vplyv rôznych faktorov na rozpustnosť. Uvádzajú, že proteíny s vyšším obsahom záporne nabitých aminokyselín (D, E) alebo nízkym obsahom aromatických aminokyselín majú tendenciu byť rozpustné. Ďalej overovali vplyv sekundárnej a terciárnej štruktúry. Sekundárna štruktúra nevykazovala žiadnu závislosť s rozpustnosťou. V prípade terciárnej štruktúry sa naopak závislosť prejavila. Tá spočívala vo výraznej prevahe jednej kategórie proteínov v niektorých typoch vinutia. Zaradenie do vinutí sa riadilo klasifikáciou SCOP [39].

V neskoršej publikácii [42] zisťovali vplyv chaperónov na rozpustnosť. Experiment uskutočnili na takmer 800 cytoplazmatických proteínoch, ktoré boli v prechádzajúcej práci [43] zaradené do kategórie proteínov náchylných k agregácii. Na expresiu bol taktiež použitý systém PURE, do ktorého postupne pridávali rôzne druhy chaperónov. Pridanie chaperónov zvýšilo rozpustnosť o viac ako 50 % v dvoch tretinách proteínov. Rozpustnosť sa nepodarilo zvýšiť o 20 a viac percent len v prípade 3 % proteínov.

Výhodou databázy eSOL je jednotný systém expresie, percentuálna hodnota rozpustnosti s a bez pridania chaperónov. Medzi nevýhody patrí spôsob merania rozpustnosti. Autori totiž pripúšťajú, že do kategórie rozpustných proteínov mohli byť zaradené aj potenciálne nerozpustné proteíny, akými sú napríklad zhluky oligomérov. Ďalšou nevýhodou je malá diverzita organizmov, z ktorých proteíny pochádzali, keďže autori exprimovali len proteíny *E. coli*. Hoci je systém PURE bezbunkový, vychádza z faktorov a látok potrebných práve pre expresiu proteínov *E. coli* [43]. Z hľadiska predikcie sú ale zaujímavé aj informácie o rozpustnosti proteínov v nevlastnom prostredí, tie eSOL bohužiaľ neobsahuje.

Databáza je dostupná online vo formáte CSV. Okrem rozpustnosti obsahuje len identifikáciu sekvencie, samotnú sekvenciu je nutné získať dodatočne z iných zdrojov.

2.2.2 TargetTrack

Databáza TargetTrack [6] slúžila na zhromažďovanie informácií z projektov PSI (*Protein Structure Initiative*), ktorých cieľom bolo získanie štruktúr veľkého množstva proteínov. Jednalo sa o tzv. vysoko priepustné projekty (*high-throughput*) štrukturálnej genomiky využívajúce rôzne technológie získania štruktúry a tvorby proteínov. Tieto projekty trvali od roku 2000 do 2015, za toto obdobie sa im podarilo zhromaždiť experimentálne informácie o 297 tisíc sekvenciách. Na uloženie týchto dát pôvodne slúžila databáza TargetDB a data-

báza PepcDB, ktorá obsahovala aj informácie o použitých protokoloch. Neskorším spojením týchto databáz vznikla databáza TargetTrack [9].

Najdôležitejší prvok databázy je tzv. cieľ (*target*). Cieľ reprezentuje skúmanú sekvenciu a obsahuje informácie o priebehu experimentov. Hoci TargetTrack neuvádza rozpustnosť proteínov, je možné ju odvodiť práve zo stavu experimentov. Ďalším dôležitým prvkom sú protokoly. Tie obsahujú detailné informácie o použitom postupe v rozličných fázach experimentu. TargetTrack obsahuje až 1 494 rôznych protokolov. Celkový počet cieľov je 335 711, každý cieľ pozostáva z experimentov, ktorých je v celej databáze TargetTrack 961 548. Rôzne ciele môžu skúmať rovnakú sekvenciu, preto je počet cieľov väčší ako počet unikátnych sekvencií (297 404). Doménové zloženie databázy je podľa pôvodu cieľov nasledovné: 64,7 % baktérie, 26,3 % eukaryoty, 5,9 % archea, 0,2 % vírusy a 2,9 % cieľov s neuvedeným alebo neidentifikovaným organizmom.[1]

Výhodou databázy TargetTrack je jej veľkosť. Tá však so sebou prináša zvýšené nároky na jej spracovanie. Jedným z problémov je určenie expresného systému, ktorý nie je explicitne uvedený. Pre identifikovanie expresného systému je nutné prechádzať texty protokolov. Tieto texty nemajú pevnú štruktúru a ich automatické spracovanie strojom nie je priamočiare. Expresný systém má veľký vplyv na rozpustnosť a jeho neznalosť by potencionálne viedla k zhoršeniu kvality dátovej sady.

Databáza TargetTrack bola uzatvorená v júly 2017 a nie je možné do nej pridávať ďalšie dáta alebo ich meniť. Celá databáza je dostupná online[2] vo formáte XML spolu s jej schémou a dokumentáciou. Podstatné informácie o schéme databázy a jej samotnom spracovaní sa nachádzajú v kapitole 3, kde je taktiež detailne popísaný spôsob odvodenia rozpustnosti a expresného systému.

2.2.3 NESG

Jedným z hlavných stredísk, ktoré prispievalo do databázy TargetTrack v rámci projektov PSI je NESG (*The Northeast Structural Genomics*) [40]. Stredisko NESG má taktiež vlastnú databázu, tá navyše obsahuje mieru rozpustnosti a expresie proteínov uvedenú vo forme celého čísla od 0 do 5, kde 5 je maximálna hodnota rozpustnosti/expresie [69].

V práci [51] použili podmnožinu NESG zloženú z 9 644 proteínov. Podľa autorov článku boli tieto proteíny vytvorené jednotným spôsobom expresie a to v období od roku 2001 do 2008. Zloženie tejto sady je podľa pôvodu cieľov nasledovné: 82 % baktérie, 12 % archea, 6 % človek a 0,3 % iné eukaryoty. Proteíny boli exprimované *in vivo* v *E. coli*. Medzi hlavné výhody tejto sady patrí: nebinárna hodnota rozpustnosti a expresie, jednotný postup a známy expresný systém.

2.2.4 PDB

Proteínová dátová banka (PDB) [5] plní účel jednotného celosvetového archívu štruktúr biologických makromolekúl. V súčasnosti obsahuje až 137 tisíc štruktúr. Hoci sa v nej priamo neuvádza informácia o rozpustnosti, je možné ju odvodiť od metódy použitej na získanie štruktúry. Niektoré z týchto metód totiž vyžadujú proteíny, ktoré sú do určitej miery rozpustné, príkladom takejto metódy je röntgenová kryštalografia (*X-Ray*). Väčšina štruktúr PDB bola získaná práve touto metódou, jedná sa cca 123 tisíc štruktúr.

[1] Vyššie uvedené číselné údaje boli získané analýzou databázy TargetTrack, viď kapitolu 3.
[2] http://dx.doi.org/10.5281/zenodo.821654

Štruktúry pridané do PDB prechádzajú procesom validácie [5], jedná sa preto o relatívne spoľahlivý zdroj dát. Z PDB je možné získať len rozpustné proteíny. Túto databázu je teda možné využiť pre doplnenie rozpustných proteínov do dátovej sady alebo na konfrontovanie ohodnotenia rozpustnosti voči iným databázam.

Databáza je dostupná online[3] a poskytuje filtre, ktorými je možné obmedziť výber sekvencií podľa zdrojového organizmu, expresného systému, použitej metódy, vlastností štruktúry a mnohých ďalších.

2.2.5 Atlas ľudských proteínov

Atlas ľudských proteínov (AĽS) [64, 63] je pôvodom švédsky projekt, ktorý si dal za cieľ zmapovať všetky ľudské proteíny v bunkách, tkanivách a orgánoch. Dáta poskytnuté od AĽS využili v [55] na tvorbu dátovej sady pre prediktor rozpustnosti. Vytvorená dátová sada spolu s prediktorom je dostupná na repozitári GitHub[4]. Sada obsahuje 16 082 sekvencií, nejedná sa však o sekvencie celých proteínov, ale len ich fragmentov. Sekvencie sú preto pomerne krátke, väčšina z nich pozostáva z 20 až 150 aminokyselín. Všetky fragmenty boli exprimované v *E. coli*. Rozpustnosť každého fragmentu je určená číslom od 1 do 5, kde medzi každými dvoma celými číslami uvedeného rozsahu existuje práve jeden medzistupeň. Rozpustnosť môže teda nadobúdať 9 rôznych hodnôt, čím vyššia je hodnota, tým lepšia je rozpustnosť.

V porovnaní s ostatnými dátovými sadami, ktoré boli v rámci práce zozbierané, mala uvedená sada fragmentov v niektorých prípadoch opačné tendencie korelácie rozpustnosti so sekvenčnými vlastnosťami. Tento jav je pravdepodobne spôsobený tým, že sa jedná len o fragmenty proteínov a nie celé proteíny.

[3] https://www.rcsb.org/
[4] https://github.com/SBRG/Protein_ML

2.3 Nástroje

V súčasnosti existuje niekoľko nástrojov na predikciu rozpustnosti, tie je možné principiálne rozdeliť do dvoch kategórií, globálne a profilovo založené. Globálne vychádzajú z vlastností, ktoré sú určené na úrovni celej sekvencie. Narozdiel od nich využívajú profilovo založené metódy vlastnosti vzťahujúce sa len k časti sekvencie. Zoskupenie hodnôt vlastností jednotlivých častí vytvára profil. Možná je taktiež kombinácia vyššie uvedených prístupov, teda niektoré vlastnosti sa vzťahujú k častiam sekvencie a iné sú globálne.

Táto podkapitola zhŕňa základne informácie o vybraných nástrojoch, popisuje ich zdroje dát, predikčnú metódu a jej presnosť. Jednotlivé nástroje sú v texte usporiadané vzostupne podľa dátumu ich publikácie.

V nasledujúcom texte, ako aj vo zvyšku práce sú pre prehľadnosť uvádzané skratkové označenia aminokyselín. Označenie aminokyselín sa riadi podľa štandardu IUPAC[5].

2.3.1 Model Wilkinson-Harrison

V článku [68] bol publikovaný jeden z prvých modelov rozpustnosti. Autori vychádzali z pomerne malej dátovej sady o veľkosti 81 proteínov. Za rozpustné považovali tie proteíny, ktoré neviedli k vzniku bunkových inklúzií. Na výber vlastností použili diskriminačnú analýzu. Najväčší vplyv na rozpustnosť malo nasledujúcich 6 vlastností: priemerný náboj, podiel aminokyselín tvoriacich otočku, zastúpenie cysteínu (C), zastúpenie prolínu (P), hydrofobicita a celkový počet aminokyselín. Uvedené vlastnosti sú zoradené vzostupne podľa hodnoty korelácie danej vlastnosti a rozpustnosti. Silnú koreláciu mali len prvé dve, zvyšné 4 vlastnosti korelovali slabo.

Neskôr bol publikovaný článok [13], kde predošlý model upravili. Ukázalo sa, že len 2 zo 6 parametrov z pôvodného modelu majú výrazný vplyv na rozpustnosť. Týmito parametrami sú priemerný náboj a podiel aminokyselín tvoriacich otočky. Tento model sa označuje ako modifikovaný model Wilkinsona-Harrisona. Modifikovaný model používa pre výpočet zložený parameter CV, ktorý v sebe zahŕňa priemerný náboj a podiel aminokyselín tvoriacich otočky. Výpočet CV je daný rovnicou 2.1.

$$CV = 15{,}43\left(\frac{N+G+P+S}{n}\right) - 29{,}56\left|\frac{(R+K)-(D+E)}{n} - 0{,}03\right| \tag{2.1}$$

Kde n je celkový počet aminokyselín v proteíne. Každé z písmen N, G, P, S, R, K, D, E predstavuje skratkové označenie aminokyselín. Prvý zlomok rovnice 2.1 predstavuje zastúpenie aminokyselín tvoriacich otočku a druhý priemerný náboj.

Rozpustnosť je určená na základe znamienka výrazu $CV - CV'$, kde CV' je diskriminant s hodnotou 1,71. Ak je daný výraz záporný potom je proteín predikovaný ako rozpustný a naopak, ak je hodnota kladná bude proteín predikovaný ako nerozpustný. Autori ďalej uvádzajú vzorec 2.2 na výpočet pravdepodobnosti (Pst) výsledku predikcie.

$$Pst = 0{,}4934 + 0{,}276|CV - CV'| - 0{,}0392(CV - CV')^2 \tag{2.2}$$

Presnosť tejto metódy sa v článku neuvádza. Tento model bol však otestovaný autormi nástroja PROSO [58] na ich dátovej sade, kde dosiahol presnosť 57,6 %, ďalej na sade nástroja SOLpro [36] s presnosťou 53,75 % a na dátovej sade vytvorenej v tejto práci získal presnosť 53,91 %.

[5] http://www.bioinformatics.org/sms2/iupac.html

2.3.2 PROSO

Nástroj PROSO [58] používa pre predikciu rozpustnosti dvojvrstvový model strojového učenia. Prvá vrstva predikuje rozpustnosť pomocou SVM s gausovským jadrom. Na tejto vrstve sa nachádzajú tri klasifikátory. Vstupom prvého SVM klasifikátoru je zastúpenie jednotlivých aminokyselín (*amino acid frequency*), vstupom druhého je zastúpenie dvojíc aminokyselín (dimérov) a podobne je vytvorený klasifikátor pre trojice aminokyselín (triméry). Druhá vrstva používa naivný bayesovský klasifikátor, ktorého vstupom sú výsledky SVM klasifikátorov prvej vrstvy.

V prípade dimérov a trimérov vzniká pomerne veľký počet vlastností (*features*), 400 pre diméry a 8 000 pre triméry. Tie predstavujú všetky možné permutácie s opakovaním, ktoré môžu z aminokyselín vzniknúť. Počet týchto vlastností znížili zhlukovaním jednotlivých aminokyselín do tzv. schém. Zhlukovanie prebiehalo na základe podobných fyzikálnochemických a štrukturálnych vlastností aminokyselín. Autori vytvorili niekoľko schém zhlukovania. Pre jednotlivé stupne k-merov vybrali najlepšiu schému. Zavedením schém docielili zmenšenie aminokyselinovej abecedy, a tým aj zníženie počtu rôznych permutácií. Jednotlivé vlastnosti boli ďalej vybrané s využitím obaľovacej (*wrapper*) metódy [31]. Najvýznamnejšou vlastnosťou bolo zastúpenie záporne nabitej dvojice aminokyselín DE. Ďalej medzi významné vlastnosti patrilo zastúpenie aminokyselín R, D, C, E, G, L, M, S a dimérov RE, QA, EG, HM a KG.

V článku stanovili presnosť tohto prediktoru na 72 % a MCC 0,434. Pri tvorbe dátovej sady čerpali hlavne z databázy TargetDB a PDB. Výsledná dátová obsahovala cca 14 000 proteínov. V prípade databázy TargetDB označili za rozpustné tie proteíny, ktoré dosiahli status rozpustný[6], ak proteín dosiahol len status exprimovaný a tento status pretrvával dlhšie časové obdobie, potom bol označený ako nerozpustný. Pri získaní sekvencií z PDB bola situácia jednoduchšia, nakoľko autori vychádzajú z predpokladu, že všetky proteíny v PBD sú rozpustné. Z proteínov získaných z TargetDB boli navyše odstránené tie proteíny, ktoré boli označené ako nerozpustné a zároveň boli identické s niektorým z proteínov PDB. Ďalej zo sady odstránili transmembránové proteíny pomocou nástroja TMHMM [33] a taktiež znížili redundanciu dátovej sady pomocou nástroja CD-HIT [19], hranica identity bola 50 %.

2.3.3 SOLpro

SOLpro [36] je ďalším nástrojom, ktorý pre predikciu používa SVM. Jedná sa o dvojfázový prediktor. V prvej fáze je vytvorených 20 SVM modelov, každý z nich je natrénovaný na inej množine vlastností. V druhej fáze sa vytvorí finálny SVM model. Jeho vstupmi sú výstupy SVM modelov prvej fázy spolu s dĺžkou sekvencie. Vstupom druhej fázy je teda 21 vlastností.

Pôvodne bolo vytvorených 23 sád vlastností, z ktorých 21 vychádza zo zastúpeniu monomérov, dimérov a trimérov, autori pri ich tvorbe použili 7 rôznych aminokyselinových abecied. Zvyšné dve sady sú označené ako *computed* (vypočítané vlastnosti) a *precomputed* (predikované vlastnosti). Sada vlastností *computed* sa skladá z vlastností, ktoré je možné vypočítať zo sekvencie. Táto sada obsahuje 6 vlastností, patrí sem dĺžka sekvencie, zastúpenie aminokyselín tvoriacich otočku, absolútna hodnota náboja, molekulárna hmotnosť, GRAVY (*grand average of hydropathy*) index a alifatický index. Oproti tomu sada *precomputed* pozostáva z vlastností, ktoré boli odvodené zo sekvencie pomocou predikčných

[6] Prípadne iné statusy, na základe, ktorých je možné o proteíne prehlásiť, že je rozpustný.

nástrojov SCRATCH [35]. Sada *precomputed* obsahuje pomer aminokyselín tvoriacich α- a β-štruktúry, počet domén a pomer vystavených aminokyselín. Autori použili na identifikovanie vhodných sád vlastností obaľovaciu metódu [31], čo viedlo k odstráneniu 3 sád. Jednalo sa o sady, ktoré vychádzali zo zastúpenia monomérov, dimérov a trimérov.

Na vyhodnotenie nástroja bola použitá metóda 10-násobnej krížovej validácie. Nástroj dosiahol presnosť 74,16 % a 0,487 MCC. Testovanie a trénovanie nástroja prebiehalo na sade SOLP, ktorú vytvorili na základe databáz TargetDB, PDB, SwissProt [61] a nepatrnú časť (175 proteínov) získali z literárnych prameňov. Dátová sada SOLP pozostáva z 17 408 sekvencií a obsahuje rovnaký počet rozpustných a nerozpustných proteínov. Pri tvorbe sady postupovali podobne ako autori nástroja PROSO [58]. Okrem transmembránových proteínov navyše vylúčili proteíny, ktorých dĺžka nebola v intervale 20 až 2 000 aminokyselín. Ďalej vylúčili tie sekvencie, ktoré obsahovali dve a viac po sebe idúcich neidentifikovaných aminokyselín. Redundanciu odstránili pomocou nástroja BLASTCLUST [4]. Sekvencia bola považovaná za redundantnú, ak miera podobnosti jej časti s častou inej sekvencie presiahla 25 %, zároveň však musela táto časť pokrývať aspoň 50 % jednej zo sekvencií.

2.3.4 ccSOL

Ďalším prediktorom, ktorý používa SVM je nástroj ccSOL [2]. Tento nástroj používa na trénovanie a testovanie databázu eSOL, ktorú rozdelili na 3 časti: hlavička – najviac rozpustné proteíny, chvost – najmenej rozpustné proteíny a zvyšok. Hlavička obsahovala 1 081 sekvencií a chvost 1 078. Pre každú z týchto sekvencií bolo následne vypočítaných 28 fyzikálnochemických vlastností. Autori postupne zisťovali, ktoré z týchto vlastností vedia najlepšie rozlišovať proteíny zo sady hlavičky od sady chvostu a naopak. Zistili, že najlepšie rozlišuje hlavičku od chvosta 11 vlastností, pričom najvýznamnejšie z nich sú 3 vlastnosti: neusporiadanosť, zastúpenie náhodných cievok (*coil*) a hydrofobicita. V ďalšom kroku vytvorili 2^{11} SVM modelov, jeden pre každú kombináciu 11-tich vlastností. Kvalitu modelov vyhodnocovali pomocou 10 násobnej krížovej validácie. Najlepšie výsledky dosahoval model založený na 6-tich vlastnostiach: zastúpení náhodných cievok, neusporiadanosti, hydrofobicite, hydrofílii, zastúpení β-skladaných listov a α-skrutkovíc. Presnosť tohto prediktoru nie je v článku uvedená. V neskoršej publikácii vzťahujúcej sa k nástroju ccSOL omics [1], na ktorej pracovali aj autori nástroja ccSOL, sa uvádza presnosť tejto metódy 76 %.

2.3.5 PROSOII

PROSOII [57] je podobne ako nástroj PROSO založený na dvoch vrstvách. Na prvej vrstve je použitá metóda Parzenovho okna a logistickej regresie. Druhá vrstva je tvorená logistickou regresiou a jej vstupmi sú výstupy modelov prvej vrstvy.

Metóda Parzenovho okna [46] je použitá za účelom vytvorenia modelu sekvenčnej podobnosti. Na určenie miery podobnosti jednotlivých sekvencií bol použitý nástroj BLAST [4] a hodnota BLAST skóre. Rozpustnosť je určená na základe vzťahu 2.3. Kde SS predstavuje mieru podobnosti sekvencie k sade rozpustných a SI k sade nerozpustných proteínov.

$$f(t) = \frac{SS}{SS + SI} \tag{2.3}$$

Ďalej je na prvej vrstve použitá logistická regresia, ktorá pracuje zo zastúpením monomérov a dimérov. Z celkového počtu 20 monomérov použili 18 a zo 400 dimérov použili len 13. V prípade monomérov nepoužili A a L. Z dimérov boli zvolené: AK, CV, EG,

GN, GH, HE, IH, IW, MR, MQ, PR, TS a WD. Výber sa riadil podľa obaľovacej metódy [31]. Autori si všimli, že až 8 z 13 vybraných dimérov obsahuje elektricky nabité postranné reťazce (negatívne D, E a pozitívne R, H, K), čo je v súlade zo zisteniami v práci Wilkinsona-Harrisona [68]. Ďalšie často sa vyskytujúce skupiny aminokyselín boli: aromatické hydrofóbne (F, W, H, Y) a alifatické hydrofóbne (I, M, P). Podľa článku [15] má vysoký podiel aromatických aminokyselín F, Y a W negatívny dopad na rozpustnosť. Autori PROSOII taktiež uvádzajú, že významnú úlohu v rozpustnosti môže mať aj vysoké zastúpenie hydrofóbnych dimérov. Pri predikcii taktiež použili alifatický index, GRAVY index a izoelektrický bod.

Nástroj bol vyhodnotený pomocou 10 násobnej krížovej validácie, jeho presnosť je 71 % a korelácia MCC 0,421. Dátová sada, ktorá bola v rámci nástroja vytvorená obsahuje až 82 000 proteínov. Táto sada vychádza hlavne z databáz PepcDB a PDB. Postup jej tvorby je principálne podobný ako v nástroji SOLpro, resp. PROSO. Sadu navyše vyvážili z hľadiska dĺžok sekvencií tak, aby bolo zastúpenie proteínov s rôznymi dĺžkami zhruba rovnaké medzi sadou rozpustných a nerozpustných proteínov.

2.3.6 SCM

Metóda, ktorú používa nástroj SCM [26] (*Scoring Card Method*), pracuje len zo zastúpením dimérov v proteíne. Keďže existuje až 20 aminokyselín celkový počet rôznych dimérov je 400. Následne je získaný počet výskytov jednotlivých dimérov zvlášť pre rozpustné a nerozpustné proteíny. Výstupom sú dve tzv. skórovacie matice rozpustnosti (SSM), jedna pre nerozpustné a jedna pre rozpustné proteíny, každá z nich obsahuje 400 hodnôt. Tieto hodnoty sú následné normalizované celkovým počtom dimérov v danej sade. V ďalšom kroku sa z týchto matíc vytvorí len jedna, ktorá vznikne rozdielom skórovacej matice nerozpustných a rozpustných proteínov. Táto matica je následne optimalizovaná pomocou evolučného algoritmu IGA [25], kritériom optimalizácie (*fitness*) bola plocha pod ROC krivkou (AUC), ktorá sa používa na ohodnotenie kvality binárnych klasifikátorov. Samotná predikcia prebieha na základe hodnoty $P(s)$, získanej pomocou vzťahu 2.4, ak je táto hodnota vyššia ako zvolený prah, potom je proteín označený za rozpustný, inak ako nerozpustný. Zastúpenie konkrétneho diméru v sekvencii reprezentuje w_i, táto hodnota je následne vynásobená prvkom skórovacej matice S_i, kde pozícia v matici i určuje typ diméru.

$$P(s) = \sum_{i=1}^{400} w_i S_i \tag{2.4}$$

Presnosť tohto prediktoru je 84 % na dátovej sade sd957 a 60 % na sade SOLproDB. Dátová sada SOLproDB vychádza z dátovej sady nástroja SOLpro, autori SCM z nej ale odstránili proteíny obsahujúce neidentifikované aminokyseliny, po tejto redukcii obsahuje 16 902 sekvencií. Dátová sada sd957 sa od SOLproDB líši tým, že obsahuje len proteíny s experimentálne potvrdenou rozpustnosťou a jednotnými experimentálnymi podmienkami. Táto sada nie je vyvážená a obsahuje 285 rozpustných a 672 nerozpustných proteínov, dokopy sa jedná o 957 sekvencií, čo je pomerne málo.

Autori vykonali experiment, v ktorom prebiehalo trénovanie SCM na sade sd957 a testovanie na sade SOLproDB. Dosiahli presnosť len 53,9 %. Taktiež uskutočnili experiment, v ktorom použili nástroj SOLpro a testovali ho na vlastnej sade (sd957). V tomto prípade bola presnosť SOLpro ešte menšia, a to 49,21 %. Podľa autorov môže byť nízka hodnota presnosti spôsobená rozdielnym charakterom dátových sadách z pohľadu experimentálnych podmienok.

2.3.7 ESPRESSO

Nástroj ESPRESSO [24] predikuje rozpustnosť dvomi rôznymi metódami. Prvá z nich je založená na sekvenčných a štrukturálnych vlastnostiach proteínu. Medzi sekvenčné vlastnosti zaradili nukleotidové, kodónové a aminokyselinové zloženie. Štrukturálne vlastnosti boli získané pomocou predikčných nástrojov. Medzi štrukturálne vlastnosti zaradili sekundárnu štruktúru, relatívne prístupný povrch (*relative accessible surface area*), neusporiadané a transmembránové regióny. Pomocou Studentovho testu vybrali tie vlastnosti, pre ktoré bolo $p < 0{,}05$. Zistili, že v prípade proteínov exprimovaných spôsobom *in vivo* je rozpustnosť spojená s nabitými aminokyselinami, ktoré zvyknú byť na povrchu proteínov. V prípade bezbunkových systémov nepozorovali závislosť medzi nabitými aminokyselinami a rozpustnosťou. Vybrané vlastnosti boli následne použité ako vstupy pre strojové učenie, konkrétne sa jednalo o SVM, ktoré dosahovalo najlepšie výsledky. Okrem SVM experimentovali s metódou neurónových sietí a náhodných lesov.

Druhý spôsob predikcie je založený na vzoroch. Autori sa snažili identifikovať často sa vyskytujúce vzory – skupiny aminokyselín, ktoré sú typické pre rozpustné a nerozpustné proteíny. Zistili, že najvhodnejšia veľkosť vzoru je 6 až 7 aminokyselín. Taktiež uvádzajú, že vzorov by nemalo byť príliš veľa, inak dôjde k pretrénovaniu a zároveň ich nesmie byť príliš málo, pretože by sa v danej sekvencii nemusel nájsť žiadny vzor. Ďalej si všimli, že vzory typické pre rozpustné proteíny sú veľmi často na povrchu proteínu, a naopak niektoré vzory nerozpustných proteínov sa nachádzajú v jadre. Samotná predikcia je založená na diskriminačnej funkcii, ktorá pracuje s počtom vzorov v dátovej sade a počtom nájdených vzorov v predikovanej sekvencii.

Pre trénovanie a testovanie použili Hirosovu dátovú sadu [23]. Táto sada čerpá z výsledkov experimentov, v ktorých merali expresiu a rozpustnosť ľudských proteínov v dvoch rôznych expresných systémoch. Prvým je systém typu *in vitro* – bezbunková expresia (*wheat germ cell-free expression system*) a druhým expresia *in vivo* v baktérii *E. coli*. Výhodou tejto sady je, že obsahuje sekvencie vo forme cDNA. Proteíny rozdelili do dvoch sád označených ako *dataset_Single* a *dataset_Multi*. Sada *dataset_Single* obsahuje len proteíny, ktorých vlastnosti boli namerané len v jednom (*single*) experimente. Naopak sada *dataset_Multi* obsahuje tie proteíny, ktoré sa vyskytli vo viacerých (*multi*) experimentoch. Ďalej pomocou nástroja CD-HIT odstránili podobné sekvencie s prahom 80 % a následne odstránili sekvencie s mierou identity >40 %. Na záver znížili identitu ešte viac a to na 25 % pomocou nástroja ALIGN0 [47]. Veľkosť sád pre rôzne expresné systémy sa pohybuje v ráde stoviek pre sadu *multi* a v ráde tisícov pre sadu *single* (cca 5 000). Výsledné sady boli nevyvážené.

Pri testovaní na proteínoch exprimovaných v *E. coli* dosiahla metóda založená na sekvenčných a štrukturálnych vlastnostiach presnosť 68 % a MCC 0,42. Druhá metóda využívajúca vzory na tom bola horšie s MCC len 0,23 a presnosťou 63 %.

2.3.8 ccSOL omics

Narozdiel od predchádzajúcich metód, ktoré vychádzajú z globálnych vlastností sekvencie, je metóda ccSOL omics [1] založená na sekvenčnom profile. Profil je zložený z hodnôt rozpustností jednotlivých fragmentov sekvencie. Jeden fragment pokrýva 21 aminokyselín a jeho rozpustnosť sa vypočíta na základe metódy ccSOL [2]. Prvý fragment začína od N-konca sekvencie, nasledujúci fragment je vždy posunutý o jednu aminokyselinu bližšie k C-koncu sekvencie. Jedná sa teda o mechanizmus posuvného okna o veľkosti 21 aminokyselín. Dĺžka profilu zaleží od dĺžky sekvencie. Aby mohol nástroj porovnávať rôzne dlhé sekvencie, transformovali profil na Fourierove koeficienty. Z nich použili prvých 100 ako vstup pre

neurónové siete. Autori stanovili presnosť tejto metódy na 74 %. Vyhodnotenie prebiehalo na testovacej sade zloženej z 31 760 proteínov, ktoré si boli vzájomné podobné z menej ako 30 %. Uvedená testovacia sada vychádzala zo sád vytvorených nástrojmi PROSOII, SOLpro a databázy eSOL.

2.3.9 DeepSol

Najnovším nástrojom v oblasti predikcie proteínov je nástroj DeepSol [29]. Na predikciu používa hlboké učenie v podobe konvolučných neurónových sietí. Autori navrhli 3 rôzne architektúry prediktoru, ktoré označili ako DeepSol S1, DeepSol S2 a DeepSol S3. Prvá z nich, DeepSol S1, predikuje rozpustnosť len na základe aminokyselinovej sekvencie. DeepSol S2 a S3 používajú navyše dodatočných 57 vlastností. Rozdiel medzi verziou S2 a S3 je v spojení vektoru vlastností a vektoru získaného zo sekvencie. DeepSol S2 tieto vektory spojí priamo, narozdiel od verzie S3, ktorá najskôr transformuje vektor vlastností pomocou dopredných neurónových sietí a až potom ho pridá k vektoru sekvencie.

Fixný počet prvkov pre konvolučné neurónové siete zabezpečili pomocou reprezentácie sekvencie v podobe postupnosti vektorov s obmedzením na maximálnu dĺžku. Každá sekvencia je prevedená na postupnosť vektorov x_1, x_2, až x_L, kde x_i je binárny vektor s 21 prvkami, 20 z nich je vyhradených pre aminokyseliny a 1 pre medzeru. Aktívny je vždy len jeden prvok (kódovanie 1 z n), aktívny prvok korešponduje s i-tou aminokyselinou sekvencie. Počet vektorov L obmedzili na 1 200. To znamená, že daný spôsob dokáže reprezentovať len sekvencie s maximálnou dĺžkou 1 200 aminokyselín, pre kratšie sekvencie bola zavedená medzera, ktorá sa doplňuje od poslednej aminokyseliny až po stanovenú dĺžku L.

Dodatočných 57 vlastností, ktoré pri predikcii rozdelili na sekvenčné a štrukturálne. Medzi sekvenčné vlastnosti zaradili: dĺžku sekvencie, molekulárnu hmotnosť, pomer aminokyselín tvoriacich otočku, alifatický index, priemernú hydropatiu a absolútny náboj. Zo štrukturálnych vlastností použili: trojstavovú sekundárnu štruktúru, osemstavovú sekundárnu štruktúru, pomer odhalených aminokyselín a pomer odhalených aminokyselín vynásobený ich hydrofobicitou. Pre výpočet štrukturálnych vlastnosti použili nástroj SCRATCH [35].

Trénovacia sada nástroja DeepSol vychádza z neupravenej dátovej sady vytvorenej autormi nástroja PROSOII [57]. V ďalšom kroku znížili redundanciu tejto sady zhlukovaním na 90 % mieru identity. Na zhlukovanie použili nástroj CD-HIT [19]. Z trénovacej sady taktiež odstránili sekvencie, ktorých miera identity bola aspoň 30 % s niektorou zo sekvencií testovacej sady. Výsledná trénovacia sada obsahovala 28 972 rozpustných a 40 448 nerozpustných proteínov. Na testovanie prediktoru použili dátovú sadu publikovanú v [8], ktorá pozostáva z 1 000 rozpustných a 1 000 nerozpustných proteínov.

Najlepšie výsledky v kontexte korelácie (MCC) a presnosti dosiahla metóda DeepSol S2, s koreláciou 0,55 a presnosťou 77 %, takmer rovnaké výsledky dosiahla verzia S3 s MCC 0,54 a totožnou presnosťou. Metóda S1, ktorá je založená len samotnej aminokyselinovej sekvencii, dosiahla 73 % presnosť a 0,46 MCC.

Kapitola 3

Tvorba dátových sád

Hlavým zdrojom dát pre dátové sady je databáza TargetTrack [9]. Táto kapitola obsahuje detailný popis jednotlivých fáz tvorby dátových sád a problémov, ktoré je nutné v jednotlivých fázach riešiť. Proces tvorby sád od samotného spracovania databázy TargetTrack až po výslednú trénovaciu a testovaciu sadu realizuje skript `create_dataset.sh` umiestnený v priečinku `targettrack_processing`.

3.1 Schéma databázy TargetTrack

Pred samotným spracovaním databázy je potrebná znalosť jej schémy. Táto časť sa zoberá vybranými prvkami XML schémy databázy TargetTrack. Pôvodné anglické názvy prvkov schémy sú uvedené v zátvorke za daným prekladom.

Cieľ

Hlavným prvkom databázy je cieľ (*target*), ktorý reprezentuje biologickú sekvenciu. Každý *target* má jednoznačnú identifikáciu (atribút *id*), ktorá sa skladá z identifikácie strediska a unikátneho identifikátora v rámci strediska, tieto dve časti sú od seba oddelené pomlčkou. *Target*, okrem iného, ďalej obsahuje zoznam sekvencií, zoznam experimentov a typ proteínu.

Typ proteínu (*targetProteinType*) kategorizuje proteíny podľa ich zloženia. Jedná sa o nepovinný prvok s vopred danou množinou hodnôt. Niektoré z možných hodnôt sú: jednodoménový proteín (*single-domain protein*), viacdoménový proteín (*multidomain protein*), oligomerický proteín (*oligomeric protein*) a membránový proteín (*membrane protein*). Dodatočné informácie môžu byť uvedené v poznámke (*remark*).

Zoznam sekvencií (*targetSequenceList*) obsahuje prvky popisujúce jednotlivé sekvencie (*targetSequence*). V tomto zozname sa typicky nachádza len jedna sekvencia. Uvedenie viacerých sekvencií je pomerne zriedkavé a používa sa napr. pri ligandoch alebo v prípade uvedenia jednej sekvencie v rôznych reprezentáciách. Prvok *targetSequence* obsahuje názov sekvencie (*sequenceName*), zdrojový organizmus (*sourceOrganism*), samotnú sekvenciu (*oneLetterCode*), chemický typ sekvencie (*sequenceChemicalType*) a ďalšie. Reprezentácia sekvencie je závislá od chemického typu, ktorý môže nadobúdať hodnoty proteín (*protein*), DNA (*dna*) alebo RNA (*rna*). Najbežnejšia je hodnota proteín, jedná sa o reprezentáciu sekvencie pomocou aminokyselín. Niekedy sa spolu s proteínov reprezentáciou uvádza aj reprezentácia DNA, ktorá pozostáva z nukleotidov.

Experiment

Experiment (*trial*) predstavuje popis postupu pre získanie štruktúry cieľa. Zoznam experimentov (*trialList*), ktorý je umiestnený v cieli, je vždy zložený z aspoň jedného experimentu. Dôležitými prvkami experimentu sú zoznamy: histórie stavov (*statusHistoryList*), protokolov (*trialProtocolList*), sekvencií (*trialSequenceList*) a výstupov (*trialOutcomeList*). V rámci experimentu môže byť uvedená iná sekvencia ako v cieli. Jedným z dôvodov môže byť neočakávaná či zámerná zmena niektorej časti sekvencie.

Zoznam protokolov neobsahuje celé protokoly, jedná sa len o zoznam referencií (*protocolRef*). Na odkazovanie sa používa identifikátor *protocolId*, ktorý je unikátny len v rámci strediska. Pre jednoznačné vyhľadanie pomocou referencie je nutné tento identifikátor spojiť s označením daného strediska. Uvedené protokoly sa viažu ku konkrétnemu experimentu. Postup tvorby a získania štruktúry proteínu sa preto medzi jednotlivými experimentmi cieľa môže líšiť.

Novo pridaným prvkom v schéme je zoznam výstupov. Výstup (*trialOutcome*) obsahuje vo všeobecnosti ľubovoľné informácie, ktoré sa podarilo zistiť z experimentu. Môže to byť napríklad percentuálne vyjadrenie miery rozpustnosti a expresie. Vzhľadom na to, že bol tento prvok zavedený pomerne neskoro, obsahuje ho len malé množstvo experimentov.

Protokoly

Jednotlivé protokoly (*protocol*) sú uvedené v zozname protokolov (*protocolList*). Každý protokol musí obsahovať identifikáciu (*protocolId*), vlastný text (*protocolText*) a typ (*protocolType*). Identifikácia je unikátna len v rámci jedného strediska. Neskôr bol k prvku protokolu zavedený atribút *id*, jeho hodnota sa skladá z názvu strediska a identifikácie v rámci strediska, ktoré sú od seba oddelené pomlčkou. Aj napriek tomu, že je tento atribút nepovinný, uvádza sa pri všetkých protokoloch. Pomocou atribútu *id* je možné jednoznačne určiť daný protokol. Pri odkazovaní z experimentov je však použitý prvok *protocolId*, ktorý je nejednoznačný. Pre prevod na jednoznačný identifikátor je potrebné spojiť označenie centra a *protocolId*. Označenie centra je možné odvodiť z identifikácie cieľa.

Prvok *text* obsahuje textový popis protokolu. Hoci existujú doporučenia týkajúce sa obsahu textu, jeho štruktúra nie je žiadnym spôsobom obmedzená. Vo výsledku sú texty protokolov štruktúrne odlišné, čo výrazne komplikuje získavanie informácií pomocou strojového spracovania.

Postup tvorby proteínu a získania jeho štruktúry je delený do niekoľkých častí. To, ku ktorej časti sa protokol viaže, určuje typ protokolu. Jednotlivé typy majú definované poradie. V tejto práci nás budú zaujímať nasledujúce typy: výber (*selection*), génová syntéza (*gene synthesis*), polymerázová reťazová reakcia (*PCR*, z anglického *polymerase chain reaction*), klonovanie (*cloning*), rast (*growth*) a expresia (*expression*). Typy sú uvedené v poradí, v akom proces tvorby proteínov prebieha. Okrem zmienených typov existuje ešte niekoľko ďalších, ktoré za nimi nasledujú. Z hľadiska rozpustnosti sú podstatné len vyššie uvedené typy protokolov, ostatné sa viažu najmä k získaniu štruktúry.

Stav

Prvok stav (*state*) vyjadruje fázu experimentu. Je typu enumerácie, teda nemôže nadobudnúť ľubovoľné hodnoty. Stav sa v priebehu času mení, pričom stavy dodržujú určitú následnosť. Príklad vývoja stavu v čase: vybraný (*selected*), naklonovaný (*cloned*), exprimovaný (*expressed*), rozpustný (*soluble*), vyčistený (*purified*), vykryštalizovaný (*crystallized*) či

Dôvod (preklad)	**Dôvod (originál)**	**Zastúpenie v %**
zlyhanie expresie	expression failed	30,7
iný dôvod	other	19,7
zlyhanie klonovania	cloning failed	13,6
zlyhanie purifikácie	purification failed	11,7
duplicitný cieľ	duplicated target found	9,2
úspešné získanie štruktúry	structure successful	6,4
slabá difrakcia	poor difraction	4,1
zlyhanie kryštalizácie	crystallization failed	3,3
PDB duplicita	PDB duplicate found	0,9
ďalšie dôvody	-	0,4

Tabuľka 3.1: Percentuálne zastúpenie dôvodov ukončenia na úrovni experimentov v databáze TargetTrack.

práca ukončená (*work stopped*). Špeciálnym stavom je testovací cieľ (*test target*), tento stav majú ciele, ktoré boli do databázy TargetTrack pridané za účelom testovania infraštruktúry databázy, testovacie ciele nemusia obsahovať validné dáta.

V databáze sa uvádza stav na troch miestach. Prvým je stav na úrovni cieľa, druhým stav experimentu, tieto stavy reprezentujú posledný stav, t. j. aktuálny stav cieľa resp. experimentu. Tretím je história stavov, kde sú uvedené všetky stavy, ktoré experiment získal spolu s dátumom ich nadobudnutia (*dateComplete*).

Okrem dosiahnutého stavu môže cieľ a experiment obsahovať dôvod ukončenia (*stopStatus*). Na úrovni cieľa sa nachádza v elemente *targetStopDetails*, v experimente sa jedná o element *stopDetails*. V oboch prípadoch je uvedenie dôvodu ukončenia nepovinné a obsahujú ho len niektoré položky. Z celkového počtu 335 771 cieľov je dôvod ukončenia uvedený v 15,1 % cieľov, z nich je viac ako v 99,9 % dôvodom ukončenia *other* (iný dôvod). Experimenty, ktorých je 961 548, majú dôvod ukončenia v 16,3 %. Zastúpenie rôznych hodnôt dôvodu ukončenia experimentov je zobrazené v tabuľke 3.1.

3.2 Predspracovanie databázy TargetTrack

V tejto fáze dochádza k určeniu rozpustnosti proteínov a výberu podstatných informácií z databázy TargetTrack. Najskôr je každý cieľ rozdelený do niekoľkých častí tak, aby bola zachovaná informácia o rôznom postupe v experimentoch. Informácie z jednotlivých experimentov sú agregované na základe rovnakého postupu. Z jedného cieľa môže teda vzniknúť niekoľko záznamov, kde každý z nich agreguje informácie obecne z niekoľkých experimentov.

Rozpustnosť je určená na úrovni jednotlivých experimentov. Záznamy obsahujú agregovanú hodnotu rozpustnosti z viacerých experimentov. Tá je vyjadrená počtom rozpustných a nerozpustných experimentov v rámci záznamu. Podobne ako v prípade nástroja PROSOII [57], prebiehala klasifikácia rozpustných proteínov na základe dosiahnutých stavov. Experiment je označený ako rozpustný, ak obsahuje v histórii stavov aspoň jeden stav z množiny stavov uvedených v tab. 3.2. Rozdiel od postupu autorov PROSOII je pri určovaní nerozpustných proteínov. V PROSOII za nerozpustné označili tie proteíny, ktorým sa nepodarilo dosiahnuť rozpustný stav počas určitého časového obdobia. V tejto práci bolo zavedené prísnejšie kritérium a za nerozpustné sú považované len experimenty ukončené z dôvodu zlyhania expresie alebo purifikácie.

Množina rozpustných stavov
soluble, purified, crystallized, hsqc, strucure, in pdb, native diffraction-data, NMR assigned, phasing diffraction-data, diffraction, in bmrb, nmr structure, crystal structure, diffraction-quality crystals

Tabuľka 3.2: Množina stavov experimentov databázy TargetTrack, na základe ktorých je možné považovať sekvenciu v experimente za rozpustnú.

Výsledný záznam obsahuje rozpustnosť, sekvenciu v aminokyselinovom formáte, postup (jednotlivé protokoly), názov proteínu, zdrojový organizmus a zoznam laboratórií, ktoré sa na tvorbe podieľali. Skript taktiež extrahuje percentuálnu informáciu o rozpustnosti z detailov experimentu. Tento údaj sa však nachádza len v malom počte položiek, navyše sa vyskytli experimenty s nulovou percentuálnou rozpustnosťou, ktoré zároveň dosiahli niektorý z rozpustných stavov. Na základe vyššie uvedených dôvodov nebol údaj o percentuálnej rozpustnosti použitý pri ďalšom spracovaní a tvorbe sady.

Vo fáze predspracovania boli z databázy TargetTrack odstránené:

- Ciele označené ako transmembránové v type alebo poznámke. Tieto proteíny sú typicky nerozpustné, a preto sa zo sád bežne odstraňujú [24].
- Skúšobné ciele (*test target*), nakoľko neobsahujú validné dáta.
- Ciele, ktoré neobsahovali práve jednu aminokyselinovú sekvenciu.
- Experimenty s aspoň jedným stavom nadobudnutým pred dátum 1. 1. 2006. Podľa [38] dochádzalo pred uvedeným obdobím k veľkému počtu experimentálnych chýb.
- Záznamy, kde nebolo možné odvodiť rozpustnosť.

3.3 Určenie expresného systému

TargetTrack nemá vyhradený špeciálny prvok alebo atribút pre uvedenie expresného systému. Túto informáciu je nutné odvodiť z použitých protokolov, konkrétne v texte daného protokolu. Pri určení boli zvolené dva prístupy. V prvom sa vyhľadávali názvy expresných organizmov v texte protokolov. Neskôr sa ukázalo, že tento spôsob je málo špecifický a dochádzalo k nesprávnej identifikácii. Preto bol vytvorený nový systém založený na hodnote skóre. Tento prístup používa taktiež kľúčové slová, tie ale nepredstavujú názov konkrétneho organizmu, ale skôr použité technológie typické pre daný spôsob expresie. Každé kľúčové slovo obsahuje navyše hodnotu skóre, ktoré určuje mieru špecifickosti kľúčového slova pre daný organizmus. Vzhľadom na povahu textu a nutnosti prehľadu v technológiách tvorby proteínov, boli kľúčové slová vytvorené s pomocou experimentátorov z Loschmidtových laboratórií. Pre experimentátorov bola vytvorená jednoduchá webová stránka, ktorej účelom bolo zjednodušenie procesu prechádzania textov protokolov a priradenia kľúčových slov.

Proces určenia expresného systému je znázornený na obr. 3.1. Vstupom programu sú konfiguračné súbory, texty protokolov a protokoly spolu s početnosťou ich výskytu. Konfiguračné súbory umožňujú, okrem definície kľúčových slov, priamo priradiť expresný systém. Výstupom skriptu je súbor obsahujúci protokoly s priradeným expresným systémom a webová stránka, ktorá umožňuje prechádzať texty protokolov a taktiež zobrazuje aktuálny stav skóre.

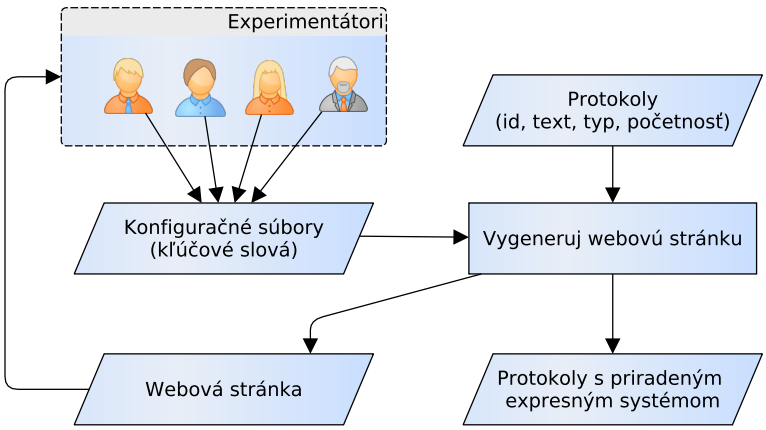
Hlavným prvkom webovej stránky je tabuľka – príklad stránky je na obr. 3.2. V riadku sú jednotlivé protokoly spolu s počtom záznamov, v ktorých sa vyskytujú. Riadky sú zoradené podľa počtu záznamov zostupne. Tabuľka obsahuje dva typy riadkov. Prvý typ pozostáva z informácií viažúcich sa k protokolu, patrí sem početnosť výskytu, celkové skóre, typ a identifikácia protokolu. Pre každý protokol existuje práve jeden riadok prvého typu. Za nimi nasleduje niekoľko riadkov druhého typu, tie pozostávajú z informácií o výskyte kľúčových slov. Druhý typ obsahuje 3 položky: nájdené kľúčové slovo, skoré kľúčového slova a kontext, v ktorom sa kľúčové slovo nachádza. Počet riadkov druhého typu odpovedá počtu nájdených kľúčových slov v protokole. K sebe patriace riadky prvého a druhého sú od ostatných opticky odlíšené podfarbením.

Obecné kľúčové slová, akým je napríklad názov baktérie (*E. coli*) majú priradenú nízku hodnotu skóre a slúžia len na zobrazenie kontextu v tabuľke. Kontext slúži experimentátorom na manuálne priradenie protokolu k expresnému systému, ak experimentátorovi kontext nestačí, môže prejsť na stránku s plným textom protokolu a to pomocou kliknutia na identifikáciu protokolu. Manuálne priradenie prebieha pomocou zaškrtávacieho políčka umiestneného vedľa početnosti výskytu. Experimentátor má možnosť zaškrtnuté protokoly uložiť a taktiež znovu načítať.

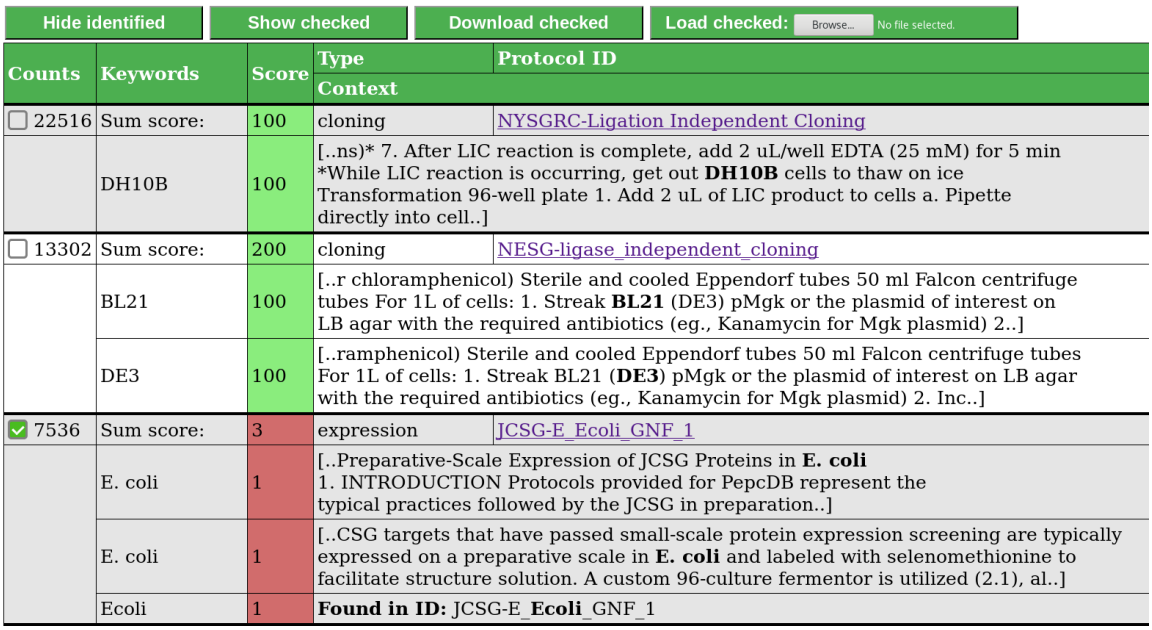
Na základe kľúčových slov a zoznamu zaškrtnutých protokolov poskytnutých od experimentátorov je k protokolu priradený expresný systém. Priradenie expresného systému sa riadni len dvomi podmienkami:

1. Súčet skóre protokolu je vyšší alebo rovný 100.
2. Protokol sa nachádza v súbore zaškrtnutých protokolov.

Ak je aspoň jedna z uvedených podmienok splnená, potom je k protokolu priradený expresný organizmus. V práci boli brané do úvahy len expresné systémy s expresným organizmom baktérie *E. coli*. Ostatné expresné organizmy nemajú v databáze TargetTrack dostatočné zastúpenie, podľa autorov nástroja PROSO [58], je v TargetDB zastúpenie proteínov vytvorených v *E. coli* zhruba 75 %.



Obr. 3.1: Iteratívny proces určenia expresného systému. Experimentátori postupne dopĺňajú obsah konfiguračných súborov na základe údajov sprístupnených webovou stránkou. Výstupom každej iterácie je nová verzia stránky a zoznam protokolov s priradeným expresným systémom.



Obr. 3.2: Stránka s tabuľkou protokolov. Hlavička tabuľky obsahuje počet záznamov (*counts*), kľúčové slová (*keywords*), skóre (*score*), typ (*type*), identifikáciu protokolu (*protocol ID*) a kontext (*context*). Nad tabuľkou sú umiestnené tlačítka na schovanie (*hide*), resp. zobrazenie (*show*) identifikovaných/zaškrtnutých (*hide identified/checked*) riadkov a tlačítka na stiahnutie/načítanie (*download/load checked*) zaškrtnutých protokolov.

3.4 Určenie živočíšnych domén

Živočíšne domény umožňujú základný pohľad na pôvod proteínov a zloženie sady. Domény boli určené na základe taxonomického delenia dostupnom na stránke NCBI[1], konkrétne sa jedná o súbor `rankedlineage.dmp`, ktorý je súčasťou archívu `new_taxdump`. Kategorizácia do domén prebiehala v troch fázach. Prvá fáza predstavuje klasifikovanie na základe zhody vo vedeckom názve, druhá vyžaduje zhodu v rodovom názve a posledná je založená len na zhode druhu. Klasifikácia končí na fáze s prvou úspešnou zhodou. Poradie fáz je dôležité, pretože druhy nie sú na úrovni domén unikátne. To znamená, že jeden druh môže byť zároveň v rôznych doménach. Pre minimalizovanie tohto problému je vhodné postupovať od najviac špecifického názvu (vedecký názov) k obecnému (druh). Týmto postupom je klasifikovaná väčšina organizmov už v prvých dvoch fázach, kde nedochádzalo k žiadnym kolíziám domén. V prípade druhu, došlo k nejednoznačnému zaradeniu do domén v menej ako 30-tich záznamoch, tie boli následne zo sady odstránené. Po všetkých fázach zostalo neroztriedených cca 3 % záznamov, jedná sa o záznamy s neuvedeným, neznámym či chybne zadaným názvom organizmu. Záznamy, ktoré sa nepodarilo klasifikovať boli zo sady odstránené.

[1] https://www.ncbi.nlm.nih.gov/taxonomy

3.5 Tvorba trénovacej a testovacej sady

Základom trénovacej aj testovacej sady je databáza TargetTrack. Testovacia sada bola obmedzená len na proteíny pochádzajúce od centra NESG. Za týmto účelom bola použitá dátová sada prezentovaná v článku [51], v ktorom autori zisťovali vplyv aminokyselinového zloženia na rozpustnosť a expresiu. Proteíny tejto sady majú priamo uvedenú rozpustnosť na škále od 0 do 5. Oproti tomu u zvyšnej časti proteínov databázy TargetTrack nie je uvedená konkrétna hodnota rozpustnosti a je nutné ju nepriamo odvodiť z dostupných záznamov. Preto bola sada NESG zvolená na testovacie účely a zvyšná časť databázy TargetTrack bola použitá len na tvorbu trénovacej sady. Takéto rozdelenie výrazne zvyšuje objektívnosť vyhodnocovania, pretože odstraňuje závislosť medzi zvoleným postupom nepriameho odvodenia rozpustnosti na trénovacej sade a samotným vyhodnotením na sade testovacej.

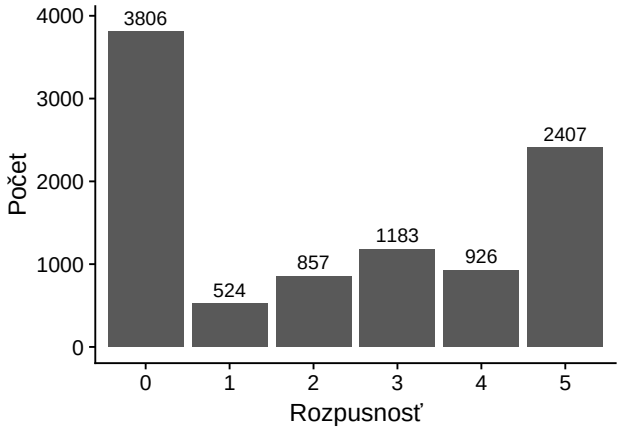
Postup spracovania testovacej a trénovacej sady je znázornený na obr. 3.5. Úvodné fázy spracovania trénovacej sady (*predspracovanie*, *určenie domén* a *expresného systému*) boli podrobne popísané v predchádzajúcich častiach tejto kapitoly. Ďalej nasleduje fáza *odstránenia prekrývajúcich záznamov*, v tejto fáze sa z trénovacej sady odstránia všetky záznamy zo sekvenciami, ktoré sa vyskytujú v testovacej sade. Prienik sa týkal 4 265 sekvencií, jednalo sa prevažne o rozpustné proteíny. Nerozpustné proteíny sa v prieniku takmer nenachádzali. Príčinou je neuvádzanie dôvodu ukončenia centrom NESG v databáze TargetTrack, na základe ktorého sa určuje nerozpustnosť proteínov. Odvodená rozpustnosť sa nezhodovala s rozpustnosťou uvedenou v NESG len v 121 prípadoch. Pri porovnaní bol pre prevod rozpustnosti NESG do binárnej podoby zvolený prah jedna, viď nižšie.

Vo fáze *spájanie záznamov, určenie rozpustnosti sekvencie* dochádza k spájaniu záznamov s rovnakou sekvenciou a určeniu rozpustnosti sekvencie. Za rozpustné sú považované tie sekvencie, ktoré boli rozpustné v aspoň jednom experimente, inak boli označené ako nerozpustné. V prípade testovacej sady (NESG) je ešte predtým upravená hodnota rozpustnosti tak, aby nadobúdala len dve hodnoty (rozpustný/nerozpustný). Keďže je dôležité hlavne to, či je proteín rozpustný alebo nerozpustný bol prah nastavený na hodnotu jedna. Teda sekvencie s nulovým ohodnoteným boli označené za nerozpustné a sekvencie s hodnotou väčšou alebo rovnou jednej za rozpustné. Zmena prahu však nehrá značnú rolu, pretože najviac zastúpenou hodnotou rozpustnosti je 0 a 5, viď histogram na obr. 3.3. V tejto fáze sú ďalej odstránené záznamy, ktoré neobsahujú žiadny experiment s uvedenou rozpustnosťou. Uvedený krok sa týkal len trénovacej sady[2]. V testovacej sade boli ohodnotené všetky záznamy, dodatočne z nej však bolo odstránených 59 záznamov[3]. Jednalo sa o sekvencie s nulovou hodnotou expresie a zároveň nenulovou hodnotou rozpustnosti, tieto sekvencie sú pravdepodobne chybne anotované, keďže meranie rozpustnosti môže prebiehať až po expresii.

Zo sád boli podobne ako v prípade nástroja PROSOII [57] odstránené príliš krátke sekvencie. Minimálna dĺžka sekvencie bola stanovená na 20 aminokyselín. Týmto kritériom prešli takmer všetky sekvencie a vylúčená bola len jedna sekvencia z trénovacej sady. Horný limit pre dĺžku sekvencie je neobmedzený. Ďalej boli odstránené sekvencie, obsahujúce aspoň jednu neznámu aminokyselinu. Sekvencií s neznámou aminokyselinou bolo len 10. Vyššie

[2] Väčšina záznamov s neuvedenou rozpustnosťou je odfiltrovaná už vo fáze *predspracovania*, niektoré z týchto záznamov sa do sady dostali, ak mali uvedenú aspoň percentuálnu rozpustnosť. Tento údaj však nebol pri určený rozpustnosti použitý, viď záverečný odstavec podkapitoly 3.2.

[3] Uvedené číselné údaje sú relatívne k fáze spracovania, v ktorej sa nachádzajú, pretože veľkosť sady sa v priebehu fáz redukuje.



Obr. 3.3: Histogram rozpustnosti testovacej sady (NESG) pred prvou fázou spracovania.

zmienené kroky reprezentujú fázy *odstránenie krátkych sekvencií* a *odstránenie sekvencií s neznámymi aminokyselinami*.

Pomocou nástroja TMHMM [33] sú *odstránené transmembránové proteíny*. Tieto proteíny bývajú spravidla nerozpustné [24], a preto sa bežne pri tvorbe sád prediktorov rozpustnosti odstraňujú. Klasifikácia je závislá od počtu špirál, ak mal proteín viac ako 18 transmembránových špirál, jedná sa pravdepodobne o transmembránový proteín a je zo sa sady vylúčený. Prah bol zvolený podľa doporučenia v článku nástroja TMHMM [33].

Vo fáze *PDB korekcia*, sú vylúčené nerozpustné sekvencie, ktoré sa zároveň nachádzajú v PDB. Použitá bola len podmnožina databázy PDB, pre výber proteínov boli stanovené dva kritériá. Prvým je obmedzenie na rovnaký expresný organizmus ako v dátových sadách (*E. coli*). Druhá podmienka je kladená na metódu získania štruktúry, zvolené boli metódy označené ako *X-RAY* [59] a *SOLUTION NMR* [54]. Tieto metódy vyžadujú pre získanie štruktúry len rozpustné proteíny. V uvedenej podmnožine PDB sa teda nachádzajú iba rozpustné proteíny. Ak je proteín v sade označený ako nerozpustný a nachádza sa vo vybraných proteínoch PDB, potom bol kvôli nesúladu v rozpustnosti vylúčený. Vylučovanie sa týka len identických sekvencií. V prípade veľmi podobných sekvencií nie možné určiť, či sa jedná o chybu v experimente alebo je nerozpustnosť spôsobená zmenou (mutáciou) aminokyselín, prípadne inými faktormi. *PDB korekciou* bolo vylúčených 35 proteínov.

Ďalej sú sady *rozdelené podľa rozpustnosti* na rozpustnú a nerozpustnú sadu. Toto rozdelenie je zavedené kvôli zhlukovaniu. Fáza *zhlukovania* odstráni zo sady podobné sekvencie. Zhlukovanie je vhodné vykonať pre rozpustné a nerozpustné proteíny zvlášť, tým budú v sade zachované podobné proteíny z rôznou rozpustnosťou. Významom zhlukovania je teda odstránenie redundantných sekvencií v rámci jednej triedy. Hlavným parametrom zhlukovania je prah identity. Jedná sa o mieru zhody aminokyselín porovnávaných sekvencií, ktorá je udávaná v percentách. V každom zhluku je práve jedna reprezentatívna sekvencia. Pre reprezentatívne sekvencie a zhluky platia dve základné pravidlá:

1. Reprezentatívne sekvencie zhlukov majú voči sebe identitu nižšiu ako stanovený prah.
2. Sekvencie v rámci zhluku majú voči reprezentatívnej sekvencii identitu väčšiu alebo rovnú ako stanovený prah.

Uvedený postup je typický pre nástroj USEARCH [14]. Pravidlá pre pridávanie sekvencií do zhluku a samotná definícia identity sa medzi rôznymi nástrojmi často líši, princíp však zostáva rovnaký. Napríklad nástroj MMseqs2 [60] používa v pravidlách okrem identity aj tzv. E-hodnotu. Táto hodnota zohľadňuje fakt, že k zhode medzi sekvenciami, resp. ich časťami, môže dôjsť náhodou. Čím nižšia je E-hodnota, tým je zhoda významnejšia.

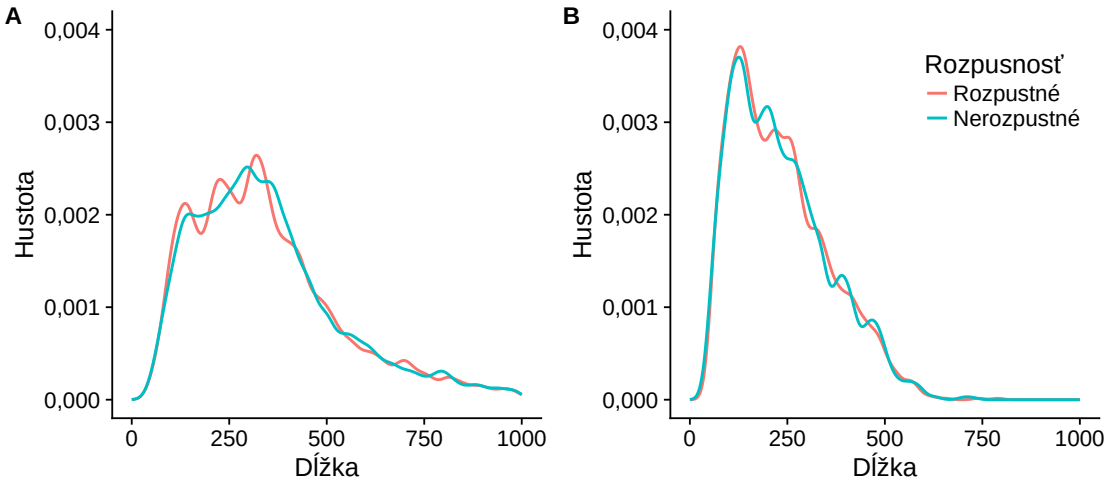
Medzi často používané nástroje na zhlukovanie patrí CD-HIT [19] a USEARCH. Tieto nástroje používajú rôzne heuristiky, ktorých účelom je najmä zvýšenie efektivity procesu zhlukovania, tie však fungujú dobre len v obmedzenom rozsahu prahu identity. V článku nástroja kClust [21] uvádzajú, že oba nástroje sú vhodné pre zhlukovanie, ak je prah nastavený aspoň na 50 % mieru identity. Pri nižších hodnotách prahu klesá rýchlosť a presnosť nástroja CD-HIT. V prípade nástroja USEARCH nezaznamenali vplyv na rýchlosť, ale klesla jeho citlivosť. Na odstránenie redundantných sekvencií sa odporúča použiť 25 % mieru identity [67]. Táto hodnota je výrazne pod prahom 50 %. Preto nie je vhodné použiť vyššie uvedené nástroje CD-HIT a USEARCH. Z tohto dôvodu bol na zhlukovanie vybraný nástroj MMseqs2. Predchodcom tohto nástroja sú nástroje MMseqs [22] a kClust. Jedná sa o pomerne nový nástroj, ktorý dokáže pracovať aj s nižším prahmi identity s dôrazom na zachovanie citlivosti a efektivity zhlukovania. Príklad zhlukovania s využitím nástroja MMseqs2:

```
$ mmseqs easy-cluster sada.fa out tmp --min-seq-id 0.25
```

Bežný postup zhlukovania s použitím MMseqs2 v sebe zahŕňa tvorbu databáze, zhlukovanie a extrahovanie reprezentatívnych sekvencií. Voľba `easy-cluster` umožňuje preskočiť tvorbu databázy a vykonať zhlukovanie s extrahovaním reprezentatívnych sekvencií v rámci jedného kroku. Voľba `easy-cluster` vyžaduje nasledovné údaje: vstupnú sadu (`sada.fa`) vo FASTA formáte[4], reťazec, ktorý bude použitý ako prefix pre výstupné súbory (`out`) a názov priečinku pre uloženie dočasných súborov (`tmp`). Výstupné reprezentatívne sekvencie sú taktiež vo FASTA formáte. Pre uvedený príklad sa budú reprezentatívne sekvencie nachádzať v súbore `out_req_seq.fasta`. Prah identity je zadaný pomocou parametru `--min-seq-id`, hodnota 0,25 značí 25 % prah identity.

Po zhlukovaní je sada rozpustných a nerozpustných proteínov opäť spojená do jednej sady. Táto sada je následne *vyvážená podľa rozpustnosti a dĺžky sekvencie*. Proteíny sú najskôr rozdelené do niekoľkých košov podľa dĺžky. Každý kôš je označený celým nezáporným číslom. Číslo koša je možné získať pomocou výrazu $\lfloor dsek/skos \rfloor$, kde *dsek* reprezentuje dĺžku sekvencie a *skos* šírku koša, ktorá má hodnotu 100. Presná hodnota šírky koša nie je príliš podstatná, vyššie hodnoty však vedú k horšej dĺžkovej vyváženosti. Naopak nízke hodnoty šírky koša zlepšujú vyváženosť, ale môže dôjsť k väčšej redukcii sekvencií, kvôli nedostatočnému zastúpeniu tried v koši. Pre každý kôš je následne získaný počet proteínov v rozpustnej a nerozpustnej triede. Z početnejšej triedy koša sú náhodne odobrané sekvencie tak, aby bol počet medzi triedami rovnaký. Týmto krokom sa stal kôš vyvážený. Podobný postup vyváženia bol použitý aj v článku PROSOII [57]. Zastúpenie jednotlivých dĺžok sekvencií po vyvážení sád znázorňuje obr. 3.4 v podobe grafov hustoty pravdepodobnosti. Z obrázku je možné vidieť, že rozpustné a nerozpustné sekvencie sú v sadách zastúpené rovnomerne a ich krivky sú takmer identické. Z pohľadu dĺžky prevládajú v testovacej sade kratšie sekvencie v porovnaní so sadou trénovacou. Obe sady však majú najväčšie zastúpenie sekvencií v intervale od 0 do 500. Vyváženie je posledným krokom tvorby sady, po ktorom je k dispozícii finálna trénovacia, resp. testovacia sada. Postupné redukovanie počtu sekvencií v závislosti od jednotlivých fáz je zobrazené v tabuľke 3.3.

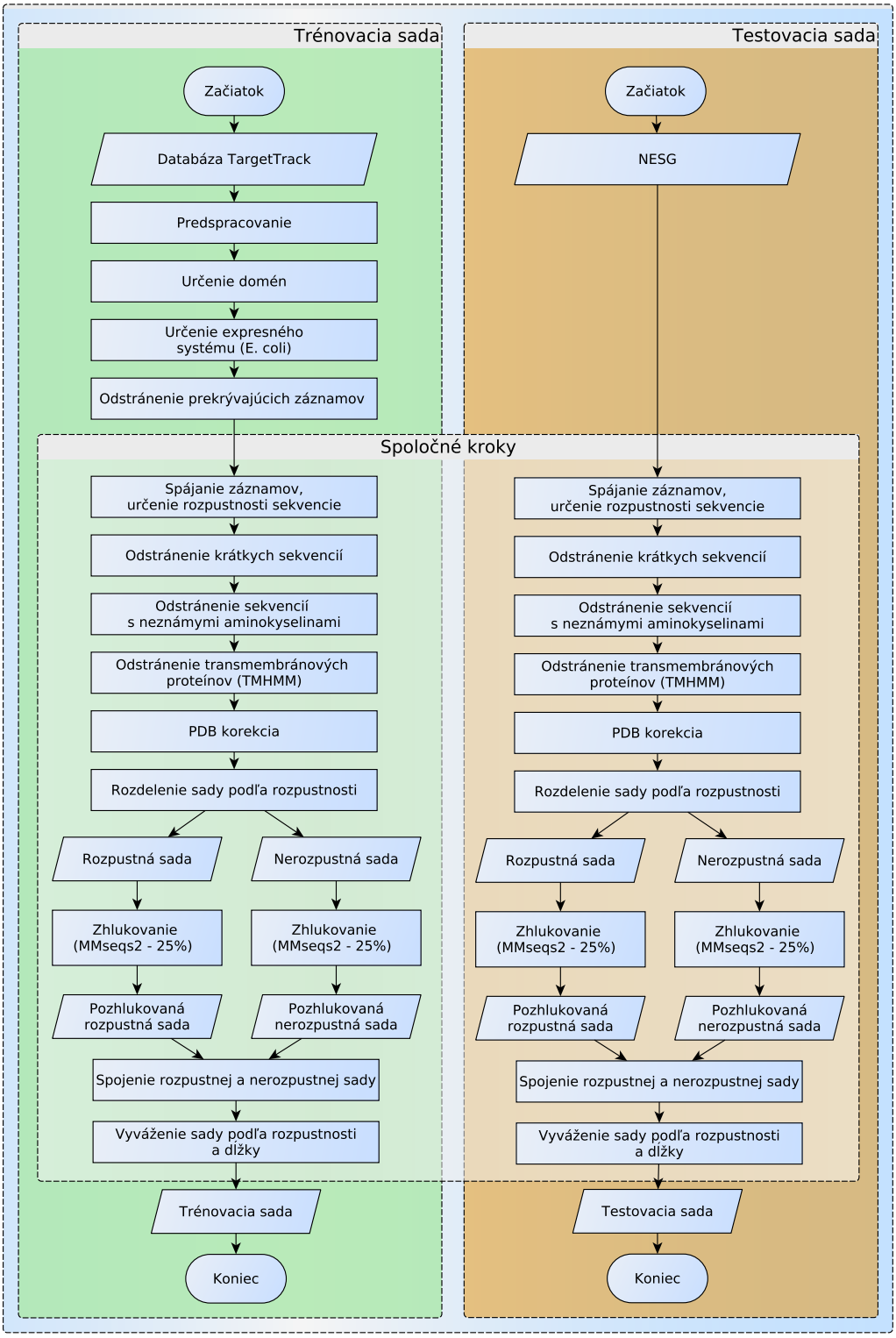
[4] https://zhanglab.ccmb.med.umich.edu/FASTA/



Obr. 3.4: Grafy hustoty pravdepodobnosti dĺžok sekvencií pre trénovaciu sadu (**A**) a testovaciu sadu (**B**). Za účelom ilustrácie vyváženosti podľa rozpustnosti je hustota uvedená zvlášť pre rozpustné a nerozpustné sekvencie. Osa dĺžky bola kvôli prehľadnosti obmedzená na hodnotu 1000. Počet sekvencií s dĺžkou väčšou ako tisíc aminokyselín je 244. Maximálna dĺžka sekvencie v trénovacej sade je 2842 a v testovacej 790 aminokyselín.

Fáza	**Trénovacia sada celkom**	•	○	**Testovacia sada celkom**	•	○
Začiatok	335 771 (C)	-	-	9 703 (Z)	-	-
Predspracovanie	114 905 (Z)	-	-	-	-	-
Určenie domén	111 648 (Z)	-	-	-	-	-
Určenie expresného systému (*E. coli*)	82 362 (Z)	-	-	-	-	-
Odstránenie prekrývajúcich záznamov, určenie rozpustnosti a spájanie záznamov	54 976	40 911	14 065	9 423	5 718	3 705
Odstránenie krátkych sekvencií	54 976	40 911	14 065	9 423	5 718	3 705
Odstránenie sekvencií s neznámymi aminokyselinami	54 969	40 910	14 059	9 420	5 715	3 705
Odstránenie transmembránových proteínov	50 813	38 334	12 479	8 681	5 355	3 326
PDB korekcia	50 793	38 334	12 459	8 666	5 355	3 311
Zhlukovanie	21 648	16 155	5 493	4 401	2 491	1 910
Vyváženie	**10 912**	5 456	5 456	**3 788**	1 894	1 894

Tabuľka 3.3: Postupný vývoj veľkostí dátových sád vzhľadom na fázu spracovania. Vyváženie je poslednou fázou spracovania a odpovedá finálnemu stavu dátových sád. Zátvorka pri hodnote vyjadruje jednotky: (C) počet v cieľoch (*target*), (Z) počet v záznamoch, bez zátvorky počet v sekvenciách. Okrem celkového počtu sekvencií je pri každej sade uvedený počet rozpustných (•) a nerozpustných (○) sekvencií.



Obr. 3.5: Postup tvorby trénovacej (zelená) a testovacej (oranžová) sady. Spoločné kroky sú podfarbené svetlejším pozadím. Text v obdĺžnikoch predstavuje názvy jednotlivých fáz spracovania. Korešpondujúce názvy fáz sú v texte podkapitoly 3.5 vyznačené kurzívou.

Kapitola 4

Experimenty

V tejto kapitole sú prezentované výsledky niekoľkých experimentov. Experimenty súvisia predovšetkým so zisťovaním vplyvu rôznych vlastností a tvorbou experimentálnych modelov. Na vyhodnotenie bol použitý Pearsonov korelačný koeficient (PCC). Vo väčšine prípadov sú predikčné modely založené na metóde regresie, takže sa jedná o koreláciu medzi spojitými hodnotami (predikované hodnoty) a binárnou rozpustnosťou. V prípadoch, kde sú binárne aj predikované hodnoty, je kvôli rozlíšeniu použité označenie MCC (Matthewov korelečný koeficient). Stále sa však jedná o PCC, pretože hodnoty MCC a PCC sú pre koreláciu medzi dvoma binárnymi vektormi ekvivalentné [12]. V uvedených experimentoch, ako aj pri tvorbe finálneho prediktoru bola použitá knižnica `sklearn` [48] v jazyku Python. Parametre modelov strojového učenia boli vo väčšine prípadov získané pomocou mriežkovej metódy (*grid search*) v kombinácii s n-násobnou krížovou validáciou na trénovacej sade.

4.1 K-mery

Najčastejšie používané k-mery sú aminokyseliny (monoméry) a ich dvojice (diméry). Vlastnosťou je ich percentuálne zastúpenie v sekvencii. Takmer všetky nástroje pracujú zo zastúpením monomérov a mnohé z nich používajú aj zastúpenie dimérov, medzi ne patrí napr. SOLpro, SCM, PROSO či PROSOII. Preto sa táto práca taktiež zaoberá vplyvom zastúpenia k-merov na rozpustnosť.

Vyššie uvedené nástroje reprezentujú diméry v podobe 400 dvojíc aminokyselín, z matematického hľadiska sa jedná o permutácie s opakovaním. Okrem tohto spôsobu reprezentácie dimérov bola v tomto experimente použitá aj reprezentácie s využitím kombinácií s opakovaním, čím zároveň klesne počet reprezentačných dvojíc z 400 na 210.

4.1.1 Priamy vplyv k-merov

Najskôr boli vyhodnotené korelácie pre zastúpenia jednotlivých monomérov a dimérov. Výsledky sú v tab. 4.1, riadky sú zoradené podľa absolútnej hodnoty korelácie dosiahnutej na trénovacej sade zostupne. Tabuľka pozostáva z 3 častí, výsledkov pre monoméry, permutácie dimérov a kombinácie dimérov. Z monomérov má najväčší vplyv na rozpustnosť zastúpenie aminokyselín R, K a E. Všetky z uvedených aminokyselín R, K a E sú elektricky nabité, kladne R a K, záporne E. Toto zistenie je v súlade s pozorovaním autorov databázy eSOL a nástrojov PROSO či ESPRESSO. Medzi desiatimi najlepšími aminokyselinami sa ale nevyskytla aminokyselina D, hoci patrí do skupiny nabitých aminokyselín. V súvislosti

Mono	**PCC**$_{TR}$	**PCC**$_{TE}$	**Di**$_P$	**PCC**$_{TR}$	**PCC**$_{TE}$	**Di**$_C$	**PCC**$_{TR}$	**PCC**$_{TE}$
R	-0,166	-0,03	RR	-0,124	-0,021	AR	-0,139	-0,008
E	0,153	0,112	AR	-0,123	0,004	EK	0,13	0,077
K	0,139	0,085	KE	0,122	0,058	RR	-0,124	-0,021
A	-0,127	-0,021	EE	0,119	0,045	DK	0,124	0,028
Q	0,12	0,048	AA	-0,117	-0,034	EE	0,119	0,045
N	0,092	0,024	GR	-0,116	-0,034	AA	-0,117	-0,034
G	-0,087	-0,046	RA	-0,116	-0,017	GR	-0,114	-0,038
Y	0,085	-0,011	AG	-0,101	-0,03	KQ	0,112	0,082
P	-0,065	-0,05	DK	0,101	0,022	AG	-0,111	-0,014
W	-0,052	-0,033	RP	-0,099	-0,031	EN	0,109	0,05

Tabuľka 4.1: Výsledky monomérov (Mono), dimérov s permutáciami (Di$_P$) a dimérov s kombináciami (Di$_C$). Vybraných bolo 10 najlepších korelácií v rámci danej kategórie. Výber sa riadil absolútnou hodnotou korelácie dosiahnutej na trénovacej sade (PCC$_{TR}$). Vedľa PCC$_{TR}$ sa nachádza korelácia testovacej sady (PCC$_{TE}$).

s aminokyselinou D sa taktiež nepotvrdil vplyv diméru DE na rozpustnosť, táto dvojica bola podľa autorov nástroja PROSO najlepšou vlastnosťou.

Korelácie namerané na testovacej a trénovacej sade sa líšia pomerne výrazne. Relatívne malý rozdiel je len medzi koreláciou monomérov E, K a kombináciou dimérov KQ a EK. Ďalej je možné pozorovať, že kombinácie dimérov dosiahli mierne lepšie výsledky ako permutácie.

4.1.2 Modely strojového učenia

Okrem zastúpenia jednotlivých k-merov boli v tomto experimente vyhodnotené skupiny k-merov ako celky, a to pomocou metód strojového učenia. Pri vyhodnocovaní boli použité tri druhy modelov: lineárna regresia, SVM a náhodné lesy. Ďalej bolo vytvorených šesť skupín vlastností: monoméry, permutácie dimérov, kombinácie dimérov, kombinácie trimérov, kombinácie trimérov s kombináciami dimérov, kombinácie dimérov s rozdelením na povrchové a skryté diméry. Vlastnosťami každej skupiny je zastúpenie jednotlivých k-merov. Posledná skupina navyše rozlišuje, či sa diméry nachádzajú na povrchu proteínu a sú prístupne prostrediu, alebo sú skryté vo vnútri proteínu. Prístupnosť aminokyselín bola spočítaná pomocou nástroja RaptorX [49]. Výsledky pre jednotlivé skupiny sú v tab. 4.2. Najlepšie výsledky dosiahli monoméry. Ostatné skupiny majú podobné výsledky, spomedzi nich sú však z výpočtového hľadiska najvhodnejšie kombinácie dimérov.

Pri procese učenia majú monoméry oproti dimérom a trimérom výhodu, ktorá spočíva v častejšom výskyte v sekvencii. Priemerná dĺžka sekvencie v trénovacej sade je 366 aminokyselín. V prípade monomérov, ktorých je len 20, je takmer isté, že sa v priemerne dlhej sekvencii vyskytne každý z nich niekoľkokrát. Oproti tomu, nebudú niektoré diméry a najmä triméry v priemerne dlhej sekvencii nájdené vôbec. Preto môžu modely založené na diméroch, resp. trimérach vyžadovať väčšie trénovacie sady.

Skupina	Najlepšia metóda	Počet vlastností	PCC 5-nás. k. v.	PCC testovacia sada
Mono	SVM	20	0,314	0,147
Di_P	Lineárna regresia	400	0,304	0,091
Di_K	SVM	210	0,294	0,117
Di_{KR}	SVM	420	0,259	0,117
Tri_K	SVM	1 540	0,259	0,121
Tri_K a Di_K	Náhodné lesy	1 750	0,266	0,117

Tabuľka 4.2: Výsledky k-merov, označenie v dolnom indexe typu k-meru určuje bližšiu špecifikáciu: P - permutácie, K – kombinácie, R – prístupnosť k-meru podľa nástroja RaptorX. Pre každú skupinu bola vybraná najlepšia konfigurácia v rámci trénovacej sady.

4.2 Vlastnosti založené na symboloch

Predošlý experiment s k-mermi je v podstate založený na zastúpení symbolov v textovom reťazci, kde symboly predstavujú aminokyseliny a reťazec reprezentuje postupnosť aminokyselín – primárnu štruktúru proteínu. Tento princíp sa dá vo všeobecnosti uplatniť na všetky také vlastnosti proteínov, ktoré je možné reprezentovať sekvenciou znakov. Pri predikcii rozpustnosti je typicky dostupná len aminokyselinová sekvencia, z nej sú následne počítané, resp. predikované ďalšie vlastnosti. Tieto vlastnosti sa preto často viažu k samotným aminokyselinám. Napríklad určujú, či je aminokyselina na povrchu alebo vo vnútri proteínu, jej prítomnosť v agregačné náchylnom, či neusporiadanom regióne, alebo časti sekundárnej štruktúry. Možná reprezentácia týchto vlastností je znázornená v príklade 4.2.1. Význam jednotlivých vlastností (abecied) vyjadrených formou sekvencie znakov je nasledovný:

Sekvencia aminokyselín: označenie podľa štandardu IUPAC.

Agregácia: `A` – časť nachýlia k agregácii, `-` – ostatné časti

Neusporiadanosť: `*` – neusporiadaná časť, `.` – usporiadaná časť

Povrchová prístupnosť: `E` – povrchová časť, `B` – skrytá časť, `M` – zvyšné časti

Sekundárna štruktúra: `H` – α-skrutkovica, `E` – β-skladaný list, `C` – náhodná cievka

Príklad 4.2.1. Rôzne vlastnosti sekvencie reprezentované symbolmi. Vlastnosti s označením R boli vypočítané pomocou nástroja RaptorX [49] a agregácia pomocou nástroja ArchCandy (AC) [3].

```
Sekvencia aminokyselín:     MNQVEMTEFPVGKPDAGGRIWAALETGRLVRLQASPRD
Agregácia (AC):             ------AAAAAAAAAAAAAAAA---------------
Neusporiadanosť (R):        ****.............................***
Povrchová prístupnosť (R):  EEEMEMEEMEBEEEEMMBMBBBMBEEMBBBBBBMEEMB
Sekundárna štruktúra (R):   CCCCCCCHHHHHHHHHHHCCCCCEEEEEECCCCCCCC
```

Takto vyjadrené vlastnosti boli následne skombinované do dvojíc. Označenie dvojice sa skladá z troch častí oddelených pomlčkou. Prvá a druhá časť udáva typ abecedy a posledná je tvorená dvojicou znakov. Význam prvého znaku určuje abeceda uvedená v prvej časti a význam druhého znaku abeceda uvedená v druhej časti názvu. Nech je vlastnosť označená ako `seq_acc_EE`, kde `seq` udáva abecedu aminokyselín a `acc` prístupnosť, potom bude daná vlastnosť predstavovať zastúpenie aminokyseliny E na povrchu proteínu.

Vlastnosť	PCC_{TR}	PCC_{TE}	Prvá časť	PCC_{TR}	PCC_{TE}	Druhá časť	PCC_{TR}	PCC_{TE}
seq_diso_R.	-0,161	-0,043	seq_R	-0,166	-0,03	diso_.	-0,008	-0,029
seq_arch_R-	-0,159	-0,028	seq_R	-0,166	-0,03	arch_-	-0,008	0,014
seq_acc_EE	0,157	0,11	seq_E	0,153	0,112	acc_E	0,055	0,071
seq_arch_E-	0,144	0,108	seq_E	0,153	0,112	arch_-	-0,008	0,014
seq_acc_RM	-0,139	-0,025	seq_R	-0,166	-0,03	acc_M	-0,007	-0,01
seq_acc_KE	0,136	0,092	seq_K	0,139	0,085	acc_E	0,055	0,071
seq_arch_K-	0,132	0,085	seq_K	0,139	0,085	arch_-	-0,008	0,014
seq_acc_RE	-0,124	-0,018	seq_R	-0,166	-0,03	acc_E	0,055	0,071
seq_diso_K.	0,124	0,07	seq_K	0,139	0,085	diso_.	-0,008	-0,029
seq_arch_A-	-0,119	-0,014	seq_A	-0,127	-0,021	arch_-	-0,008	0,014

Tabuľka 4.3: Výsledky vlastností založených na symboloch. Korelácia na trénovacej sade – PCC_{TR}, korelácia na testovacej sade PCC_{TE}, aminokyselinová sekvencia – seq, neusporiadanosť – diso, agregácia – arch, prístupnosť – acc. Vybraných bolo 10 najlepších výsledkov podľa korelácie dvojice symbolov na trénovacej sade. Tabuľka je rozdelená na tri časti, prvá pre koreláciu dvojice a zvyšné dve pre korelácie častí, ktoré ju tvoria.

Taktiež boli vytvorené vlastnosti založené na opakovaní určitého počtu rovnakých znakov bezprostredne za sebou v sekvencii (segmenty). Označenie týchto vlastností sa skladá zo 4 častí: označenie abecedy, znak abecedy, označenie segmentovej vlastnosti a dĺžka segmentu. Napríklad vlastnosť `arch_A_seg_5` reprezentuje zastúpenie segmentov s dĺžkou 5 aminokyselín, ktoré sa zároveň nachádzajú v agregačne náchylnej časti. Skratka `arch` určuje, že sa jedná o agregačnú abecedu a `seg` je označenie segmentovej vlastnosti, za ktorou nasleduje dĺžka segmentu.

Výsledky vlastností založených na symboloch sú uvedené v tab. 4.3. Vybraných bolo 10 vlastností s najvyššou absolútnou hodnotou korelácie na trénovacej sade. Okrem dvojíc vlastností je v tabuľke uvedená aj individuálna korelácia vlastností, ktoré dvojicu tvoria. Je možné pozorovať, že korelácia jednej z častí je v mnohých prípadoch vyššia ako korelácia dvojice. Prínos dvojíc teda nie je významný. Príčinou môže byť nedostatočná presnosť nástrojov, ktorými boli vlastnosti získané. Nástroj RaptorX je však pomerne presný, takže nepresnosti by mali byť minimálne. Získané výsledky preto indikujú, že príliš nezáleží na kontexte, v ktorom sa aminokyselina nachádza, či je na povrchu, v neusporiadanom regióne alebo v agregačne náchylnej časti.

4.3 Sekvenčné vzory

Nasledujúci experiment vychádza z metódy vzorov nástroja ESPRESSO [24]. Metóda vzorov je založená na klasifikácii sekvenčných vzorov na negatívne a pozitívne. Negatívne vzory sú typické pre nerozpustné proteíny a pozitívne pre rozpustné. Okrem metódy ESPRESSO bol na vzoroch natrénovaný a vyhodnotený model SVM, ktorý je prezentovaný v druhej časti experimentu.

4.3.1 Metóda ESPRESSO

Najskôr sú získané všetky vzory z trénovacej sady. Pre každý vzor je uchovaná informácia o počte rozpustných a nerozpustných sekvencií, v ktorých sa vzor našiel. V prípade viacnásobného výskytu vzoru v rámci jednej sekvencie je vzor započítaný len raz. V ďalšom kroku

je ku každému vzoru priradené skóre S pomocou vzorca 4.1,

$$S = \frac{V_p/N_p}{V_p/N_p + V_n/N_n} = \frac{V_p}{V_p + (N_p/N_n)V_n} \tag{4.1}$$

kde V_p a V_n udáva počet výskytov vzoru V v rozpustných, resp. nerozpustných sekvenciách. Výrazy N_p a N_n sú do vzorca 4.1 zavedené za účelom normalizácie. Výraz N_p udáva celkový počet vzorov nájdených v rozpustných sekvenciách trénovacej sady a N_n celkový počet výskytov vzorov v nerozpustných sekvenciách. Hodnotu N_p je možné taktiež vyjadriť pomocou sumy:

$$N_p = \sum_{i=1}^{n} V_{pi} \tag{4.2}$$

Kde V_{pi} predstavuje počet výskytov vzoru V_i v rozpustných sekvenciách a n je počet všetkých vzorov. Podobný postup, ale s použitím nerozpustných výskytov platí pre N_n.

Takto zostavené skóre môže nadobúdať hodnoty od 0 do 1. Hodnoty blízke jednej znamenajú, že vzor je zastúpený viac v rozpustných než nerozpustných sekvenciách. Podobne hodnoty blízke nule predstavujú väčšie zastúpenie v nerozpustných sekvenciách. Ak má vzor rovnaký počet (normalizovaných) výskytov v rozpustných aj nerozpustných sekvenciách, potom bude hodnota skóre 0,5. Klasifikácia vzorov spočíva v nastavení dolného prahu d a horného prahu h. Kde pre d, h platí $0 \leq d < 0{,}5$ a $0{,}5 < h \leq 1$. Vzor je negatívny, ak pre skóre vzoru S platí $S < d$, podobne je vzor klasifikovaný ako pozitívny pre $S > h$.

Samotná predikcia rozpustnosti je založená na vzorci 4.3. Označenie S_R predstavuje skóre rozpustnosti, N_{psek} a N_{nsek} udávajú počet rozpustných, resp. nerozpustných vzorov nájdených v klasifikovanej sekvencii. Výrazy N_{pdat} a N_{ndat} udávajú počet rozpustných, resp. nerozpustných vzorov v rámci celej trénovacej sady.

$$S_R = N_{psek} - \frac{N_{pdat}}{N_{ndat}} N_{nsek} \tag{4.3}$$

Binárna rozpustnosť je odvodená z hodnoty S_R, ak je táto hodnota záporná potom bude sekvencia klasifikovaná ako nerozpustná a pre nezápornú hodnotu S_R ako rozpustná. Rozhodovací prah je teda nastavený na nulovú hodnotu S_R. Nulová hodnota skóre vyjadruje rovnaké zastúpenie pozitívnych a negatívnych vzorov v klasifikovanej sekvencii vzhľadom na pomerné zastúpenie pozitívnych a negatívnych sekvencií v celej sade.

Pred samotnou klasifikáciou je nutné redukovať počet vzorov. Na redukciu boli zvolené dva hlavné kritéria, prvým je nastavenie vhodných prahov a druhým odstránenie vzorov s nízkym výskytom. Za týmto účelom bol vytvorený algoritmus, ktorý postupne vyberá rôzne sady prahov. Vstupom tohto algoritmu je minimálny počet výskytov, horný a dolný prah skoré. Algoritmus najskôr odstráni vzory, ktoré nedosahujú požadovaný minimálny počet výskytov. Ďalej algoritmus postupne znižuje dolný a zvyšuje horný prah tak, aby platili 2 podmienky:

1. Celkový počet vybraných vzorov musí byť nižší ako 10 000.

2. Pomer medzi počtom pozitívnych a negatívnych vzorov nesmie presiahnuť hodnotu 2,5. Toto pravidlo taktiež platí pre pomer negatívnych a pozitívnych vzorov.

Ak nie je splnená prvá podmienka, je horný/spodný prah, zvýšený/znížený o hodnotu 0,0025 (0,25 %). V prípade porušenia druhej podmienky je zvýšený/znížený len prah početnejšej skupiny. Obmedzenie na pomer vzorov zabraňuje výraznej prevahe negatívnych či

Minimálny počet	Horný prah	Dolný prah	Pozitívne vzory	Negatívne vzory	MCC trénovacia sada	MCC testovacia sada
100	0,595	0,405	5 481	4 081	0,318	0,076
	0,648	0,36	610	255	0,265	0,027
300	0,57	0,43	5 233	2 223	0,224	0,069
	0,638	0,395	151	61	0,202	0,029
700	0,55	0,45	5 668	2 447	0,195	0,089
	0,595	0,428	243	98	0,171	0,031
1 000	0,55	0,453	3 019	1 292	0,189	0,094
	0,578	0,44	295	139	0,165	0,078
2 000	0,535	0,465	2 778	1 297	0,183	0,092
	0,553	0,455	408	200	0,169	0,092

Tabuľka 4.4: Výsledky metódy vzorov založenej na ESPRESSO, šedo podfarbené riadky obsahujú aspoň 1 000 vzorov, nepodfarbené menej ako 1 000. Vybrané boli najlepšie konfigurácie v rámci daného minimálneho počtu výskytov. Kritériom výberu bola korelácia dosiahnutá na trénovacej sade.

pozitívnych vzorov. Hodnoty 10 000 a 0,0025 boli zvolené experimentálne. Ak sú splnené obe podmienky, potom budú vybrané vzory použité na predikciu. Po predikcii a uchovaní výsledkov dochádza opäť k zvýšeniu/zníženiu prahu a ďalšiemu vyhodnoteniu. Tento proces sa opakuje až kým nie je počet rozpustných alebo nerozpustných vzorov nulový. Algoritmus bol spustený päťkrát s nasledujúcimi minimálnymi počtami výskytov: 2 000 (18 %), 1 000 (9 %), 700 (6 %), 300 (3 %), 100 (1 %). V zátvorkách je uvedené percentuálne zastúpenie vzhľadom na trénovaciu sadu. Veľkosť trénovacej sady bola 10 912 sekvencií, čo je zároveň maximálny počet výskytov, ktoré môže vzor dosiahnuť. V tabuľke 4.4 sú zhrnuté výsledky pre rôzne kombinácie prahov. Ku každému minimálnemu počtu sú uvedené dva výsledky. Prvý z nich predstavuje najlepšiu konfigurácia, v ktorej bolo zároveň pri predikcii použitých aspoň 1 000 vzorov. Druhý riadok predstavuje najlepšiu konfiguráciu s počtom vzorov menším ako 1 000. Všetky s uvedených konfigurácií vzorov boli vybrané na základe výsledkov dosiahnutých na trénovacej sade.

Najvyššiu hodnotou MCC 0,094 získala konfigurácia s 3 019 pozitívnymi a 1 292 negatívnymi vzormi, z výsledkov uvedených v tab. 4.4 je možné vidieť, že podobnú koreláciu majú aj konfigurácie s výrazne nižším počtom vzorov.

Metóda vzorov implementovaná v tejto práci dosahuje nižšiu koreláciu ako výsledky prezentované v publikácii nástroja ESPRESSO, kde uvádzajú hodnotu korelácie MCC 0,23. Na testovacej sade tejto práce však prediktor ESPRESSO získal hodnotu korelácie MCC len 0,051. Výrazný pokles pri vyhodnotení môže súvisieť s odlišným postupom tvorby sady ESPRESSO a testovacej sady prezentovanej v tejto práci. Dátová sada nástroja ESPRESSO bola zhlukovaná ako celok, bez rozlíšenia rozpustných a nerozpustných proteínov. Tento spôsob zhlukovania predikčný problém zjednodušuje tým, že zo sady môžu byť odstránené podobné sekvencie s rôznou rozpustnosťou. Takéto sekvencie by pravdepodobne zdieľali určitý počet rovnakých vzorov, keďže majú tieto sekvencie rôznu rozpustnosť, zdieľané vzory budú mať malú rozhodovaciu schopnosť. Ďalším faktorom je pravdepodobne nedostatočná veľkosť testovacej sady ESPRESSO. Testovacia sada ESPRESSO pozostáva len zo 178 sekvencií, čo je výrazne menej ako veľkosť testovacej sady vytvorenej v tejto práci, ktorá obsahuje 3 788 sekvencií.

Min. počet	Horný prah	Dolný prah	Počet vzorov	MCC_{SVM} 5-nás. k. v.	MCC_{SVM} testovacia sada	$MCC_{ESPRESSO}$ testovacia sada
100	0,648	0,36	865	0,307	0,02	0,027
300	0,638	0,395	212	0,211	0,038	0,029
700	0,595	0,428	341	0,183	0,052	0,031
1 000	0,578	0,44	434	0,202	0,054	0,078
2 000	0,553	0,455	608	0,188	0,078	0,092

Tabuľka 4.5: Porovnanie metódy vzorov založenej na SVM a ESPRESSO. Sady vzorov boli vybrané na základe tab. 4.4 s obmedzením na maximálne 1 000 vzorov.

4.3.2 Metóda s využitím SVM

Nasledujúca metóda nerozlišuje vzory na pozitívne a negatívne. Odvodenie vzťahu medzi rozpustnosťou a vzormi je ponechané na SVM. Najskôr bolo nutné vzory vhodne reprezentovať. Ku každej sekvencii bol priradený vektor, ktorého dĺžka zodpovedá počtu vzorov. Hodnota jedna v príslušnom prvku poľa značí výskyt daného vzoru v sekvencii a naopak nula vyjadruje, že sa vzor v sekvencii nenachádza. Počet všetkých prvkov matice, ktorá je vstupom SVM bude teda daný súčinom počtu sekvencií a počtu vzorov. Pretože je vzorov príliš veľa, je potrebné ich počet redukovať. Výber sa riadil podľa výsledkov dosiahnutých metódou ESPRESSO, viď tab. 4.4. V ďalšom kroku boli pomocou 5-násobnej krížovej validácie vyhodnotené rôzne parametre SVM. Pre každú sadu vzorov boli vybrané parametre s najlepšou priemernou koreláciou 5-násobnej krížovej validácie na trénovacej sade. Tieto parametre boli použité na natrénovanie finálnych SVM modelov, ktoré boli následne vyhodnotené na testovacej sade. Výsledky metódy SVM sú zobrazené v tab. 4.5. Na základe uvedených výsledkov, je možné konštatovať, že SVM neprinieslo žiadne zlepšenie. Čo indikuje, že vzory získané metódou ESPRESSO majú pomerne malú rozlišovaciu schopnosť medzi rozpustnými a nerozpustnými proteínmi.

4.3.3 Získanie vzorov

Každý vzor má pevnú veľkosť. Podľa výsledkov experimentov autorov ESPRESSO [24] je ideálna veľkosť vzoru 6 až 7 aminokyselín. V tejto práci bola zvolená dĺžka 6, podobne ako v prediktore ESPRESSO. Vzory nie sú tvorené aminokyselinami, namiesto toho sa na mieste aminokyseliny nachádza názov skupiny, do ktorej aminokyselina patrí. Zaradenie aminokyselín do skupín súvisí s ich fyzikálno-chemickými vlastnosťami, jednotlivé skupiny sú uvedené v tab. 4.6. Vzory sú získané pre každú šesticu aminokyselín sekvencie. Sekvencia je prechádzaná od začiatku po koniec s využitím metódy posuvného okna. Základom získavania vzorov je prevod šestice aminokyselín na šesticu názvov skupín. Jedna aminokyselina môže byť zároveň vo viacerých skupinách, viď tab. 4.6. K jednej šestici je teda možné vytvoriť niekoľko vzorov. Počet vzorov je daný počtom všetkých rôznych šestíc názvov skupín, ktoré je možné vytvoriť tak, aby platilo, že i-ta aminokyselina v šestici aminokyselín patrí do skupiny danej i-tým názvom v šestici názvov skupín. Príklad vzorov pre dvojicu aminokyselín PI, by bol nasledovný: (prolín, hydrofóbna), (prolín, alifatická), (malá, hydrofóbna), (malá, alifatická). Pre uvedenú dvojicu existujú 4 vzory. So zvyšujúcim počtom prvkov n-tice však počet možných vzorov stúpa exponenciálne.

Typ aminokyselín	I	L	V	C	A	G	M	F	Y	W	H	K	R	E	Q	D	N	S	T	P
hydrofóbne	●	●	●	●	●	●	●	●	●	●	●	●	○	○	○	○	○	○	●	○
polárne	○	○	○	○	○	○	○	○	○	●	●	●	●	●	●	●	●	●	●	○
malé	○	○	●	●	●	●	○	○	○	○	○	○	○	○	○	●	●	●	●	●
prolín	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	●
drobné	○	○	○	○	●	●	○	○	○	○	○	○	○	○	○	○	○	●	○	○
alifatické	●	●	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
aromatické	○	○	○	○	○	○	○	●	●	●	●	○	○	○	○	○	○	○	○	○
pozitívne	○	○	○	○	○	○	○	○	○	○	●	●	●	○	○	○	○	○	○	○
negatívne	○	○	○	○	○	○	○	○	○	○	○	○	○	●	○	●	○	○	○	○
nabité	○	○	○	○	○	○	○	○	○	○	●	●	●	●	○	●	○	○	○	○

Tabuľka 4.6: Rozdelenie aminokyselín pre metódu vzorov ESPRESSO [24], ● – aminokyselina patrí do danej skupiny, ○ – aminokyselina do skupiny nepatrí. Skupina – typ aminokyseliny.

4.4 Terciárna štruktúra

V tomto experimente bol skúmaný vplyv terciárnej štruktúry na rozpustnosť. Súvislosť medzi rozpustnosťou a typom terciárnej štruktúry taktiež zisťovali autori databázy eSOL [43]. Uvádzajú, že niektoré SCOP vinutia obsahujú výrazný nepomer medzi rozpustnými a nerozpustnými proteínmi.

Na klasifikáciu terciárnych štruktúr proteínov bol použitý nástroj InterPro [17]. InterPro zjednocuje až 14 rôznych databáz. Tieto databázy sa medzi sebou často líšia v spôsobe klasifikácie terciárnej štruktúry a úrovni klasifikácie, ktorou môže byť napr. nadrodina, rodina či doména.

4.4.1 Predikcia pomocou SVM

Terciárna štruktúra bola použitá v kombinácii s modelom SVM. Najskôr bol redukovaný počet možných typov štruktúr (napr. rodín) zavedeným podmienky na minimálny počet výskytov v trénovacej sade. V ďalšom kroku bol k sekvenciám priradený binárny vektor s prvkom pre každú z vybraných štruktúr. Jednotka v binárnom vektore znamená výskyt danej štruktúry v proteíne a naopak prvok vektoru nadobúda hodnotu nula, ak štruktúru neobsahuje. Uvedená reprezentácia je podobná kódu 1 z n, narozdiel od neho sa vo vektore môže vyskytnúť hodnota jedna viackrát, príkladom sú viacdoménové proteíny. Jednotka sa ale nemusí vo vektore vyskytnúť ani raz, to platí pre proteíny, ktorým nebolo možné priradiť štruktúru alebo ich štruktúra nebola vybraná v rámci trénovania.

Pomocou mriežkového vyhľadávania parametrov a 5-násobnej krížovej validácie boli vyhodnotené rôzne varianty SVM. Najlepšie výsledky dosiahli analýzy založené na databázach SUPERFAMILY [44], Gene3D [34] a Pfam [18]. Nástroje SUPERFAMILY a Gene3D klasifikujú štruktúru na úrovni nadrodín a Pfam na úrovni rodiny. Uvedené nástroje sa ďalej líšia v spôsobe klasifikácie, kde SUPERFAMILY vychádza zo SCOP, Gene3D z CATH [45] a Pfam používa vlastnú klasifikáciu. Pfam taktiež rozlišuje vyššiu úroveň klasifikácie v podobe tzv. klanov, do ktorých jednotlivé rodiny patria. Zaradenie rodín do klanov je možné získať na stránke Pfam[1]. Je však dôležité podotknúť, že klany neexistujú pre všetky rodiny,

[1] ftp://ftp.ebi.ac.uk/pub/databases/Pfam

Analýza	Úroveň	PCC 5-násobná k. v.	PCC testovacia sada
SUPERFAMILY	nadrodiny	0,267	0,034
Gene3D	nadrodiny	0,263	0,024
Pfam	klany	0,239	0,038
Pfam	rodiny	0,197	0,008

Tabuľka 4.7: Výsledky metódy založenej na klasifikácii terciárnej štruktúry v kombinácii s SVM. Tabuľka obsahuje priemernú hodnotu 5-násobnej krížovej validácie na trénovacej sade a vyhodnotenie na sade testovacej.

takže sa po ich priradení zvýši počet proteínov bez uvedeného typu štruktúry. Klany sú však obecnejšie a majú väčšiu šancu dosiahnuť minimálny počet. V konečnom dôsledku môže byť teda celkový počet sekvencií s určenou štruktúrou pri trénovaní a testovaní väčší.

Výsledky modelov SVM sú uvedené v tab. 4.7. Pre každú z analýz (klasifikácií) je uvedená priemerná hodnota 5-násobnej krížovej validácie najlepšej konfigurácie vzhľadom na trénovaciu sadu. Pre všetky metódy bol najlepší prah minimálneho počtu 10 sekvencií, okrem neho boli testované prahy pre 20, 30, 40 a 50 sekvencií. Najvhodnejším jadrom pre SVM bola radiálna bázová funkcia, trochu horšie výsledky získalo lineárne jadro a sigmoid.

Najlepšie konfigurácie boli natrénovanie s využitím celej trénovacej sady a následne vyhodnotené na testovacej sade. Ako je možné vidieť z výsledkov tab. 4.7, na testovacej sade je korelácia takmer nulová. Hoci je korelácia malá aj v prípade v prípade 5-násobnej krížovej validácie, nie je zanedbateľná a rozdiel je pomerne výrazný. Po vykonaní experimentu a hľadaní možnej príčiny nízkej korelácie sa ukázalo, že jednotlivé centrá si proteíny snažia rozdeliť tak, aby pochádzali z rôznych rodín [41]. Keďže proteíny testovacej sady pochádzajú len z proteínov vytvorených v rámci jedného centra, prienik medzi štruktúrami vybranými vo fáze trénovania a štruktúrami v testovacej sade nebol príliš veľký. Po natrénovaní nemalo v testovacej sade podľa klasifikácie SUPERFAMILY priradenú nadrodinu 53 % sekvencií, v prípade nadrodín Gene3D 51 %, Pfam rodiny nemalo až 83 % sekvencií a Pfam klany 50 %. V každom z uvedených prípadov nemá pri testovaní priradenú štruktúru aspoň 50 % sekvencií. Aj napriek tejto skutočnosti je prepad korelácie pomerne výrazný. Ďalšou príčinou poklesu môže byť výber sekvencií centrom NESG. Podľa NESG wiki [41] sa výber proteínov riadil výsledkami rôznych nástrojov. Na základe čoho je možné predpokladať, že sú vybrané len proteíny s určitou šancou na úspešnú tvorbu. Ľahko predikovateľné nerozpustné proteíny by sa v takejto sade nevyskytli. Naopak by sa v nej vyskytli nerozpustné proteíny, ktoré mali podľa zvolených vlastností vyššiu šancu na rozpustnosť. Čo môže vo všeobecnosti robiť túto sadu pomerne obtiažnou z hľadiska predikcie rozpustnosti. Posledne zmienený dôvod je len hypotézou, pretože je otázne, či bol tento výber uplatnený aj v prípade proteínov, z ktorých je vytvorená testovacia sada. Uvedený postup mal byť použitý v rámci fázy PSI-2, ktorá prebiehala od roku 2005 do 2010. Proteíny testovacej sady boli vytvorené v období od roku 2001 do 2008, takže táto možnosť nie je vylúčená.

4.4.2 Viacnásobná lineárna regresia

V tomto experimente bol vytvorený model pozostávajúci z mnohých dielčich modelov lineárnej regresie, ktoré boli vytvorené pre každý typ štruktúry. Model využíva len zastúpenie monomérov spolu s tromi fyzikálno-chemickými vlastnosťami, medzi ktoré patrí izoelektrický bod, GRAVY index a priemerná flexibilita. Experiment vznikol na základe informácií od experimentátorov z Loschmidtových laboratórií. Experimentátori pozorovali, že

na niektorých doménach funguje modifikovaný model Wilkinsona a Harrisona (mWH) pomerne dobre. Model mWH je pomerne jednoduchý a využíva len zastúpenie jednotlivých monomérov, bližšie je popísaný v podkapitole 2.3.1.

Pri tvorbe modelu boli najskôr vybrané vhodné štruktúry, kritériom výberu bol minimálny počet výskytov, najlepšie výsledky získal model s obmedzením na 10 výskytov. Pre každú štruktúru, napr. rodinu bolo následne natrénovaných 10 modelov lineárnej regresie. Modely lineárnej regresie boli trénované na proteínoch s rovnakou štruktúrou. Každá štruktúra mala teda vlastnú trénovaciu sadu. Pri tvorbe jednotlivých modelov bolo do sady štruktúry navyše pridaných niekoľko náhodne vybraných sekvencií. Počet náhodne vybraných sekvencií bol trojnásobne väčší ako počet sekvencií v sade štruktúry. Architektúra niekoľkých modelov v kombinácii s náhodným výberom bola zavedená z dôvodu minimalizácie pretrénovania. Pre sekvencie bez priradených štruktúr bol vytvorený osobitný model podobným postupom.

Predikcia spočíva vo vyhodnotení všetkých lineárnych modelov pre všetky štruktúry nájdené v sekvencii a konečným výsledkom je ich priemer. Napríklad pre sekvenciu patriacu do dvoch rodín bude vyhodnotených až 20 modelov (10 pre každú z rodín) a výslednou hodnotou rozpustnosti bude priemer hodnôt všetkých 20 modelov.

Pre porovnanie boli vytvorené dva ďalšie modely, ktoré vychádzali z rovnakej množiny vlastností, ale boli natrénované bez ohľadu na typ štruktúry. Prvým z nich je lineárna regresia a druhým model náhodných lesov. Všetky tri modely boli vyhodnotené pomocou 10-násobnej krížovej validácie. Najlepšie dopadol zložený model s klasifikáciou podľa Gene3D nadrodín[2]. Oproti náhodným lesom a samostatnej lineárnej regresii sa však jednalo len o zanedbateľné zlepšenie. Výsledky 10-násobnej krížovej validácie dopadli nasledovne: zložený model (Gene3D) 0,334, náhodné lesy 0,306, lineárna regresia 0,305. Uvedené hodnoty predstavujú mieru korelácie PCC. Z výsledkov 10-násobnej krížovej validácie je vidieť, že zložená metóda nie je výrazne lepšia ako metódy natrénované na celej sade.

Zložená metóda je však výrazné limitovaná veľkosťou sady, pretože dochádza k učeniu len v rámci jednej nadrodiny, rodiny či klanu, počet sekvencií určených na trénovanie bude pomerne malý. Tento problém bol čiastočne minimalizovaný pridaním náhodných sekvencií do sád štruktúr. Do akej miery ovplyvňuje tento faktor daný model je otázne.

[2]Zvyšné typy klasifikácií dopadli podobne a pohybovali sa v rozmedzí 0,32 až 0,33 PCC.

Kapitola 5

Prediktor rozpustnosti

V nasledujúcej kapitole je popísaná tvorba prediktoru rozpustnosti spolu s jeho vyhodnotením. Výsledný prediktor je ďalej v texte označovaný ako Solpex. Názov je skratkovým označením anglických slov *soluble protein expression*, ktoré v preklade znamenajú *expresia rozpustných proteínov*. Toto označenie bolo zvolené na základe zdrojov dát pre trénovanie a testovanie prediktoru. Tie v sebe okrem rozpustnosti zahŕňajú aj vplyv expresie. Napríklad, ak je expresia nulová, potom nie je vytvorený žiadny proteín a nameraná rozpustnosť bude taktiež nulová.

Solpex bol implementovaný v jazyku Python. Pri návrhu prediktoru Solpex bolo vyskúšaných niekoľko rôznych architektúr a modelov. Najlepšie výsledky dosiahol model založený na náhodných lesoch, ktorý sa stal základom prediktoru Solpex. Na tvorbu modelov strojového učenia bola použitá knižnica `sklearn` [48]. Na výpočty bola ďalej použitá knižnica `biopython` [11], `pandas` [37] a `NumPy`, ktorá je súčasťou ekosystému `SciPy` [27].

5.1 Výber vlastností

V prvom kroku tvorby prediktoru Solpex bol uskutočnený výber vlastností, ktorý prebiehal na trénovacej sade. Výber vlastností realizuje skript `rf_model.py` umiestnený v priečinku `experimental_models`. Skupiny vlastností, na ktorých bol uskutočnený výber spolu s ich počtom (uvedeným v zátvorke) je nasledovný: monoméry (20), kombinácie dimérov (210), fyzikálno-chemické vlastnosti (14), priemerná flexibilita (1) – DynaMine [10], sekundárna štruktúra (3) – FESS [50], priemerná neusporiadanosť (1) – Espritz [66], transmembránové regióny (3) – TMHMM [33] a maximálna identita (1) – USEARCH [14]. Okrem nich bol vyskúšaný aj vplyv vlastností nástrojov RaptorX a ArchCandy z experimentu v podkapitole 4.2. Tie však nejavili žiadny prínos, a preto boli odstránené ešte pred automatickým výberom. Vstupom tejto fázy je teda celkovo 253 vlastností.

V prípade nástroja USEARCH je vlastnosťou maximálna hodnota identity k sekvenciám v databáze PDB. Nejedná sa však o celú databázu PDB, ale len o jej podmnožinu, ktorá je takmer identická s podmnožinou použitou pri korekcii rozpustnosti v podkapitole 3.5. Od podmnožiny použitej pri tvorbe dátových sád sa líši len tým, že z nej boli odstránené všetky sekvencie, ktoré sú zhodné so sekvenciami testovacej sady. Tieto sekvencie boli odstránené preto, aby nedošlo k trénovaniu na sekvenciách testovacej sady.

Medzi fyzikálno-chemické vlastnosti patrí: pomer nabitých aminokyselín, termostabilita, pomer aminokyselín K a R, pomer aminokyselín typických pre α-skrutkovicu, pomer aminokyselín typických pre β-skladaný list, pomer aminokyselín typických pre náhodné

Skupina vlastností (počet)	**Zoznam vlastností skupiny**
Zastúpenie monomérov (12)	A, C, E, G, I, K, L, N, Q, R, S, T
Zastúpenie kombinácii dimérov (12)	AI, AL, DK, DT, EE, EN, EV, GK, IL, IS, LN, LQ
Espritz (1)	priemerná neusporiadanosť
FESS (2)	pomer aminokyselín v β-skladaných listoch, pomer aminokyselín v náhodných cievkach
USEARCH (1)	maximálna identita k PDB
TMHMM (1)	počet aminokyselín v transmembránových špirálach v prvých 60 aminokyselinách (Exp60)
Ostatné (7)	pomer aminokyselín K a R, pomer aminokyselín typických pre α-skrutkovicu, pomer aminokyselín typických pre β-skladaný list, molekulárna hmotnosť, pomer molekulárnej hmotnosti a dĺžky sekvencie, izoelektrický bod, priemerná hodnota flexibility

Tabuľka 5.1: Finálna sada vlastností. Jednotlivé vlastnosti sú v tabuľke organizované do skupín podľa podobného charakteru alebo použitého nástroja. Za názvom skupiny je uvedený počet vlastností danej skupiny.

cievky, molekulárna hmotnosť, dĺžka sekvencie, pomer molekulárnej hmotnosti a dĺžky sekvencie, index nestability, izoelektrický bod, priemerná hodnota flexibility, GRAVY index a aromaticita.

Pri výbere boli najskôr odstránené podobné vlastnosti. Odstraňovanie prebiehalo tak, že zo skupiny vzájomne podobných vlastností bola ponechaná vždy len najlepšia z vlastností. Kritériom podobnosti bola absolútna hodnota Pearsonovej korelácie. Prah bol experimentálne nastavený na hodnotu 0,75. Vlastnosti boli klasifikované ako podobné, ak ich absolútna hodnota korelácie presiahla daný prah. Rozhodnutie, ktoré vlastnosti budú ponechaná a odstránené sa riadilo modelom náhodných lesov (`RandomForestRegressor`). Tento model bol natrénovaný s použitím všetkých vlastností. Následne boli z modelu získané hodnoty skóre. Skóre vyjadruje percentuálny prínos jednotlivých vlastností pre model. Spomedzi podobných vlastností bola ponechaná len vlastnosť s najvyššou hodnotou skóre. Hodnoty skóre je z modelu možné získať pomocou atribútu `feature_importances_`.

Z predošlého kroku bola získaná sada unikátnych vlastností. Ďalej boli odstránené vlastnosti s nízkym prínosom. Na tento účel bol taktiež vytvorený model náhodných lesov. Ponechané boli len vlastnosti, ktorých skóre bolo väčšie alebo rovné stanovenému prahu. V procese 10 násobnej krížovej validácie boli skúmané rôzne prahy skóre. Najlepšie výsledky získali prahy, ktoré sa pohybovali v okolí hodnoty danej výrazom $1/pv$, kde pv predstavuje počet všetkých unikátnych vlastností. Ako bolo uvedené vyššie, skóre vyjadruje percentuálny prínos vlastností, teda ak by mali všetky vlastnosti rovnaký prínos, boli by na základe vyššie uvedeného výrazu vybrané všetky. Keďže ale majú niektoré vlastnosti výrazne väčší vplyv, dochádza na tomto prahu k pomerne veľkej redukcii ich počtu. Týmto procesom bolo vybraných 36 vlastnostní. Vybrané vlastnosti sú spolu s ich zaradením do skupiny a počtom uvedené v tabuľke 5.1.

5.2 Implementácia

Základom prediktoru Solpex je model náhodných lesov (`RandomForestRegessor`), ktorého vstupom je 36 vlastností uvedených v tab. 5.1. Optimalizácia hyper-parametrov modelu prebiehala pomocou 10-násobnej krížovej validácie a mriežkového vyhľadávania. Optimalizované boli nasledovné parametre modelu: minimálny počet pozorovaní pre rozdelenie (`min_samples_split`), minimálny počet pozorovaní v liste (`min_samples_leaf`) a maximálna hĺbka stromu (`max_depth`). Finálne hyper-parametre boli zvolené podľa hodnoty korelácie a presnosti. Okrem nich bol zohľadnený aj vplyv pretrénovania, ktorý je reprezentovaný rozdielom medzi koreláciou, resp. presnosťou, validačnej a trénovacej časti trénovacej sady. Uprednostňované boli všeobecnejšie modely, teda modely s malým rozdielom. Model, ktorý bol vybraný pre prediktor Solpex dosiahol v 10-násobnej krížovej validácii nasledovné (priemerné) hodnoty: korelácia na trénovacej sade 0,448, korelácia na validačnej sade 0,375, presnosť na trénovacej sade 69 % a presnosť na validačnej sade 65,3 %.

Zvolený model bol následne uložený pomocou knižnice `joblib` a nachádza sa v priečinku `data` v súbore `rf_model.pkl`. Nejedná sa však priamo o triedu náhodných lesov, tá je obalená dodatočnou triedou `RandomForestModel`. Táto trieda uchováva okrem modelu aj zoznam vlastností určených pre model spolu s ich poradím a priemerné hodnoty vlastností na trénovacej sade. Priemerná hodnota je použitá v prípade, ak pri predikcii nie je možné získať hodnotu danej vlastnosti. Príkladom je pomer aminokyselín K a R, kde menovateľ R, môže nadobudnúť nulovú hodnotu, hoci tento prípad nie je príliš častý, môže nastať najmä pri predikcii krátkych sekvencií.

Hlavná časť prediktoru je implementovaná v súbore `solpex.py` triedou `Predictor`. Mnohé programy pre výpočet vlastností už existovali, pretože boli vytvorené vo fáze výberu vlastností a experimentovania. Tieto skripty boli v prediktore použité formou modulov. Moduly sú umiestnené v priečinku `feature_scripts`. Jedná sa predovšetkým o skripty, ktoré zaobaľujú spúšťanie dodatočných nástrojov a spracovanie ich výsledkov. Pre výpočet fyzikálno-chemických vlastností bola použitá knižnica `biopython`.

Vstupom prediktoru Solpex je zoznam sekvencií vo FASTA formáte. Výstupom je súbor vo formáte CSV, ktorý obsahuje 3 stĺpce: unikátny identifikátor sekvencie v rámci behu (`runtime_id`), identifikátor korešpondujúci s označením sekvencie vo FASTA súbore (`fa_id`) a predikovaná rozpustnosť (`soluble`). Prediktor taktiež vyžaduje zvolenie priečinku určeného pre výpočty nástrojov tretích strán a uloženie ich výsledkov. Výsledky dodatočných nástrojov sú mapované na identifikátor behu. Príklad spustenia prediktoru Solpex:

```
# python3 solpex.py --i_fa vstup.fasta --o_csv vysledky.csv --tmp_dir /tmp
```

Okrem uvedených volieb je taktiež možné zmeniť predvolenú cestu k modelu a dodatočným nástrojom. Tieto cesty by však mali byť absolútne alebo relatívne na umiestnenie skriptu prediktoru.

5.3 Vyhodnotenie

Vyhodnotenie prediktoru Solpex prebiehalo na testovacej sade vytvorenej v tejto práci. Postup tvorby tejto sady je popísaný v podkapitole 3.5. Za účelom porovnania bolo na testovacej sade vyhodnotených ďalších šesť prediktorov rozpustnosti: DeepSol, ccSOL omics, PROSOII, SOLpro, ESPRESSO a modifikovaný model Wilkinsona a Harrisona (mWH). Mnohé z uvedených nástrojov však neuvádzajú trénovacie sady, v týchto prípadoch by nebolo možné odstrániť prekrývajúce sekvencie z testovacej sady. Preto neboli odstránené

Nástroje	PCC	MCC (NAJ)	MCC (PRED)	MCC (PUB)	Presnosť (NAJ)	Presnosť (PRED)	Presnosť (PUB)	Prah (NAJ)	Prah (PRED)
Solpex	0,261	0,166	0,158	-	0,582	0,579	-	0,53	0,5
PROSOII	0,208	0,153	0,145	0,421	0,572	0,569	0,71	0,61	0,6
DeepSol	0,162	0,108	0,094	0,55	0,541	0,532	0,77	0,443	0,5
ESP. vlastnosti	0,164	0,111	0,098	0,42	0,555	0,548	0,68	0,561	0,5
ESP. vzory	-	-	0,051	0,23	-	0,523	0,63	-	-
SOLpro	0,068	0,047	0,029	0,487	0,523	0,515	0,742	0,376	0,5
ccSOL omics	0,025	0,039	-	-	0,517	-	0,74	0,75	-
mWH	0,128	0,083	0,072	-	0,539	0,535	-	0,561	0,5

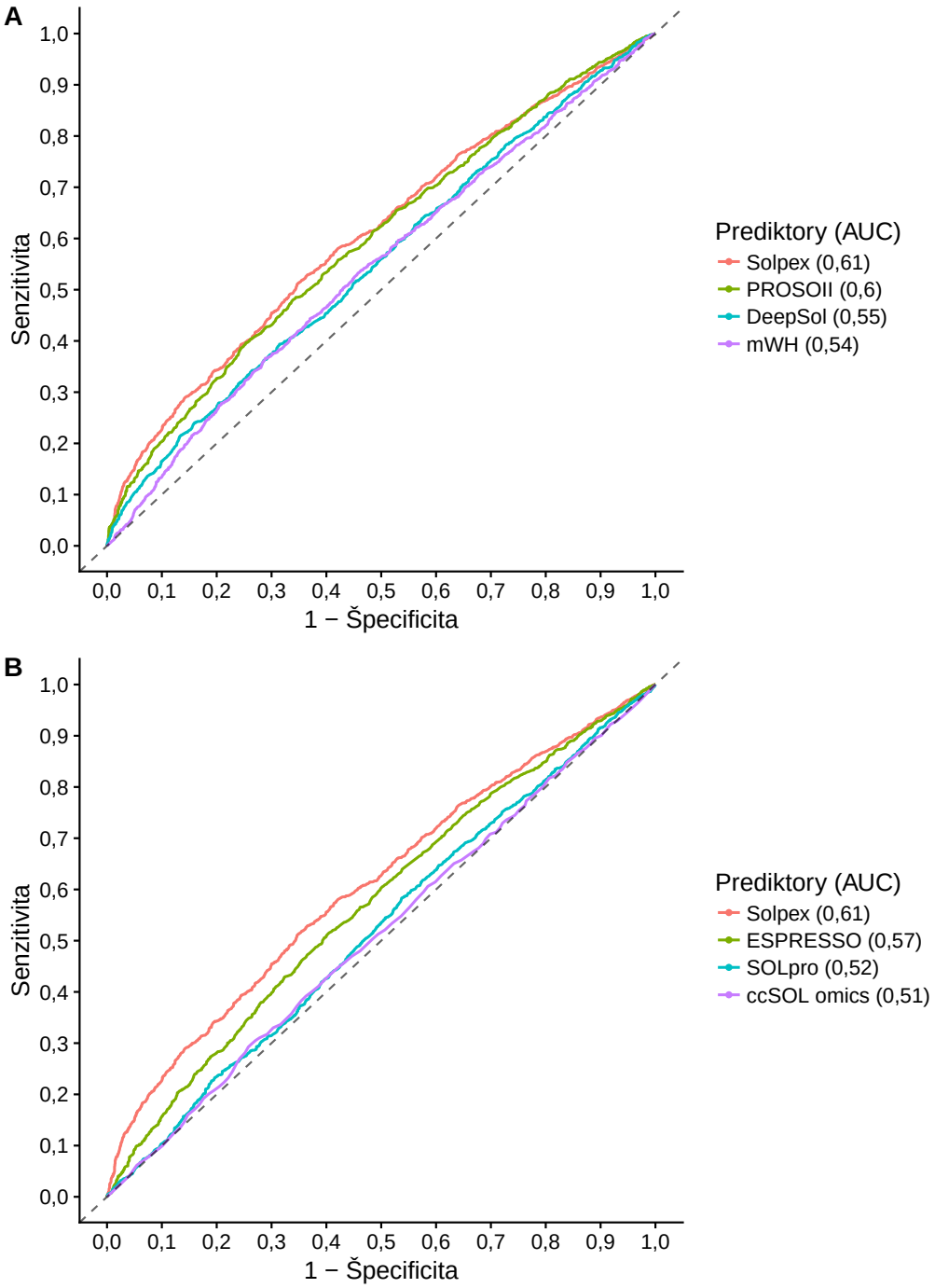
Tabuľka 5.2: Výsledky prediktorov na testovacej sade. NAJ a PRED určujú typ prahu. NAJ – najlepší prah vzhľadom na presnosť, PRED – prah určený autormi nástroja, PUB – výsledky publikované v článku daného nástroja. ESP. vzory – ESPRESSO metóda vzorov. ESP. vlastnosti – ESPRESSO metóda štrukturálnych a sekvenčných vlastností.

prekrývajúce sekvencie pri žiadnom z nástrojov. Trénovacia sada prediktoru Solpex samozrejme žiadne prekrývajúce sekvencie neobsahuje, keďže sa jedná o testovaciu sadu vytvorenú pre tento nástroj. Výsledky jednotlivých nástrojov sú zobrazené v tab. 5.2. Okrem presnosti sa v nej nachádzajú dva typy korelácie, Matthewov korelačný koeficient (MCC) a Pearsonov korelačný koeficient (PCC).

Koeficient MCC bol použitý pre meranie korelácie medzi binárnymi hodnotami rozpustnosti. Niektoré nástroje však neposkytujú binárnu hodnotu rozpustnosti. Preto obsahuje tabuľka niekoľko hodnôt MCC v závislosti na prahu, ktorý bol zvolený pri prevode spojitých hodnôt rozpustnosti na binárne. Ak bola daná hodnota vyššia ako zvolený prah potom bol proteín zaradený do rozpustnej kategórie, inak do nerozpustnej. Hodnoty označené ako PRED vychádzajú z binárnych hodnôt priamo poskytnutých prediktorom, jedná sa teda o predvolený prah daný autormi nástroja. Predvolený prah však nemajú všetky nástroje, preto bol zavedený prah NAJ. Tento prah rozdeľuje predikované hodnoty tak, aby bola dosiahnutá maximálna presnosť. Spôsob získania prahu NAJ nie je z pohľadu vyhodnotenia úplne korektný, pretože bol získaný na testovacej sade. Hodnoty presnosti a korelácie sú teda pre tento prah mierne nadhodnotené. Hodnoty s prahom NAJ sú vo vyhodnotení použité hlavne za účelom porovnania nástrojov a nie ako absolútna miera ich presnosti, resp. korelácie. Podobným postupom ako prah NAJ bola odvodená aj hodnota prahu PRED uvedená v tab. 5.2, s tým rozdielom, že ako referenčná rozpustnosť bola použitá predikovaná hodnota rozpustnosti (binárna). Hodnota prahu PRED je v tabuľke uvedená len za účelom porovnania s prahom NAJ. Tento prah bolo potrebné odvodiť, pretože ho nástroje typicky neuvádzajú, namiesto neho poskytujú hodnoty, ktoré už sú v binárnom tvare. Pre samotné vyhodnotenie však bola použitá binárna rozpustnosť poskytnutá danými nástrojmi a prah PRED nebol použitý v žiadnej časti vyhodnotenia.

Pearsonov koeficient bol použitý v prípade korelácie medzi predikovanými spojitými hodnotami rozpustnosti a rozpustnosťou podľa ohodnotenia NESG. Teda ohodnoteným, v ktorom je rozpustnosť vyjadrená celým číslom v rozmedzí od 0 do 5.

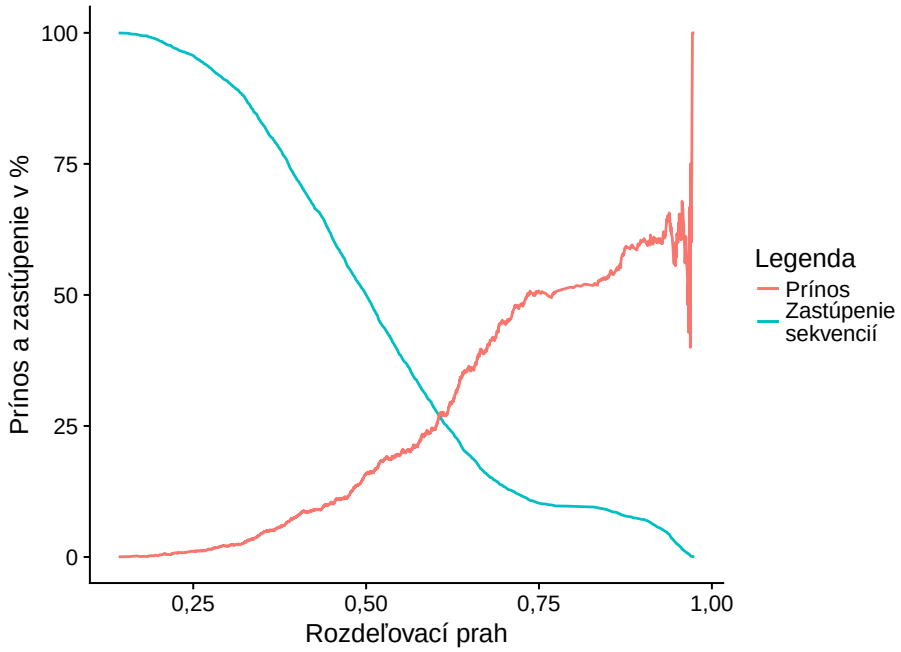
Ďalej boli pre jednotlivé nástroje zostrojené ROC krivky, ktoré sa používajú pomerne často na vyhodnotenie kvality binárnych prediktorov [67]. ROC krivka vyjadruje vzťah medzi špecificitou a senzitivitou v závislosti na rôznych prahoch. S ROC krivkou sa zvykne uvádzať aj plocha pod touto krivkou tzv. AUC, kde hodnota 1 predstavuje perfektný klasifikátor a 0,5 náhodný. Výsledné krivky ROC sú spolu s hodnotou AUC zobrazené na obrázku 5.1.



Obr. 5.1: ROC krivky jednotlivých nástrojov, pre prehľadnosť rozdelené do dvoch grafov. Graf **A** zobrazuje porovnanie nástroja Solpex s nástrojmi PROSOII, DeepSol a modifikovaným modelom Wilkinsona a Harrisona (mWH). V grafe **B** je zobrazené porovnanie nástroja Solpex s nástrojmi ESPRESSO (model vlastností), SOLpro a ccSOL omics. Za každým nástrojom je v legende uvedená plocha pod ROC krivkou (AUC).

Z výsledkov tab. 5.2 a ROC kriviek na obr. 5.1 je vidieť, že podľa uvedených metrík dosahuje najlepšie výsledky prediktor Solpex, za ktorým teste nasleduje nástroj PROSOII. Najväčší prepad oproti publikovaným výsledkom zaznamenal pomerne nedávno publikovaný nástroj DeepSol, ktorý je založený na konvolučných neurónových sieťach. Jedným z dôvodov môže byť jeho pretrénovanie na určité typy rodín. Autori nástroja totiž tvrdia, že sa neurónové siete naučili rozpoznávať aj vyššie úrovne štruktúry, akými sú napríklad vinutia [29]. Testovacia sada vytvorená v tejto práci však pozostáva len z proteínov získaných centrom NESG. Podľa [41] si jednotlivé centrá proteíny rozdelili tak, aby medzi nimi nedochádzalo k prieniku skúmaných rodín. Veľkosť preniku trénovacej sady DeepSolu a testovacej sady použitej v tejto práci činí len 166 proteínov. Pomerne malý prienik indikuje, že mohlo dôjsť k pretrénovaniu na určité typy štruktúr.

Je taktiež dôležité podotknúť, že centrum NESG nevyberalo proteíny náhodne a používalo rôzne predikčné nástroje [41], do akej miery bol takýto výber uplatnený aj v prípade proteínov testovacej sady je otázne. Bližšie je tento problém popísaný v závere podkapitoly 4.4.1. Takýto výber by totiž mohol robiť sadu obtiažnejšou z hľadiska predikcie rozpustnosti. Čo môže byť jedným z dôvodov prudkého poklesu korelácie a presnosti medzi publikovanými hodnotami nástrojov a hodnotami nameranými v rámci tejto práce na testovacej sade. Prediktor Solpex má však tento pokles najnižší. V prípade 10-násobnej krížovej validácie na trénovacej sade dosiahol Solpex presnosť 65,3 % a pri vyhodnotení na testovacej sade 57,9 %, rozdiel je teda 7,4 %. Z tabuľky 5.2 je vidieť, že ostatné nástroje majú oproti publikovaným výsledkom vyšší prepad ako Solpex, napríklad nástroj PROSOII zaznamenal stratu 14,1 % a DeepSol až 23,8 %. Pomerne malý prepad prediktoru Solpex indikuje, že má menšie sklony k pretrénovaniu a je obecnejšie platným modelom.



Obr. 5.2: Graf prínosu (červená krivka) a zastúpenia sekvencií (modrá krivka) vzhľadom na zvolený rozhodovací prah klasifikácie na testovacej sade. Zastúpenie predstavuje pomer sekvencií s hodnotou väčšou ako zvolený prah k celkovému počtu sekvencií. Nestabilné zákmity krivky prínosu v pravej časti grafu boli spôsobené nízkym počtom položiek.

Predikcia rozpustnosti je binárnym klasifikačným problémom. To znamená, že na vyváženej dátovej sade sa bude presnosť náhodnej predikcie blížiť 50 %. Táto hodnota nie je o moc menšia ako hodnota najvyššej presnosti 57,9 %, ktorú dosiahol prediktor Solpex. Hoci je percentuálny zisk pomerne malý, stále môže viesť k efektívnejšiemu vynaloženiu finančných prostriedkov pri výrobe proteínov. K úspore môže dôjsť najmä v prípadoch, v ktorých sú vyhradené zdroje len na tvorbu malej časti z celkového počtu kandidátnych proteínov. Ak má napríklad experimentátor vybrať 100 proteínov z celkového počtu 1000 proteínov, môže nastaviť rozdeľovací prah výrazne vyššie. Závislosť medzi počtom proteínov a rozdeľovacím prahom je znázornená na obr. 5.2. Z obrázku je možné vidieť, že s postupným zvyšovaním prahu sa počet sekvencií znižuje. Na obrázku je ďalej znázornený prínos prediktoru Solpex pre rôzne hodnoty prahu. Prínos predstavuje pomer vyjadrený výrazom 5.1. Kde R_{Solpex} určuje počet rozpustných proteínov s hodnotou predikcie Solpex vyššou ako daný prah a R_{Nah} predstavuje počet rozpustných proteínov pri náhodnom výbere z rovnakého počtu proteínov, akým je počet proteínov s hodnotou predikcie Solpex vyššou ako zvolený prah.

$$\frac{R_{Solpex} - R_{Nah}}{R_{Nah}} \tag{5.1}$$

Podľa obrázku 5.2 dosahuje prediktor Solpex zaujímavé hodnoty prínosu na prahu 0,75, pre ktorý je prínos 50 %. Zastúpenie sekvencií na tomto prahu zodpovedá zhruba 10 % proteínov, čo sa zhoduje s vyššie uvedeným príkladom, kde bolo potrebné vybrať 100 proteínov z 1000. Solpex by v takejto situácii správne vybral 75 rozpustných proteínov, čo je o 50 % viac ako v prípade náhodného výberu, kde by bol počet vybraných rozpustných proteínov len 50. Cena výroby a základnej charakterizácie jedného proteínu je podľa skúseností Loschmidtových laboratórií okolo 20 000 Kč. Aby bolo pri náhodnom výbere získaných 75 rozpustných proteínov, muselo by byť v laboratóriu overených až 150 proteínov. Pri použití prediktoru Solpex stačí pre získanie rovnakého počtu rozpustných proteínov overiť len 100 z nich. Ak je teda cieľom získať 75 rozpustných proteínov, potom prediktor Solpex ušetrí náklady na tvorbu 50-tich proteínov, čo predstavuje milión českých korún.

Kapitola 6

Záver

V úvode práce bolo predstavených niekoľko existujúcich prístupov k predikcii rozpustnosti. Väčšina z uvedených nástrojov používa globálne vlastnosti sekvencie. Medzi nimi sa však vyskytla aj profilovo založená metóda ccSOL omics či nástroj DeepSol, ktorý predstavuje akúsi kombináciu týchto prístupov.

Pomerne veľká časť práce bola vyhradená tvorbe trénovacej a testovacej sady. Hlavným zdrojom dát bola databáza TargetTrack. Samotný postup tvorby sád a odvodenie rozpustnosti z databázy TargetTrack bolo inšpirované mnohými publikáciami a existujúcimi prístupmi. Táto práca sa tieto prístupy pokúsila skombinovať s dôrazom na kvalitu a nie na kvantitu položiek dátovej sady. Preto boli zavedené pomerne prísne kritéria, ktoré položky sady museli spĺňať. Prísnosť kritérií odráža aj výsledný stav trénovacej sady. Po všetkých fázach spracovania obsahovala trénovacia sada len 10 912 sekvencií z pôvodných takmer 300 tisíc sekvencií databázy TargetTrack. Najväčšia strata dochádza pri odvodení rozpustnosti sekvencií a to najmä v prípade nerozpustných proteínov. Hoci je možné predpokladať, že je v databáze TargetTrack viac nerozpustných ako rozpustných proteínov, pri spracovaní sady došlo k opačnému javu a rozpustných sekvencií bolo pred záverečnou fázou vyvažovania takmer trojnásobne viac ako nerozpustných. Tento jav je zapríčinený hlavne nedostatočným uvedeným dôvodu ukončenia v databáze TargetTrack, na základe ktorého sa určuje nerozpustnosť proteínov. Ak by bol ale určený aspoň taký počet nerozpustných proteínov ako rozpustných, potom by bola veľkosť sady takmer trojnásobná oproti veľkosti aktuálnej sady. Preto vzniká otázka, či je možné nerozpustné proteíny odvodiť aj inými spôsobmi. Prípadne je taktiež možné zvážiť použitie existujúcich menej prísnych postupov určenia nerozpustných proteínov. Tým môže byť napríklad postup autorov nástroja PROSO. V takom prípade by však bolo vhodné stanoviť chybovosť týchto prístupov a na základe nej rozhodnúť o vhodnosti ich použitia tak, aby nedošlo k prílišnému zníženiu kvality sady.

Na testovacie účely bola použitá podmnožina databázy TargetTrack, pochádzajúca od centra NESG. Táto podmnožina má oproti zvyšku TargetTrack výhodu v tom, že obsahuje priamo hodnotu rozpustnosti. Samotné vyhodnocovanie teda nie je zaťažené zvoleným postupom odvodenia rozpustnosti, ktorý sa medzi rôznymi sadami existujúcich nástrojov často líši. Ďalšou výhodou testovacej a zároveň aj trénovacej sady je rovnaký expresný systém, ktorým je baktéria *E. coli*. Týmto sú sady vytvorené v tejto práci pomerne jedinečné, nakoľko sady mnohých nástrojov nerozlišujú expresný systém použitý v databáze TargetTrack. Namiesto toho vychádzajú z predpokladu, že väčšina proteínov je vytvorená práve v baktérii *E. coli*. Podľa [58] je baktéria *E. coli* použitá v 75 % prípadov. Teda zhruba štvrtina položiek týchto dátových sád bola vytvorená v inom expresnom organizme.

Druhá časť práce sa venovala najmä experimentom a tvorbe samotného prediktoru. Experimenty sa týkali zastúpeniu k-merov, symbolových vlastností, vzorov a domén. Z výsledkov experimentov vyplynulo, že najvýznamnejšou vlastnosťou je zastúpenie k-merov, z ktorých dosiahli najvyššiu koreláciu nabité monoméry a to najmä kyselina glutámová (E) a lyzín (K).

Po vykonaní experimentov bol zostavený prediktor rozpustnosti Solpex. Novo vytvorený prediktor bol následne porovnaný s vybranými existujúcimi prediktormi na testovacej sade vytvorenej v tejto práci. Spomedzi porovnávaných nástrojov získal Solpex najvyššiu presnosť, koreláciu a plochu pod ROC krivkou. Za prediktorom Solpex nasleduje v tesnom zástupe nástroj PROSOII. Pomerne výrazný prepad v presnosti zaznamenal nástroj DeepSol. V prípade nástroja DeepSol pravdepodobne došlo k pretrénovaniu na abstraktnejších úrovniach štruktúry, akými sú napríklad vinutia. Aj napriek tomu, že prediktor Solpex dosiahol najvyššiu presnosť, jej hodnota je len 57,9 %. Problém predikcie rozpustnosti teda ešte stále nie je vyriešený.

Predikcia prediktoru Solpex je podobne ako väčšina existujúcich nástrojov založená na globálnych vlastnostiach sekvencie. Tento prístup však nezachováva informáciu o poradí medzi aminokyselinami v sekvencii. V prípade vlastností založených na monoméroch je úplne jedno ako budú usporiadané, ak dôjde k zmene ich poradia hodnota globálnej vlastnosti sa nezmení. Pri globálnych prediktoroch je možné očakávať, že budú obecnejšie, ich presnosť sa ale zdá byť limitovaná. Naopak profilovo založená metóda ccSOL omics či kombinovaná metóda DeepSol, ktoré zohľadňujú poradie aminokyselín v sekvencii vykazujú známky silného pretrénovania. Z pohľadu predikcie rozpustnosti to ale vyzerá tak, že bez zakomponovania poradia aminokyselín proteínu sa presnosť prediktorov zvyšovať nebude. Pri týchto metódach však vzniká problém s reprezentáciou sekvencie alebo profilu pevným počtom hodnôt, nakoľko je dĺžka sekvencie premenlivá a strojové učenie vyžaduje konštantný počet vlastností. V prípade nástroja ccSOL omics transformovali profil na Fourierove koeficienty, z ktorých použili prvých 100. Oproti tomu nástroj DeepSol reprezentuje na pevný počet miest už samotnú sekvenciu. Reprezentácia nástroja DeepSol však vyžaduje zavedenie obmedzenia na maximálnu dĺžku sekvencie a tzv. medzery, ktorými sú doplnené kratšie sekvencie. Pomerne zaujímavý je prístup prezentovaný v [70], kde na reprezentáciu sekvencie využívajú algoritmus, ktorý bol vytvorený za účelom prevodu textových dokumentov na vektor desatinných čísiel. Autori v danom článku ukázali, že s využitím tejto reprezentácie je možné predpovedať rôzne proteínové vlastnosti. Druhý problém profilových metód, pretrénovanie, je možné limitovať vhodným rozdelením sekvencií dátových sád tak, aby boli niektoré vinutia či rodiny vylúčené z procesu učenia a ponechané len na testovacie účely.

Literatúra

[1] Agostini, F.; Cirillo, D.; Livi, C. M.; ai.: ccSOL omics: a webserver for solubility prediction of endogenous and heterologous expression in Escherichia coli. *Bioinformatics*, ročník 30, č. 20, Júl 2014: s. 2975–2977, doi:10.1093/bioinformatics/btu420.

[2] Agostini, F.; Vendruscolo, M.; Tartaglia, G. G.: Sequence-Based Prediction of Protein Solubility. *Journal of Molecular Biology*, ročník 421, č. 2-3, August 2012: s. 237–241, doi:10.1016/j.jmb.2011.12.005.

[3] Ahmed, A. B.; Znassi, N.; Château, M.-T.; ai.: A structure-based approach to predict predisposition to amyloidosis. *Alzheimer's & Dementia*, ročník 11, č. 6, Jún 2015: s. 681–690, doi:10.1016/j.jalz.2014.06.007.

[4] Altschul, S.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, ročník 25, č. 17, September 1997: s. 3389–3402, doi:10.1093/nar/25.17.3389.

[5] Berman, H. M.: The Protein Data Bank. *Nucleic Acids Research*, ročník 28, č. 1, Január 2000: s. 235–242, doi:10.1093/nar/28.1.235.

[6] Berman, H. M.; Gabanyi, M. J.; Kouranov, A.; ai.: Protein Structure Initiative - TargetTrack 2000-2017 - all data files [online]. Posledná zmena 5. júl 2017 [cit. 2018-04-08].
Dostupné z: https://doi.org/10.5281/zenodo.821654

[7] Caldwell, G. W.; Ritchie, D. M.; Masucci, J. A.; ai.: The New Pre-Preclinical Paradigm: Compound Optimization in Early and Late Phase Drug Discovery. *Current Topics in Medicinal Chemistry*, ročník 1, č. 5, November 2001: s. 353–366, doi:10.2174/1568026013394949.

[8] Chang, C. C. H.; Song, J.; Tey, B. T.; ai.: Bioinformatics approaches for improved recombinant protein production in Escherichia coli: protein solubility prediction. *Briefings in Bioinformatics*, ročník 15, č. 6, August 2013: s. 953–962, doi:10.1093/bib/bbt057.

[9] Chen, L.; Oughtred, R.; Berman, H. M.; ai.: TargetDB: a target registration database for structural genomics projects. *Bioinformatics*, ročník 20, č. 16, Máj 2004: s. 2860–2862, doi:10.1093/bioinformatics/bth300.

[10] Cilia, E.; Pancsa, R.; Tompa, P.; ai.: From protein sequence to dynamics and disorder with DynaMine. *Nature Communications*, ročník 4, November 2013, doi:10.1038/ncomms3741.

[11] Cock, P. J. A.; Antao, T.; Chang, J. T.; ai.: Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, ročník 25, č. 11, Marec 2009: s. 1422–1423, doi:10.1093/bioinformatics/btp163.

[12] Cross Validated: Relation between the phi, Matthews and Pearson correlation coefficients? [online]. Posledná zmena 19. máj 2013 [cit. 2018-04-13]. Dostupné z: https://stats.stackexchange.com/q/59343

[13] Davis, G. D.; Elisee, C.; Newham, D. M.; ai.: New fusion protein systems designed to give soluble expression inEscherichia coli. *Biotechnology and Bioengineering*, ročník 65, č. 4, November 1999: s. 382–388, doi:10.1002/(sici)1097-0290(19991120)65:4<382::aid-bit2>3.0.co;2-i.

[14] Edgar, R. C.: Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, ročník 26, č. 19, August 2010: s. 2460–2461, doi:10.1093/bioinformatics/btq461.

[15] Edwards, A. M.; Arrowsmith, C. H.; Christendat, D.; ai.: Structural proteomics of an archaeon. *Nature Structural Biology*, ročník 7, č. 10, Október 2000: s. 903–909, doi:10.1038/82823.

[16] EMBL-EBI: What are protein domains? [online] [cit. 2018-04-24]. Dostupné z: https://www.ebi.ac.uk/training/online/course/introduction-protein-classification-ebi/protein-classification/what-are-protein-domains

[17] Finn, R. D.; Attwood, T. K.; Babbitt, P. C.; ai.: InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Research*, ročník 45, č. D1, November 2016: s. D190–D199, doi:10.1093/nar/gkw1107.

[18] Finn, R. D.; Coggill, P.; Eberhardt, R. Y.; ai.: The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, ročník 44, č. D1, December 2015: s. D279–D285, doi:10.1093/nar/gkv1344.

[19] Fu, L.; Niu, B.; Zhu, Z.; ai.: CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, ročník 28, č. 23, Október 2012: s. 3150–3152, doi:10.1093/bioinformatics/bts565.

[20] Hartl, F. U.: Molecular chaperones in cellular protein folding. *Nature*, ročník 381, č. 6583, Jún 1996: s. 571–580, doi:10.1038/381571a0.

[21] Hauser, M.; Mayer, C. E.; Söding, J.: kClust: fast and sensitive clustering of large protein sequence databases. *BMC Bioinformatics*, ročník 14, č. 1, 2013: str. 248, doi:10.1186/1471-2105-14-248.

[22] Hauser, M.; Steinegger, M.; Söding, J.: MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. *Bioinformatics*, ročník 32, č. 9, Január 2016: s. 1323–1330, doi:10.1093/bioinformatics/btw006.

[23] Hirose, S.; Kawamura, Y.; Yokota, K.; ai.: Statistical analysis of features associated with protein expression/solubility in an in vivo Escherichia coli expression system and a wheat germ cell-free expression system. *The Journal of Biochemistry*, ročník 150, č. 1, Apríl 2011: s. 73–81, doi:10.1093/jb/mvr042.

[24] Hirose, S.; Noguchi, T.: ESPRESSO: A system for estimating protein expression and solubility in protein expression systems. *PROTEOMICS*, ročník 13, č. 9, Apríl 2013: s. 1444–1456, doi:10.1002/pmic.201200175.

[25] Ho, S.-Y.; Shu, L.-S.; Chen, J.-H.: Intelligent Evolutionary Algorithms for Large Parameter Optimization Problems. *IEEE Transactions on Evolutionary Computation*, ročník 8, č. 6, December 2004: s. 522–541, doi:10.1109/tevc.2004.835176.

[26] Huang, H. L.; Charoenkwan, P.; Kao, T. F.; ai.: Prediction and analysis of protein solubility using a novel scoring card method with dipeptide composition. *BMC Bioinformatics*, ročník 13 Suppl 17, 2012: str. S3, doi:10.1186/1471-2105-13-S17-S3.

[27] Jones, E.; Oliphant, T.; Peterson, P.; ai.: SciPy: Open source scientific tools for Python [online] [cit. 2018-04-27].
Dostupné z: http://www.scipy.org/

[28] Khan Academy: Orders of protein structure [online] [cit. 2018-04-24].
Dostupné z: https://www.khanacademy.org/science/biology/macromolecules/proteins-and-amino-acids/a/orders-of-protein-structure

[29] Khurana, S.; Rawi, R.; Kunji, K.; ai.: DeepSol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics*, Marec 2018, doi:10.1093/bioinformatics/bty166.

[30] Kitagawa, M.; Ara, T.; Arifuzzaman, M.; ai.: Complete set of ORF clones of Escherichia coli ASKA library (A Complete Set of E. coli K-12 ORF Archive): Unique Resources for Biological Research. *DNA Research*, ročník 12, č. 5, Január 2006: s. 291–299, doi:10.1093/dnares/dsi012.

[31] Kohavi, R.; John, G. H.: Wrappers for feature subset selection. *Artificial Intelligence*, ročník 97, č. 1-2, December 1997: s. 273–324, doi:10.1016/s0004-3702(97)00043-x.

[32] Kramer, R. M.; Shende, V. R.; Motl, N.; ai.: Toward a Molecular Understanding of Protein Solubility: Increased Negative Surface Charge Correlates with Increased Solubility. *Biophysical Journal*, ročník 102, č. 8, Apríl 2012: s. 1907–1915, doi:10.1016/j.bpj.2012.01.060.

[33] Krogh, A.; Larsson, B.; von Heijne, G.; ai.: Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. Cohen. *Journal of Molecular Biology*, ročník 305, č. 3, Január 2001: s. 567–580, doi:10.1006/jmbi.2000.4315.

[34] Lam, S. D.; Dawson, N. L.; Das, S.; ai.: Gene3D: expanding the utility of domain assignments. *Nucleic Acids Research*, ročník 44, č. D1, November 2015: s. D404–D409, doi:10.1093/nar/gkv1231.

[35] Magnan, C. N.; Baldi, P.: SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics*, ročník 30, č. 18, Máj 2014: s. 2592–2597, doi:10.1093/bioinformatics/btu352.

[36] Magnan, C. N.; Randall, A.; Baldi, P.: SOLpro: accurate sequence-based prediction of protein solubility. *Bioinformatics*, ročník 25, č. 17, Jún 2009: s. 2200–2207, doi:10.1093/bioinformatics/btp386.

[37] McKinney, W.: Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, editácia S. van der Walt; J. Millman, 2010, s. 51 – 56.

[38] Mizianty, M. J.; Kurgan, L.: Sequence-based prediction of protein crystallization, purification and production propensity. *Bioinformatics*, ročník 27, č. 13, Jún 2011: s. i24–i33, doi:10.1093/bioinformatics/btr229.

[39] Murzin, A. G.; Brenner, S. E.; Hubbard, T.; ai.: SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, ročník 247, č. 4, Apríl 1995: s. 536–540, doi:10.1016/s0022-2836(05)80134-2.

[40] NESG - NorthEast Structural Genomics consortium: Project overview [online] [cit. 2018-04-8].
Dostupné z: http://www.nesg.org/

[41] NESG Wiki: Target selection [online]. Posledná zmena 10. december 2009 [cit. 2018-04-12].
Dostupné z: http://www.nmr2.buffalo.edu/nesg.wiki/Target_selection

[42] Niwa, T.; Kanamori, T.; Ueda, T.; ai.: Global analysis of chaperone effects using a reconstituted cell-free translation system. *Proceedings of the National Academy of Sciences*, ročník 109, č. 23, Máj 2012: s. 8937–8942, doi:10.1073/pnas.1201380109.

[43] Niwa, T.; Ying, B.-W.; Saito, K.; ai.: Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. *Proceedings of the National Academy of Sciences*, ročník 106, č. 11, Február 2009: s. 4201–4206, doi:10.1073/pnas.0811922106.

[44] Oates, M. E.; Stahlhacke, J.; Vavoulis, D. V.; ai.: The SUPERFAMILY 1.75 database in 2014: a doubling of data. *Nucleic Acids Research*, ročník 43, č. D1, November 2014: s. D227–D233, doi:10.1093/nar/gku1041.

[45] Orengo, C.; Michie, A.; Jones, S.; ai.: CATH – a hierarchic classification of protein domain structures. *Structure*, ročník 5, č. 8, August 1997: s. 1093–1109, doi:10.1016/s0969-2126(97)00260-8.

[46] Parzen, E.: On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, ročník 33, č. 3, September 1962: s. 1065–1076, doi:10.1214/aoms/1177704472.

[47] Pearson, W. R.: Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.*, ročník 132, 2000: s. 185–219.

[48] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; ai.: Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, ročník 12, 2011: s. 2825–2830.

[49] Peng, J.; Xu, J.: Raptorx: Exploiting structure information for protein alignment by statistical inference. *Proteins: Structure, Function, and Bioinformatics*, ročník 79, č. S10, 2011: s. 161–171, doi:10.1002/prot.23175.

[50] Piovesan, D.; Walsh, I.; Minervini, G.; ai.: FELLS: fast estimator of latent local structure. *Bioinformatics*, ročník 33, č. 12, Február 2017: s. 1889–1891, doi:10.1093/bioinformatics/btx085.

[51] Price, W. N.; Handelman, S. K.; Everett, J. K.; ai.: Large-scale experimental studies show unexpected amino acid effects on protein expression and solubility in vivo in E. coli. *Microbial Informatics and Experimentation*, ročník 1, č. 1, 2011: str. 6, doi:10.1186/2042-5783-1-6.

[52] Riès-kautt, M.; Ducruix, A.: [3] Inferences drawn from physicochemical studies of crystallogenesis and precrystalline state. *Methods in Enzymology*, ročník 276, 1997: s. 23–59, doi:10.1016/s0076-6879(97)76049-x.

[53] Rosano, G. L.; Ceccarelli, E. A.: Recombinant protein expression in Escherichia coli: advances and challenges. *Frontiers in Microbiology*, ročník 5, Apríl 2014, doi:10.3389/fmicb.2014.00172.

[54] Ruschak, A. M.; Rose, J. D.; Coughlin, M. P.; ai.: Engineered solubility tag for solution NMR of proteins. *Protein Science*, ročník 22, č. 11, September 2013: s. 1646–1654, doi:10.1002/pro.2337.

[55] Sastry, A.; Monk, J.; Tegel, H.; ai.: Machine learning in computational biology to accelerate high-throughput protein expression. *Bioinformatics*, ročník 33, č. 16, Apríl 2017: s. 2487–2495, doi:10.1093/bioinformatics/btx207.

[56] Shimizu, Y.; Inoue, A.; Tomari, Y.; ai.: Cell-free translation reconstituted with purified components. *Nature Biotechnology*, ročník 19, č. 8, August 2001: s. 751–755, doi:10.1038/90802.

[57] Smialowski, P.; Doose, G.; Torkler, P.; ai.: PROSO II - a new method for protein solubility prediction. *FEBS Journal*, ročník 279, č. 12, Máj 2012: s. 2192–2200, doi:10.1111/j.1742-4658.2012.08603.x.

[58] Smialowski, P.; Martin-Galiano, A. J.; Mikolajka, A.; ai.: Protein solubility: sequence based prediction and experimental verification. *Bioinformatics*, ročník 23, č. 19, December 2006: s. 2536–2542, doi:10.1093/bioinformatics/btl623.

[59] Smyth, M. S.; Martin, J. H.: x ray crystallography. *MP, Mol. Pathol.*, ročník 53, č. 1, Február 2000: s. 8–14.

[60] Steinegger, M.; Söding, J.: MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, ročník 35, č. 11, Október 2017: s. 1026–1028, doi:10.1038/nbt.3988.

[61] The UniProt Consortium: UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, ročník 45, č. D1, November 2016: s. D158–D169, doi:10.1093/nar/gkw1099.

[62] Trainor, K.; Broom, A.; Meiering, E. M.: Exploring the relationships between protein sequence, structure and solubility. *Current Opinion in Structural Biology*, ročník 42, Február 2017: s. 136–146, doi:10.1016/j.sbi.2017.01.004.

[63] Uhlen, M.; Fagerberg, L.; Hallstrom, B. M.; ai.: Tissue-based map of the human proteome. *Science*, ročník 347, č. 6220, Január 2015: s. 1260419–1260419, doi:10.1126/science.1260419.

[64] Uhlen, M.; Oksvold, P.; Fagerberg, L.; ai.: Towards a knowledge-based Human Protein Atlas. *Nature Biotechnology*, ročník 28, č. 12, December 2010: s. 1248–1250, doi:10.1038/nbt1210-1248.

[65] Vendruscolo, M.; Knowles, T. P. J.; Dobson, C. M.: Protein Solubility and Protein Homeostasis: A Generic View of Protein Misfolding Disorders. *Cold Spring Harbor Perspectives in Biology*, ročník 3, č. 12, August 2011: s. a010454–a010454, doi:10.1101/cshperspect.a010454.

[66] Walsh, I.; Martin, A. J. M.; Domenico, T. D.; ai.: ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics*, ročník 28, č. 4, December 2011: s. 503–509, doi:10.1093/bioinformatics/btr682.

[67] Walsh, I.; Pollastri, G.; Tosatto, S. C. E.: Correct machine learning on protein sequences: a peer-reviewing perspective. *Briefings in Bioinformatics*, ročník 17, č. 5, September 2015: s. 831–840, doi:10.1093/bib/bbv082.

[68] Wilkinson, D. L.; Harrison, R. G.: Predicting the Solubility of Recombinant Proteins in Escherichia coli. *Nature Biotechnology*, ročník 9, č. 5, Máj 1991: s. 443–448, doi:10.1038/nbt0591-443.

[69] Xiao, R.; Anderson, S.; Aramini, J.; ai.: The high-throughput protein sample production platform of the Northeast Structural Genomics Consortium. *Journal of Structural Biology*, ročník 172, č. 1, Október 2010: s. 21–33, doi:10.1016/j.jsb.2010.07.011.

[70] Yang, K. K.; Wu, Z.; Bedbrook, C. N.; ai.: Learned protein embeddings for machine learning. *Bioinformatics*, Marec 2018, doi:10.1093/bioinformatics/bty178.

Príloha A

Obsah priloženého DVD

Na priloženom DVD sa nachádzajú nasledovné súbory a zložky:

thesis/ text bakalárskej práce a jeho zdrojové kódy

- **xmarus07-predikce-rozpustnosti-proteinu.pdf** text bakalárskej práce formát PDF
- **src/** zdrojové kódy textu bakalárskej práce vo formáte LaTeX

solpex/ zdrojové kódy prediktoru

- **solpex.py** prediktor rozpustnosti Solpex
- **data/** dáta pre prediktor
  - **rf_model.pkl** model náhodných lesov
  - **Ecoli_xray_nmr_pdb_no_nesg.fa** podmnožina databázy PDB bez NESG
- **targettrack_processing/** spracovanie databázy TargetTrack a tvorba dátových sád
  - **create_dataset.sh** skript na tvorbu dátových sád – všetky fázy tvorby
  - **scripts/** skripty jednotlivých fáz tvorby dátových sád
    - **html/** skripty pre tvorbu webovej stránky expresného systému
  - **data/** dáta pre tvorbu dátových sád
    - **nesg/** sada NESG [51]
    - **targettrack/** databáza TargetTrack [9]
    - **taxonomy/** taxonomické delenie
    - **pdb_ecoli/** podmnožina databázy PDB
    - **protocols/** konfiguračné súbory pre webovú stránku expresného systému
    - **pre_calculated/** predpočítané dáta pre tvorbu dátových sád
- **datasets/** trénovacia a testovacia dátová sada
- **feature_scripts/** skripty na výpočet vlastností, spúšťanie a spracovanie výsledkov programov tretích strán
- **features/** predpočítané vlastnosti
- **seq_id_map.csv** mapovanie predpočítaných vlastností na aminokyselinovú sekvenciu
- **solubility_predictors_wrappers/** skripty pre automatické zaslanie sekvencií a spracovanie výsledkov webových serverov nástrojov PROSOII, ccSOL omics a ESPRESSO
- **experimental_models/** experimentálne modely k-merov, terciárnej štruktúry a vzorov. Priečinok ďalej obsahuje skript **rf_model.py**, ktorý slúži na výber vlastností prediktoru a tvorbu výsledného modelu.
- **manual.md** popis spustenia prediktoru a tvorby dátových sád