VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV INFORMAČNÍCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF INFORMATION SYSTEMS

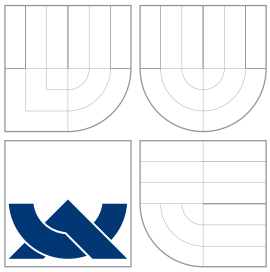
STROJOVÉ UČENÍ V ÚLOZE PREDIKCE VLIVU NUKLEOTIDOVÉHO POLYMORFISMU

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

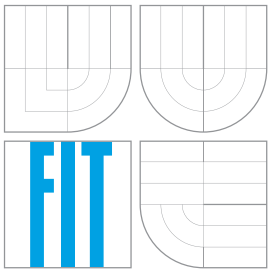
AUTOR PRÁCE
AUTHOR

Bc. ONDŘEJ ŠALANDA

BRNO 2015



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV INFORMAČNÍCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF INFORMATION SYSTEMS

STROJOVÉ UČENÍ V ÚLOZE PREDIKCE VLIVU NUKLEOTIDOVÉHO POLYMORFISMU

PREDICTION OF THE EFFECT OF NUCLEOTIDE SUBSTITUTION USING MACHINE LEARNING

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR

Bc. ONDŘEJ ŠALANDA

VEDOUCÍ PRÁCE
SUPERVISOR

Ing. JAROSLAV BENDL

BRNO 2015

Abstrakt

Tato práce prezentuje nový přístup k predikci efektu nukleotidového polymorfismu v lidském genomu. Cílem je vytvoření nového klasifikátoru, který kombinuje výsledky již existujících softwarových nástrojů. Tohoto konsenzu nad dílčími výsledky je dosaženo experimentováním s metodami strojového učení, přičemž výsledný model pak tvoří nejúspěšnější z nich. Závěrečné komplexní srovnání výsledků metaklasifikátoru s dílčími nástroji ukazuje průměrné navýšení obsahu plochy pod ROC křivkou o 3,4 a eskalaci normované přesnosti až o 7 %. Vytvořený prediktor je zpřístupněn prostřednictvím webového rozhraní na adrese http://ll06.sci.muni.cz:6232/snpeffect/.

Abstract

This thesis brings a new approach to the prediction of the effect of nucleotide polymorphism on human genome. The main goal is to create a new meta-classifier, which combines predictions of several already implemented software classifiers. The novelty of developed tool lies in using machine learning methods to find consensus over those tools, that would enhance accuracy and versatility of prediction. Final experiments show, that compared to the best integrated tool, the meta-classifier increases the area under ROC curve by 3,4 in average and normalized accuracy is improved by up to 7 %. The new classifying service is available at http://ll06.sci.muni.cz:6232/snpeffect/.

Klíčová slova

Deoxyribonukleová kyselina, protein, mutace, polymorfismus, klasifikace, trénovací dataset, strojové učení, konsenzuální predikce

Keywords

Deoxyribonucleic acid, protein, mutation, polymorphism, classification, training dataset, machine learning, ensemble learning

Citace

Ondřej Šalanda: Strojové učení v úloze predikce vlivu
nukleotidového polymorfismu, diplomová práce, Brno, FIT VUT v Brně, 2015

Strojové učení v úloze predikce vlivu nukleotidového polymorfismu

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně pod vedením pana Ing. Jaroslava Bendla

.....
Ondřej Šalanda
26. května 2015

Poděkování

Vřele děkuji svému vedoucímu panu ing. Jaroslavu Bendlovi za vedení a konzultace při vypracovávání této diplomové práce. Díky jeho vstřícnému přístupu a odborné pomoci jsem byl schopen vyřešit mnohé problémy. Děkuji také organizaci MetaCentrum za poskytnutí přístupu k distribuované výpočetní infrastruktuře nezbytné pro provedení dostatečně velkého množství experimentů. V neposlední řadě děkuji panu Bc. Janu Štouračovi za pomoc s instalací virtuálního serveru a nasazením webové aplikace.

Obsah

Kapitola 1

Úvod

Predikce vlivu mutací v DNA patří mezi hlavní odvětví bioinformatického studia. Rozmach nových metod sekvenování genomu v posledních letech měl za následek identifikování velkého množství nových mutací různých druhů. Efekt těchto mutací je však v naprosté většině neznámý a spolehlivě jej lze stanovit pouze laboratorní cestou. Tento přístup je ovšem velmi nákladný a zdlouhavý, proto byla vyvinuta řada softwarových nástrojů, které dokáží tento efekt s různou přesností predikovat.

Předmětem této práce je mimo studia existujících metod návrh a implementace nového metaklasifikátoru kombinujícího výsledky již existujících nástrojů za účelem zvýšení přesnosti a robustnosti predikce. K nalezení vhodné konsenzuální funkce bude experimentováno se zástupci všech hlavních metod strojového učení, z nichž bude vybrána ta nejúspěšnější. Klasifikátor by měl zvládat klasifikace mutací všech druhů a na celém genomu. Jedná se v tomto odvětví o ojedinělý projekt, neboť většina existujících nástrojů je omezena na některý typ mutace, případně na specifický genomický region.

Teoretická část práce je rozdělena do sedmi kapitol. Druhá kapitola se zabývá stavbou a složením DNA, která je nosičem genetické informace. Ve třetí kapitole jsou popsány procesy vzniku proteinů podle genů v DNA a také popis kompozice lidského genomu. Ve čtvrté kapitole jsou popsány mutace v nukleotidovém řetězci a možné příčiny jejich vzniku. Také jsou zde zmíněny některé mechanismy, které naopak vzniku mutací předcházejí. Následující pátá kapitola je věnována existujícím přístupům k predikci efektu mutací a možnostem jejich srovnání. Z prostudovaných nástrojů jsou v šesté kapitole uvedeny ty, které splnily požadavky a jejich výsledky budou vstupem implementovaného konsenzuálního metanástroje, v sedmé kapitole jsou poté rozebrány dostupné zdroje mutací v lidské DNA s důrazem na možnosti vytvoření nového nezávislého datasetu. Osmá kapitola je věnována strojovému učení. Jsou zde uvedeny typické úlohy, k jejichž řešení je tento koncept použit a různé metody, kterými je možné řešení realizovat.

Poslední tři kapitoly práce jsou věnovány vývoji nového klasifikátoru. Devátá kapitola popisuje celý proces návrhu a implementace nového prediktoru. Je zde popsán výběr nástrojů, zdrojů mutací a konstrukce trénovacích datasetů. Následuje popis trénovacího procesu a implementace webového rozhraní. V desáté kapitole je provedena analýza výsledků a vlastností nového metanástroje. Důraz je kladen zejména na množství statistických metrik, kterými lze nástroje hodnotit a na jejich základě porovnat dílčí klasifikátory s novým metanástrojem. V závěrečné části je poté uvedeno krátké shrnutí výsledků a dosažených cílů a také diskuse dalších možných vylepšení pro budoucí navazující práci.

Kapitola 2

DNA a její vlastnosti

Deoxyribonukleová kyselina, zkráceně DNA, je klíčová sloučenina pro život organismů. Je v ní totiž uchována genetická informace, která pak tvoří soubor genů, tzv. genom. Geny pak prostřednictvím své exprese vytváří proteiny, které definují strukturu těla organismu, jeho fungování a chování. Studium DNA a jejích vlastností je proto zcela stěžejní záležitostí a výzkum v tomto směru je prováděn s velkou intenzitou [3, 35].

2.1 Chemické složení

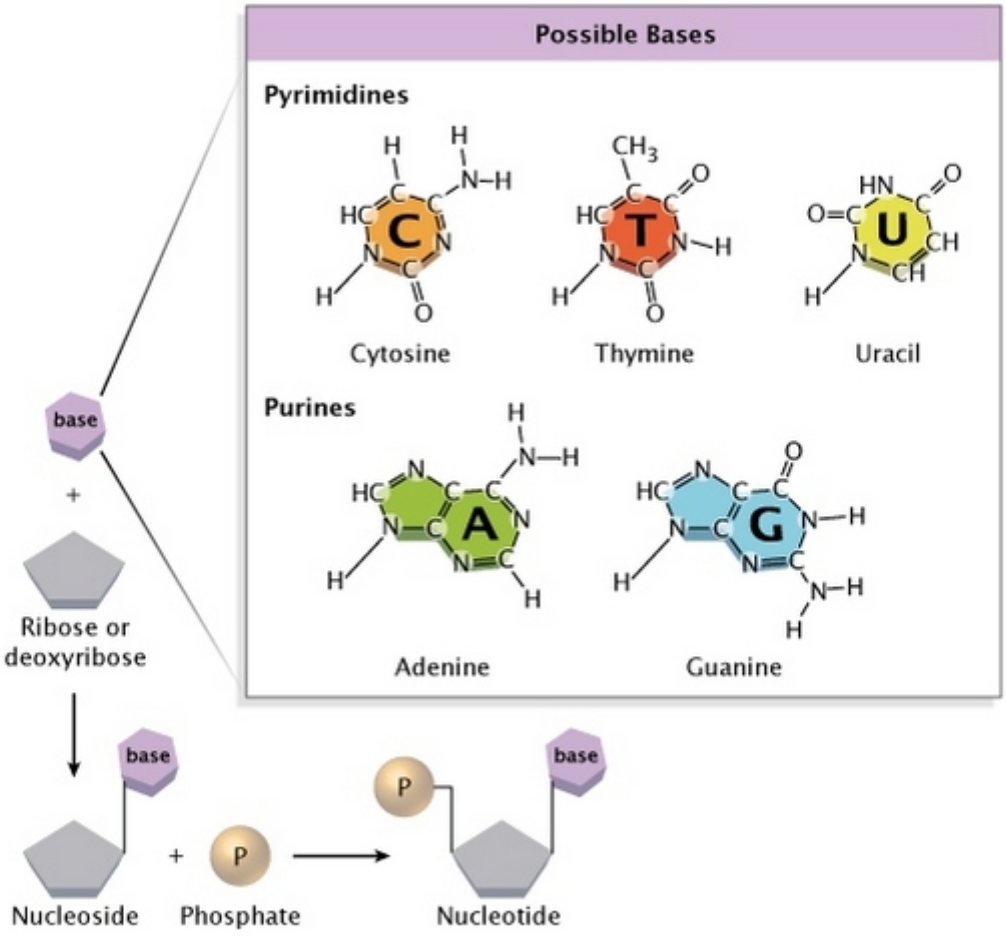
Chemickou podstatou molekuly DNA je spojování konkrétních menších molekul do řetězce. V řetězci se na sebe váží vždy sacharid, fosfát a dusíkatá báze. Tyto tři sloučeniny tvoří takzvaný nukleotid, jeden článek řetězce. Zatímco fosfátová skupina a sacharid jsou u všech nukleotidů stejné (jedná se vždy o deoxyribózu a fosfátovou skupinu HPO_4^{2-}), dusíkaté báze rozlišujeme celkem čtyři. Jsou to adenin (zkratka A), thymin (T), cytozin (C) a guanin (G). Báze jsou si svou chemickou stavbou vzájemně podobné, jak je vidět na obrázku 2.1, ne však natolik, aby byly rozdíly zanedbatelné. V této práci se zaměřuji právě na rozdíly nukleotidů, respektive, jaký vliv by měla na daném místě záměna, vložení či odebrání jednoho a více nukleotidů z řetězce. Zmíněné dusíkaté báze dělíme do dvou skupin podle jejich vlastností, jsou to:

- puriny, do kterých patří adenin a guanin,
- pyrimidiny, mezi které řadíme thymin, uracil a cytozin [38, 3].

Uracil je báze, která je řazena rovněž k pyrimidinům, ovšem vyskytuje se pouze u molekul RNA (ribonukleová kyselina), a to místo thyminu. V této práci se zabývám pouze DNA, bázi uracil tedy nebudu uvažovat [35].

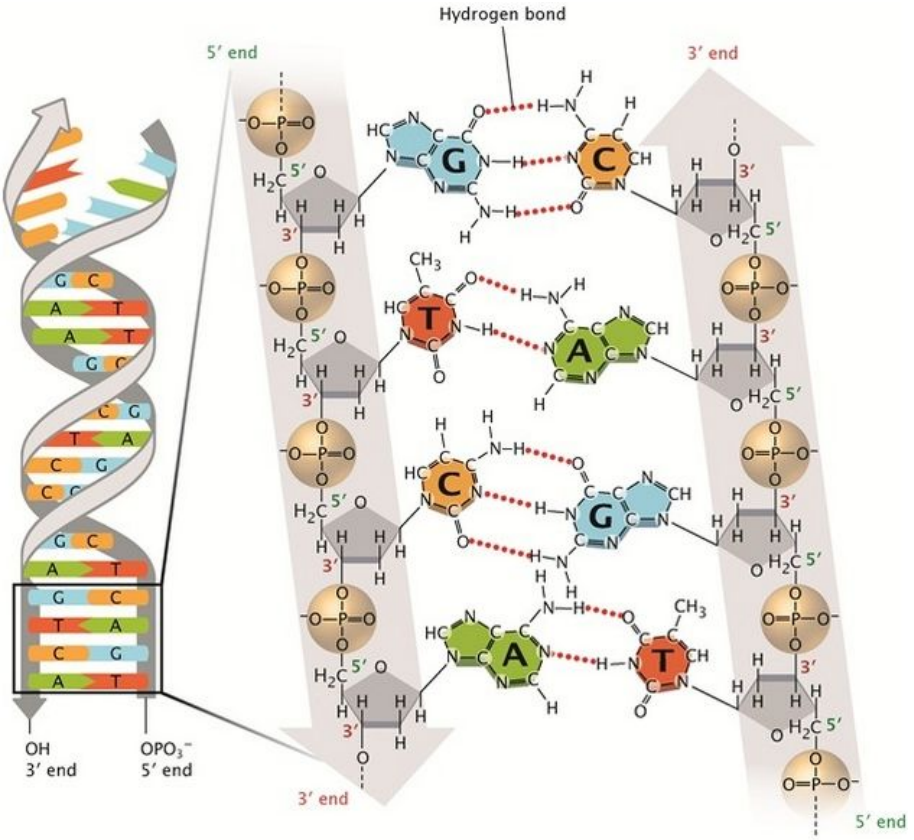
V 50. letech 20. století vědci James Watson a Francis Crick publikovali článek [55], ve kterém odhalili strukturu molekuly DNA. To byl průlomový okamžik genetického výzkumu, neboť na něj navazovaly další výzkumy, např. přenos genetické informace, mutace DNA a podobně. Důležitost objevu lze doložit i tím, že za něj byla zmíněným vědcům roku 1962 udělena Nobelova cena.

Na obrázku 2.2 v jeho pravé části je zobrazena část řetězce DNA, tedy několik nukleotidů včetně jejich částí. Je vidět, že molekula DNA se skládá ze dvou nukleotidových vláken, které jsou přes dusíkaté báze propojeny vodíkovými vazbami a stočeny do typické dvoušroubovice. V této souvislosti je nutné zmínit, že spojení dusíkatých bází je přesně definováno systémem komplementarity. Spárován je vždy purin s pyrimidinem, a to tak, že adenin je vždy spojen s thyminem a cytozin vždy s guaninem. Jedním z faktů, které



Obrázek 2.1: Chemické složení dusíkatých bází [38].

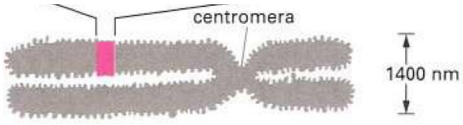
podporují tuto tezi, je skutečnost, že počty pyrimidinů a purinů si v molekule DNA vždy odpovídají [38]. Vazba A-T je evidentně slabší, neboť je tvořena pouze dvěma vodíkovými můstky, zatímco vazba C-G třemi. Dále je z obrázku patrné, že kostru vlákna nukleotidů tvoří střídající se sacharid a fosfátová skupina. Dusíkatá báze je připojena spíše do prostoru tak, aby bylo umožněno její spojení s komplementárním vláknem. Řetězec sacharidů a fosfátů je proto nazýván cukr-fosfátovou kostrou vlákna DNA, její stabilitu zajišťují kovalentní vazby. Důsledkem tohoto spojení se na jednom konci vlákna vždy nachází volný sacharid (označení 5') a na druhém fosfát (označení 3') [3, 38]. Toto pojmenování je zavedeno zejména proto, aby byl jednoznačně určen směr syntézy DNA (od 5' k 3').



Obrázek 2.2: Dvoušroubovice DNA a ukázka párování bází [38].

2.2 Lidský genom

Tato práce se zabývá mutacemi v lidském genomu. Jedná se o kompletní genetickou informaci člověka, která je celá obsažena v každé lidské buňce, resp. v jejím jádře. Délka lidského genomu je přibližně 3,3 miliardy párů bází ($3{,}3 \cdot 10^9$ bp), což tvoří šroubovici dlouhou přibližně jeden metr. K uložení této informace do jedné buňky, jejíž rozměry se pohybují v řádu desítek mikrometrů, musí být toto vlákno ze své dvoušroubovicové struktury efektivně sbaleno. V lidských buňkách dojde postupným balením k tomu, že DNA zaujme tvar podle obrázku 2.3. Takto vzniklý útvar se nazývá chromozom a lidskou DNA v každém jádře buňky tvoří 23 párů těchto chromozomů. Chromozomové páry jsou až na poslední označeny svými pořadovými čísly od 1 do 22. Poslední pár je u žen tvořen dvěma chromozomy X, muži mají jeden chromozom X a druhý Y [3, 9, 38].



Obrázek 2.3: Ukázka sbalené DNA do chromozomu [35].

Informace o délce genomu lze získat jeho sekvenováním. Je to proces zjištění přesné posloupnosti nukleotidů (A, C, G, T) vlákna DNA. To může být provedeno řadou různých technik, z nichž nejznámější je dnes již překonaná Sangerova metoda z roku 1977. V současné době se používají výrazně přesnější a rychlejší next-generation a third-generation metody [3].

První genom člověka byl nasekvenován v roce 2003. Od té doby se počet osekvenovaných lidských genomů značně rozrostl, což umožnilo specializaci výzkumu na genovou variabilitu a její fenotypové projevy. Pro jakési sjednocení náhledu se používá tzv. referenční genom. Jedná se o obecný model lidského genomu, který je vytvářen a aktualizován konsorciem GRC (The Genome Reference Consortium). Tento model se mění na základě nových informací z nasekvenování nových lidských jedinců. Aktuálně používané jsou referenční genomy s označením GRCh37 (použit i v této práci) a GRCh38 [11, 9].

2.2.1 The Genome Reference Consortium

GRC je mezinárodní organizace, která spravuje, a dohlíží na aktualizace referenčního genomu člověka a některých modelových organismů (myš, dánio pruhované[1]). Tyto referenční genomy zahrnují jednak lineární reprezentaci chromozomů, dále nasekvenované sekvence, které se zatím nepodařilo do genomu umístit a také alternativní sekvence pro některé regiony, které jsou pro lineární reprezentaci příliš komplexní. Aktualizace referenčních genomů se provádí buď formou malých oprav (*patches*), hlavní aktualizace pak znamená nové vydání genomu (*releases*) [11].

2.3 Vazba DNA na protein

Genetická informace v DNA se navenek projevuje vznikem proteinu podle předlohy, takzvaného genu. Proteiny jsou základními stavebními látkami lidského organismu, v němž plní řadu buněčných funkcí (enzymatické, regulační, transportní, pohybové, zásobní a další). Aby mohl protein vzniknout, musí dojít k takzvané genové expresi, kdy se proteinu odpovídající gen, tedy sekvence nukleotidů DNA, nakopíruje a následně přeloží. Je tedy třeba, aby DNA byla schopná předávat svou informaci dále. Přenos genetické informace je reprezentován třemi základními ději: replikací, transkripcí a translací. Replikace se uplatňuje při dělení buněk, kdy nově vzniklá buňka musí obsahovat opět celou lidskou DNA, totožnou s mateřskou buňkou [40]. Pro zajištění přesnosti replikace existuje řada buněčných mechanismů a enzymů, které kontrolují, zda při replikaci nedošlo k chybě. Při detekci špatně nakopírovaného nukleotidu tyto opravné mechanismy vynucují reparaci [12].

[1] Dánio pruhované (*zebrafish*) je sladkovodní kaprovitá ryba s charakteristickým pruhovaným zbarvením.

Kapitola 3

Vznik proteinů

Pro důkladnější pochopení mutací DNA je nutná detailní znalost toho, jak se z genetické informace stane protein. Tento proces probíhá ve dvou fázích, transkripci a translaci.

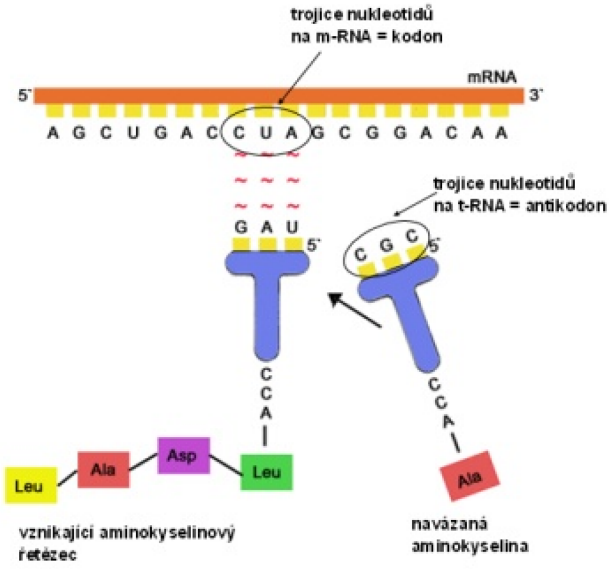
3.1 Transkripce

Transkripcí neboli přepisem genu vzniká vlákno mRNA (mediátorová RNA), komplementární k nukleotidové sekvenci genu v DNA. Je to proces, na kterém se opět podílí enzymy, konkrétně RNA-polymeráza. Pro pochopení procesu je nutné zmínit, že gen, tedy sekvence nukleotidů, k němuž vzniká komplementární mRNA vlákno, je na DNA ohraničen. Před ním se v určené vzdálenosti nachází promotor, což je posloupnost nukleotidů typická pro všechny eukaryotické buňky. Promotor se netranskribuje, slouží pouze k tomu, aby na něj nasedl enzym RNA-polymeráza, která vytváří nové vlákno mRNA. Dále pak se na DNA ještě mimo gen a promotor nachází regulační oblasti, které umožňují nasednout dalším podpůrným proteinům (enzymům), které se nazývají transkripční faktory. Dokud nejsou navázány všechny nutné transkripční faktory i RNA-polymeráza, transkripce nemůže začít a protein z genu nevznikne. Tím je umožněno buňce regulovat koncentraci proteinů, které vytváří [13, 3].

Výsledkem transkripce je primární transkript, což je nukleotidové vlákno mRNA, které je komplementární k sekvenci nukleotidů genu. Tento transkript se ale ještě skládá z protein nekódujících části, takzvaných intronů, a kódujících částí - exonů. Zároveň platí, že celková délka intronů je řádově větší než délka exonů. Pro extrahování pouze exonové sekvence dochází k takzvanému sestřihu. Hranice intronu a exonu se vyznačují typickou posloupností nukleotidů, která je identifikována specifickými enzymy. Tyto enzymy vystřihnou intronové části a spojí exonové sekvence do finálního transkriptu, který je vstupem translace [13, 35].

3.2 Translace

Zatímco transkripce probíhá u eukaryotických buněk v jádře, translace probíhá v cytoplazmě na jiných buněčných organelách, a to ribozomech. Ribozomy jsou tvořeny vlákny rRNA (ribozomální RNA), na jejich povrchu translační proces překládá transkript mRNA na řetězec aminokyselin, jak je vidět na obrázku 3.1. Mechanismus funguje na následujícím principu: Řetězec je čten po trojicích nukleotidů, takzvaných kodonech. Molekuly tRNA (transferová RNA) přináší k ribozomu aminokyseliny, které jsou navázány na antikodon, což je opět triplet nukleotidů. Na kodony mRNA se váží komplementární antikodony tRNA a vznikající řetězec aminokyselin se spojuje v protein [14].



Obrázek 3.1: Průběh proteosyntézy, konkrétně translace [3].

Algoritmus překódování kodonu na aminokyselinu je shrnut na obrázku 3.2. Genetický kód, jak je tento algoritmus nazýván, je degenerovaný, neboť vidíme, že některé aminokyseliny jsou kódovány více než jedním kodonem. Speciální kodony existují jak pro začátek translace (start kodon – aminokyselina methionin), tak i pro konec (tzv. terminační kodony). Jelikož funkce i trojrozměrná struktura proteinu jsou přesně dány pořadím a typem aminokyselin, při procesu transkripce a translace je bezpodmínečně nutné, aby zúčastněné enzymy pracovaly bez chyb [14, 12].

	U		C		A		G	
U	UUU	fenylalanin	UCU	serin	UAU	tyrosin	UGU	cystein
	UUC	fenylalanin	UCC	serin	UAC	tyrosin	UGC	cystein
	UUA	leucin	UCA	serin	UAA	**stop**	UGA	**stop**
	UUG	leucin	UCG	serin	UAG	**stop**	UGG	tryptofan
C	CUU	leucin	CCU	prolin	CAU	histidin	CGU	arginin
	CUC	leucin	CCC	prolin	CAC	histidin	CGC	arginin
	CUA	leucin	CCA	prolin	CAA	glutamin	CGA	arginin
	CUG	leucin	CCG	prolin	CAG	glutamin	CGG	arginin
A	AUU	izoleucin	ACU	treonin	AAU	asparagin	AGU	serin
	AUC	izoleucin	ACC	treonin	AAC	asparagin	AGC	serin
	AUA	izoleucin	ACA	treonin	AAA	lysin	AGA	arginin
	AUG	**metionin**	ACG	treonin	AAG	lysin	AGG	arginin
G	GUU	valin	GCU	alanin	GAU	kys.	GGU	glycin
	GUC	valin	GCC	alanin	GAC	asparagová	GGC	glycin
	GUA	valin	GCA	alanin	GAA	kys.	GGA	glycin
	GUG	valin	GCG	alanin	GAG	glutamová	GGG	glycin

Obrázek 3.2: Kódování aminokyselin pomocí kodonů mRNA [35].

3.3 Kompozice genomu

Z hlediska predikce vlivu mutací je nutné klasifikovat jednotlivé úseky DNA, protože predikční přístupy k mutacím se v různých částech genomu liší. Rozlišujeme zejména dvě základní kategorie, a to kódující a nekódující část genomu.

Kódující úseky jsou ty sekvence nukleotidů, které jsou součástí exonu nějakého genu, přímo tedy kódují aminokyselinový řetězec proteinu. Je patrné, že jsou pro život organismu naprosto zásadní, přesto zaujímají jen 1,5–2 % délky genomu [15, 51].

Zbylých cca 98 % délky jsou části takzvaně nekódující. Patří sem intronové oblasti včetně sekvenčních motivů zodpovědných za rozpoznání hranice exonu a intronu, promotory, regulační oblasti a další. Vzhledem k rozsáhlosti těchto částí je pochopitelné, že se zde může rovněž objevit velké množství mutací. I když tyto mutace neovlivňují přímo sekvenci výsledných proteinů, protože nepodléhají transkripci, jejich funkční anotace má velký význam. Nachází-li se totiž například v promotoru nebo na klíčovém regulačním místě, mohou ovlivnit, zda k vytvoření proteinu vůbec dojde. Mutace na hranici intronu a exonu pak může ovlivnit výsledek sestřihu. Tomuto a dalším tématům je věnována následující kapitola [15].

Kapitola 4

Mutace DNA a jejich vliv na organismus

V této kapitole se budu věnovat obecně mutacím v lidské DNA, jak, kdy a kde se objevují, jaký je jejich charakter, co mohou obecně způsobovat. Tím bude doložena důležitost predikce jejich vlivu na proteiny a celý organismus. Komplexní znalosti variant v lidském genomu pak totiž přispějí k možnosti účelné léčby genetických chorob, ještě většímu rozmachu genového a proteinového inženýrství či možnosti vytváření umělých organismů.

4.1 Klasifikace mutací

Nukleotidové varianty v DNA podléhají základnímu dělení na tyto třídy [35]:

- inzerce,
- delece,
- substituce.

Rozdělení mutací je demonstrováno na obrázku 4.1. Inzercí rozumíme vložení sekvence nukleotidů délky alespoň 1 (délka shora je v podstatě neomezená) na určité místo genomu. V tomto rozdělení mezi inzerce řadíme i takzvané duplikace či n-násobná opakování, která jsou v některé literatuře popsána zvlášť. Podmínkou duplikace či násobení je, že se kopírovaná sekvence musí vložit hned za zdrojový templát. Delece způsobí naopak odebrání jednoho nebo více nukleotidů z DNA sekvence. Opět lze odebírat i poměrně rozsáhlé úseky, které ovšem, jak bude ještě zmíněno, nemusí nutně mít signifikantní vliv na funkci organismu. V případě substituce se intuitivně jedná o záměnu jednoho nukleotidu za jiný [30].

Další dělení v sobě odráží místo, kde k mutaci dochází. V předchozí kapitole jsem uvedl, že genom se dělí na kódující a nekódující úseky. Stejně tak můžeme rozdělit i mutace zkráceně na kódující a nekódující. Mutace kódujících částí mohou přímo měnit primární strukturu vznikajícího proteinu. Na základě vědomostí o degenerovanosti genetického kódu je však nutné provést ještě podrobnější rozdělení těchto mutací do těchto kategorií [35, 30]:

- Synonymní (*synonymous*) mutace: Jedná se o takové substituce, které sice způsobí změnu nukleotidu, ovšem čtený kodon při translaci kóduje stále stejnou aminokyselinu. Pochopitelně, taková změna má sama o sobě velmi malý vliv na fungování proteinu, proto i při analýze škodlivosti velké množství prediktorů takové mutace rovnou vyřazuje a označuje za neutrální.

Obrázek 4.1: Demonstrace principu základních typů mutací [35].

- Nesmyslné (*nonsense*) mutace: V tomto případě dochází k tomu, že místo původní trojice nukleotidů kódující aminokyselinu se detekuje terminační kodon. Taková změna může, ale nemusí mít význam. Záleží zejména na tom, o jakou část byl tímto původní protein zkrácen a zda vzniklá zkrácená část obsahuje všechna důležitá vazebná místa.
- Nesynonymní (*nonsynonymous*) mutace: Do této kategorie spadají takové změny, které mění výslednou aminokyselinu proteinu. Predikce jejich vlivu vyžaduje poněkud komplexnější přístup, který bude rozveden dále.

Varianty v nekódujících částech genomu můžeme dělit opět dle regionu, ve kterém se nachází, tedy hranice intronů a exonů, zbylé části intronů, regulační oblasti a další.

4.2 Příčiny mutací, mutagenní faktory

Mutace se mohou v DNA objevovat jistým způsobem samovolně či náhodně, tedy bez zásahu člověka, například vlivem prostředí. Existují však i postupy, kdy lze molekulu DNA upravit a mutace způsobit uměle, v laboratoři.

4.2.1 Mutace při vzniku nové buňky

K výskytu mutace bez zásahu člověka může dojít v několika případech. Prvním z nich je replikace DNA při dělení buňky. Lidské buňky se vzhledem k pohlavnímu rozmnožování dělí pomocí meiózy. Zjednodušeně lze o tomto procesu říci, že rodiče vyprodukují každý jednu sadu 23 chromozomů do haploidní buňky, takzvané gamety. Obě gamety se pak spojí a vytvoří zárodeční buňku nového organismu, zvanou zygota. Jejím dělením pak dojde k vzniku celého organismu. Je třeba poznamenat, že chromozomy jsou v gametě poskládány náhodně a při vzniku zygoty se chromozomy přes sebe překládají a kombinují. Ve zkratce, některé geny přejímá nový jedinec od matky, jiné od otce. Tato náhodnost je velkým zdrojem genetické rozmanitosti, neboť nikdy nedojde k tomu, že je potomek

přesnou kopií svého rodiče. To je tedy první zdroj odlišností v DNA [39, 3], byť se v tomto případě nejedná o projev mutace.

K mutacím jako takovým může docházet při jednorázové replikaci DNA buňky (například při vzniku gamety) a to tehdy, když replikační mechanismus udělá chybu. V buňce pro tyto účely existují opravné mechanismy, které přinutí hlavní replikační enzym, DNA polymerázu, aby chybně připojené nukleotidy odstranila. Nicméně, neopravené chyby se ve fázi replikace objevují, a to přibližně po každých 10^7 připojených nukleotidech. Naprostá většina (99 %) z těchto chyb je poté ještě eliminována systémem oprav chybného párování bází. Tento systém detekuje špatně spárované dusíkaté báze a daný úsek odštěpí. DNA polymeráza pak dosyntetizuje dle komplementárního vlákna správnou sekvenci, která je do šroubovice připojena pomocí dalšího enzymu, DNA ligázy. Ani poté však nelze garantovat bezchybný výsledek replikace, opravné mechanismy navzdory své specifičnosti některé chyby nedetekují a mutace v DNA přetrvá. K tomuto jevu dochází zřídka, je ovšem nezbytný pro genetický vývoj organismů. Zároveň je nutné poznamenat, že opravující enzymy také vznikají na základě DNA předpisu a mohou být tedy mutacemi poškozeny a jejich přesnost se může snížit [39, 12, 3].

4.2.2 Fyzikální a chemické faktory

Ke vzniku mutací buněk přispívají také další mutagenní faktory. Z fyzikálních faktorů jsou nejúčinnější záření, a to jak ionizující, tak neionizující. Stupeň poškození DNA je úměrný množství pohlceného záření. Ionizující záření (například gama, neutronové, protonové) způsobuje zejména zlomy řetězců DNA a vytváření kovalentních vazeb mezi dusíkatými bázemi namísto vodíkových. Neionizující záření, například ultrafialové, působí poněkud specifičtěji. Může způsobovat substituce nukleotidů na určitých místech a také vytvářet chemicky stabilní dimery dvou sousedních nukleotidů, které pak brání transkripci [35].

Dále mohou být zdrojem mutací chemické faktory, které mění strukturu DNA, poškozují nebo zaměňují nukleotidy pomocí chemické reakce. Sekvenci nukleotidů mohou také měnit některé viry či retroviry, které například mohou způsobit, že se buňka začne nekontrolovatelně dělit a na daném místě tkáně pak může vznikat nádor. A právě léčba nádorových onemocnění je velkou motivací pro další výzkum DNA a mutací [35].

Kapitola 5

Predikce vlivu mutací

Největším problémem výzkumu mutací je určení, jak se daná mutace projeví na fenotypu. Úvodem je třeba vymezit, jak se k tomuto problému přistupuje. Každá mutace může obecně mít na DNA, proteiny a posléze i organismus buď neutrální vliv (nic se jejím výskytem nezmění), nebo může působit škodlivě. Mezi škodlivé (*deleterious*) mutace řadíme všechny, které negativně působí na organizmus. Za projev negativního působení lze považovat nejen potlačení funkce proteinů, ale v řadě případů i její zesílení [42, 15, 25].

5.1 Přesnost predikce

Nejpřesněji lze vliv mutace určit laboratorní metodou. Pomocí specifických enzymů (zejména restrikčních endonukleáz) lze na daném místě vlákno DNA rozštěpit, provést mutaci, vlákna opět spojit a pozorovat efekt [3, 35]. Tento postup je velmi přesný a o každé mutaci může spolehlivě říci, jaký má význam, ovšem je časově i finančně vysoce náročný. Už jen kdybychom uvažovali pouze bodové substituce, tedy záměnu jednoho nukleotidu, pak na celém lidském genomu bychom jich museli provést přibližně 9 miliard. Pokud uvažujeme ještě alespoň krátké substituce a delece, počet nám ještě několikanásobně naroste. Proto vznikly nové výzkumné směry, které se soustředí na využití výpočetní síly počítačů k co nejpřesnější predikci vlivu mutací bez nutnosti provádět laboratorní ověření. Tyto přístupy sice nemohou garantovat stoprocentní přesnost predikce, ale lze je s výhodou použít jako filtr pro vyřazení mutací s jistotou označených jako škodlivé. Cílem je zúžení sady mutací vhodných k experimentálnímu ověření na rozumnou velikost [15].

5.2 Bioinformatika a přístupy k predikci

Bioinformatika je věda, zabývající se shromažďováním, uchováváním, organizací a analýzou biologických dat. Identifikace škodlivých mutací v množině všech přípustných variant v lidském genomu je jedním z jejích hlavních cílů. I když existuje velké množství metod, které komplexně popisují a sumarizují lidskou genetickou diverzitu, identifikace škodlivých mutací je stále velkou výzvou. Díky rozvoji sekvenovacích metod a narůstajícího množství nasekvenovaných lidských jedinců lze pozorovat poměrně značné množství substitucí, krátkých delecí a inzercí v populaci. Většina z nich je poměrně běžná, ale vyskytují se i mutace vzácnějšího charakteru. V tuto chvíli je hlavním problémem klasifikace a anotace mutací a míst, kde se nacházejí, spíše než sběr dat [15].

5.3 Výpočetní predikční metody

Při měření škodlivosti mutací se zatím stále vychází nejvíce z evolučních informací. Většina metod bere v úvahu zejména fakt, že sekvence DNA, které jsou pozorované v nezměněné podobě u více živých druhů organismů, jsou pro život a zdraví důležité, protože se i přes genetickou rozmanitost zachovaly a nebyly odstraněny v rámci přírodního výběru. Komparativní studie dovedou při dostatečném množství vstupních dat takové sekvence odhalit a stávají se vítaným zdrojem informace pro odhad škodlivosti. Mutace v evolučně konzervovaných úsecích budou s velkou pravděpodobností škodlivé. Tento přístup však není zcela přesný a má své podstatné limity. Konzervační analýza, jak je hledání evolučně konzervovaných sekvencí označováno, není pro stanovení škodlivosti mutace sama o sobě dostačující, neboť kvantitativní vyjádření evoluční konzervovanosti je jen jednou z vlastností daného místa DNA a k přesnější predikci je třeba přidružit větší množství informací. Druhým podstatným faktem je fylogenetický rozsah aplikace určující sadu organizmů, s kterými bude analyzovaná DNA sekvence porovnávaná [15, 42].

Například, při porovnání lidského genomu a genomu kvasinek (*yeast*) lze samozřejmě pozorovat velké odlišnosti. Pokud se ale objeví sekvence vysoké podobnosti, je zřejmé, že jsou evolučně již velmi dlouho udržované v konstantním stavu a proto mutace v nich budou pravděpodobně škodlivé. Tento přístup tedy vykazuje vysokou specificitu vzhledem k výběru škodlivých mutací, senzitivita však bude nižší, neboť všechny mutace mimo konzervované úseky jsou brány jako neutrální, čímž se zvyšuje počet falešně negativních výsledků. Naopak, při porovnání lidského a šimpanzího genomu, které vykazují až 98.8 % podobnost, dochází ke značné ztrátě specificity, protože se pak vyskytuje velké množství falešně pozitivních predikcí [51, 15].

Tyto přístupy, které jsou založené na škodlivostní anotaci pozic genomu, předpokládají, že pozice budou mít detekovatelnou historii tzv. purifikační selekce (*purifying selection*), tedy mechanismu, který nedovoluje vytvářet mutace v konzervovaných úsecích a tím genetickou informaci chrání před většinou škodlivých genetických změn. Ovšem, například lidský organismus, resp. jeho DNA, získal větší množství genů a regulačních sekvencí teprve v nedávné historii, takže je nelze v komparativních studiích detekovat.

Metody predikce se liší obecně podle toho, do jakého regionu mutace spadá. Většina nástrojů předpokládá minimálně rozdělení na mutace v kódující části (exon) a v nekódující (intron, regulační sekvence apod.) [15].

5.3.1 Predikce v kódujících úsecích

Tyto metody jsou obecně propracovanější než metody pro nekódující části, neboť jsou vyvíjeny již delší dobu. Exonová část genomu je krátká (cca 2 % délky) a do této doby už poměrně dobře prozkoumaná a anotovaná. Obecně se předpokládá, že většina mutací, které způsobí posun tříbodového čtecího okna při translaci (inzerce, delece), případně způsobují předčasné ukončení translace (vznikne terminační kodon) jsou škodlivé. Neplatí to však vždy, jak dokazují některé nové studie [31]. Co se poté týče nesynonymních mutací, tedy těch, které mění aminokyselinu, přistupuje se vedle evolučních informací i k analýze, jak drastická biochemická změna v proteinu nastala, tedy, jak se liší nová aminokyselina od původní. Díky delšímu výzkumu je k dispozici větší množství přístupů, ty hlavní budou nyní krátce rozvedeny [15].

Existují metody, které vycházejí z konkrétních informací o daném místě, tedy například vyčíslí jeho evoluční konzervovanost a podle ní predikují (*tzv. first-principle metody*) [15]. Druhým typem jsou metody, které využijí trénování na známé množině mutací (*tzv. knowledge-based metody*). Princip jejich práce spočívá ve vytvoření skupiny pravidel,

které podle vlastností mutace (evolučních, chemických, biologických a fyzikálních) potom predikují její škodlivost. Rozdíl obou přístupů je zjevný: First-principle metody vychází z „pravdivých dat“. Bývají tedy obecně přesnější, jelikož nemohou být negativně ovlivněny trénovacími daty. Při nedostatku informací však přesnost jejich predikce dramaticky klesá, protože nejsou tak robustní. Jsou využity například v bioinformatických nástrojích MAPP [50], SIFT [26] a PANTHER [53]. Naopak, knowledge-based klasifikátory se obecně vyznačují vyšší robustností predikce, která sice není na určitých DNA variantách tak přesná, ovšem jistým způsobem garantuje stejnou chybovost na celé exonové části genomu. Velmi zde záleží na trénovací množině dat a procesu učení. Tento problém je řešen i v praktické části této práce v dalších kapitolách. Příkladem takových nástrojů jsou SNAP [7], PolyPhen-2 [2] a PhD-SNP [8].

Jak již bylo řečeno, k predikci jsou často využívány i informace o biochemických vlastnostech aminokyselin, dále např. informace o vazebných místech proteinů a strukturní informace (přítomnost beta skládaného listu). Zahrnutí dalších znalostí do predikčního procesu výrazně zvyšuje její přesnost, což je v podstatě uvedeno již v úvodu této kapitoly.

5.3.2 Predikce v nekódujících úsecích

V předchozích kapitolách bylo uvedeno, že i nekódující úseky DNA mají významný vliv na proteiny, zejména na jejich vytváření a posttranskripční úpravy. Navíc i díky dominanci délky nekódující DNA je počet mutací v nich značný [17]. Proto se výzkum DNA v poslední době více zabývá i mutacemi v nekódujících částech DNA, ukázalo se totiž, že mohou být rovněž zdrojem genetických chorob [49]. Dále, evoluční analýza ukázala, že v nekódující části DNA se nachází co do délky přibližně pětkrát více konzervovaných úseků, tedy těch, které podléhají čistící selekci.

Stejně jako v předchozím případě, i pro nekódující varianty je stěžejní analýza konzervovanosti. S tím rozdílem, že v jejich souvislosti lze pozorovat výrazně menší podobnost mezi druhy, lidský genom lze komparativně analyzovat pouze s některými obratlovci, diverzita proti například bakteriím je již příliš vysoká. Velice často se proto při analýze lidské DNA přistupuje k zarovnání například s myší, primátů, posléze i dalších savců. Tímto způsobem bylo odhaleno, že téměř 8 % délky nekódující DNA je evolučně konzervováno [15].

Základní princip metod predikce je velice podobný u všech existujících nástrojů a přístupů. Jedná se o kvantifikování průměrného počtu mutací na daném místě DNA na základě fylogenetické informace a zarovnání sekvencí DNA zvolených organismů, v našem případě obratlovců. Průměrná rychlost výskytu mutace se pak porovnává s obvyklou četností mutací pro neutrální pozice. Pokud je někde četnost mutací výrazně nižší, je nasnadě, že tato pozice podléhá purifikační selekci, tudíž mutace v ní budou pravděpodobně škodlivé. Predikční metody se liší hlavně tím, jak s touto informací naloží, zda ji ještě něčím obohacují. Některé z nich (binCons [32], phastCons [48]) používají mechanismy takové, že celkové skóre dané pozice je ovlivněno jak evoluční informací o této konkrétní pozici, tak informací od sousedů. Vychází z předpokladu, že konzervované nukleotidy se vyskytují většinou v řetězcích vedle sebe. Naproti tomu metody GERP [17], phyloP [36] a SCONE [5] analyzují každou pozici zvlášť, což při dostatku vstupních dat k zarovnání není na škodu. Naopak, jejich výsledky nejsou ovlivněny ničím jiným než faktickými znalostmi o konkrétní pozici. Nově vznikající nástroje se snaží k informaci o konzervovanosti přidat další informace a anotace a zvýšit tak přesnost predikce [15, 42].

5.3.3 Shrnutí metod

Z hlediska přesnosti se jeví jako nejlogičtější postup při predikci zjistit, kde se daná mutace nachází a dle toho použít příslušnou metodu. Pro kódující mutace totiž mohou být využity i informace o aminokyselinách. Dle nedávného výzkumu [15, 51] tomu tak ale obecně nemusí být, nekódující predikční metody mohou být natolik obecné, že na kódujících mutacích mají srovnatelnou úspěšnost s metodami přímo pro exonové varianty. Vysvětlení může být více, první je to, že analýza vycházející pouze z komparace lidského genomu se savčími neuvažuje některé paralogy a funkcionálně vzdálené ortologní sekvence, které se vyskytují u fylogeneticky vzdálenějších druhů. Při analýze na proteinové úrovni se totiž taková omezení obvykle nekladou, což může snížit predikční sílu. Dále, absence funkcionální informace proteinové úrovně může být prospěšná, neboť zatím není vše důkladně prozkoumáno a tyto informace mohou být zavádějící. Příkladem je synonymní mutace na úrovni aminokyseliny. Přestože nedojde k její záměně, na nukleotidové úrovni mohou být poškozeny hranice intronu a exonu a tím i sestřih proteinu.

Velkou, ne-li největší výzvou výzkumu v oblasti předpovědi efektu mutací je vytvoření prediktoru, který dokáže určovat vliv mutací na celé lidské DNA bez ohledu na pozici a typ mutace, ke které dochází.

Kapitola 6

Nástroje pro predikci nukleotidového polymorfismu

V této kapitole budou rozebrány konkrétní softwarové nástroje sloužící k predikci vlivu DNA variant. Dle literatury byly nastudovány jejich přístupy a principy, na základě přesnosti, rychlosti a dostupnosti jsem poté vybral pět z nich, které nejlépe splňovaly požadavky. Tyto nástroje jsem poté použil pro vytvoření nového prediktoru v praktické části práce.

Prvním cílem bylo získat maximální přehled o existujících metodách a jejich výsledcích, aby bylo možné vybrat ty nástroje, které se liší přístupem a mají dle testování přijatelnou přesnost predikce. V tabulce 6.1 jsou uvedeny už pouze ty nástroje, které budou použity v praktické části.

Z tabulky je patrné, že některé nástroje poskytují pouze část znalostí o škodlivosti mutací, jedná se o phastCons [48], phyloP [36] a GERP++ [17]. Všechny tři provádí různým způsobem totéž, snaží se pouze označit na lidském genomu konzervovaná místa. Jak bylo ovšem uvedeno v kapitole s rozborem metod, tato informace je nedostačující. Tyto nástroje tedy nebudou přímou součástí výsledného konsenzu, jejich výsledky se ale promítnou. Skóre konzervovanosti, které udává nástroj GERP++ je totiž součástí anotací, které používá nástroj GWAVA [41]. Webový nástroj MutationTaster2 [45] zase využívá výsledků nástrojů phyloP a phastCons. Nástroj SIFT (SIFT DNA) patří do skupiny prediktorů aminokyselinových substitucí. V předchozích kapitolách již ale bylo zmíněno, že záměna aminokyseliny je způsobena nukleotidovou substitucí. SIFT sice nedovede predikovat efekt mutace na celém genomu, avšak na rozdíl od ostatních nástrojů ke klasifikaci aminokyselinových substitucí má rozšířené možnosti vstupu, lze u něj zadat mimo proteinu a mutace rovněž lidský chromozom, pozici a nukleotidovou mutaci. Princip jeho činnosti je odlišný od nukleotidových prediktorů, a proto byl rovněž vybrán do výsledného konsenzu.

Nástroj	Vlastnosti	Popis	Algoritmus, parametry
phastCons PhyloP	Pouze hledají konz. úseky. Výsledky ke stažení.	Porovnávají četnost mutací z vícenásobného zarovnání (zkr. MSA) chromozomů 45 obratlovců s informacemi z fylogenetického modelu a hledají odchylky.	phastCons využívá HMM, PhyloP mechanismus hledání odchylek očekávané a pozorované četnosti mutací. Lze omezit délku hledaných konzervovaných úseků.
GERP++	Hledá konz. úseky, nepredikuje škodlivost. Volně ke stažení.	Na základě fylogenetického stromu odhadne neutrální počet mutací na dané pozici, poté vypočítá z MSA chromozomů 33 obratlovců rozdíl počtu mutací proti odhadu. V druhé fázi generuje kandidátní úseky, spojuje a finální vybírá od statisticky nejvýznamnějších.	Pro běh lze nastavit minimální a maximální délku úseku (typicky 4b–2kb) a parametry pro výpočet minimálního skóre.
GWAVA	Určuje škodlivost substitucí v DNA. Ke stažení.	Integruje genomické a epigenomické anotace pozic. Pracuje na celém genomu.	Pro kombinaci dostupných anotací místa využívá tři náhodné lesy. Podle regionu pozice mutace je lze parametricky volit.
MutationTaster2	Určuje škodlivost mutací všech typů. Webová služba.	Kombinuje skóre konzervovanosti s dalšími znalostmi, pro nejčastější mutace má předpočítaný výsledek, což zvyšuje rychlost.	Tři Bayesovské klasifikátory, nástroj sám vybere podle typu a místa mutace. Některé rovnou určí jako škodlivé či neutrální.
CADD	Poskytuje předpočítaná skóre škodlivosti. Výsledky ke stažení.	Integruje funkční anotace, patogenicitu, regulační efekt, konzervovanost a další. Poskytuje předpočítané skóre pro všechny možné substituce i krátké inzerce/delece.	Deset SVM modelů (*support vector machine*), natrénovaných a otestovaných se snaží zkombinovat různé anotace daného místa/mutace do jednotného skóre.
FATHMM-MKL	Poskytuje předpočítaná skóre škodlivosti. Výsledky ke stažení.	Klasifikuje na základě funkčních anotací a konzervovanosti pozic. Poskytuje předpočítaná skóre pro substituce na celém genomu.	Dva SVM modely pro kódující, resp. nekódující část genomu. Kombinují anotace, které nejlépe oddělují neutrální od škodlivých mutací.
SIFT DNA	Predikuje efekt nesynonymních substitucí. Ke stažení.	Predikuje efekt pouze nesynonymních substitucí. Znalosti získává z databází, nikoli učením.	V databázi hledá podobné proteiny, na základě jejich MSA určí relativní pravděpodobnost mutace a z ní skóre škodlivosti.

Nástroj	Vstup	Výstup	Srovnání, testování
phastCons [48] **phyloP** [36]	Vícenásobné zarovnání, fylog. model, phastCons navíc parametry HMM	Soubor ve formátu WIG, pro každou pozici pravděpodobnost konzervovanosti	PhyloP: Fylog. i statistický přístup - celkem 4 testy, všechny vykazovaly srovnatelnou významnost.
GERP++ [17]	Sekvenční zarovnání, parametry, fylogenetický strom	Soubor s řádkem pro každý konz. element ve tvaru: start end length RS-score p-value	Srovnání s phastCons: GERP++ označuje více úseků za konzervované, hlavně v exonech. Výsledky se ale z 80 % překrývají.
GWAVA [41]	BED soubor s číslem chromozomu, pozicí a RS# identifikátorem mutace	BED vstupní soubor rozšířený o predikční skóre tří modelů	Na nezávislé sadě úspěšnější než MutationTaster.
MutationTaster2 [45]	Lze nahrát VCF soubor nebo použít webový formulář	Binární klasifikace mutace (neutrální x škodlivá)	Překonání úspěšnosti nástrojů SIFT, PolyPhen-2 a PROVEAN na testovací sadě složené z exonových mutací.
CADD [25]	není	TSV soubor, na řádku pozice v genomu, mutace a C-skóre na škále <0,99>	Porovnání kódujících mutací s SIFT a PolyPhen-2, CADD úspěšnější
FATHMM-MKL [47]	není	TSV soubor, na řádku pozice v genomu, mutace a skóre na škále <0,1>	Srovnání s GWAVA a CADD zvlášť na nekódující a kódující množině mutací.
SIFT DNA [26]	CSV soubor ve formátu: chromozom, pozice, orientace, mutace	TSV soubor mutací včetně anotací a SIFT skóre na škále <0,1>	Vzhledem k původní povaze nástroje bylo srovnání prováděno jen s dostupnými aminokyselinovými prediktory.

Tabulka 6.1: Shrnutí vlastností nástrojů.

Hlavními faktory při výběru nástrojů pro výsledný konsenzus byly: (i) úspěšnost jejich predikce, (ii) sofistikovanost metod, (iii) velikost trénovací sady, (iv) dostupnost. Na základě těchto znalostí byly tedy pro výsledný konsenzus vybrány nástroje GWAVA, CADD, FATHMM-MKL, SIFT DNA a MutationTaster2. V následujících kapitolách proto bude detailněji popsán jejich princip, výsledky a validace.

6.1 GWAVA

Nástroj GWAVA (*Genome-Wide Annotation of Variants*) je schopen vyhodnocovat substituce, a to jak v kódujících, tak i nekódujících úsecích lidského genomu. Vývoj nástroje vychází z následujícího předpokladu: Dříve implementované nástroje, které se zabývaly zejména mutacemi v kódujících sekvencích DNA, vycházely hlavně z analýzy konzervovanosti. To bylo možné, neboť kódující sekvence se nemění tak často a důležité úseky jsou velmi dobře evolučně chráněny. Rozšířením analýzy na celý genom už ale konzervovanost ztrácí svou sílu, neboť neodhalí všechny klíčové úseky, které při zásahu mutací negativně ovlivňují fenotyp. Příkladem jsou regulační sekvence, které jsou neméně důležité jako samotné kódující úseky, ale vyznačují se oproti nim zvýšenou četností přirozených změn. Snahou nástroje GWAVA bylo vedle informací o konzervovanosti začlenit do procesu nové znalosti, které by upřesnily predikci zejména v nekódujících částech genomu.

Tyto znalosti GWAVA získává z projektu ENCODE [18] (*Encyclopedia of DNA elements*). Z tohoto zdroje je možné získat velké množství anotací DNA pro lidský genom. Hlavním problémem tedy nebyl sběr dat, nýbrž stanovení, které z těchto anotací korelují se škodlivostí mutací a mohou tedy být použity k predikci. Díky experimentům a trénování se nakonec podařilo vybrat 10 anotací, z nichž největší podíl na výsledku predikce měly: (i) konzervovanost - zde bylo využito skóre poskytované programem GERP++, (ii) vzdálenost k nejbližší TSS (místo začátku transkripce, *Transcription start site*), (iii) modifikace histonů, (iv) hypersensitivita DNase I a (v) CpG ostrůvky [41].

6.1.1 Trénování

Rozhodovacím modelem GWAVA je náhodný les (*random forest*). Byl vybrán díky své možnosti zpracovávat více vstupních informací různého typu, zvládat rozdílnou velikost škodlivé a neutrální složky mutací trénovací sady a robustnosti k přítomnosti informací, které nepřispívají k predikci. Celkem byly natrénovány tři rozhodovací modely o přibližně 100 stromech, každý s jinou trénovací sadou. Složka škodlivých mutací byla pro všechny stejná, bylo použito 1 614 mutací z databáze HGMD [49] (The Human Gene Mutation Database). Neutrální mutace byly vybírány z databáze projektu 1KG [1] (The 1000 Genomes Project). Pro první les byl vybrán náhodný vzorek (*unmatched model*), pro druhý mutace, které se nachází v blízkosti nějaké TSS (*tss model*), protože mutace z HGMD jsou z 75 % také kolem TSS) a pro třetí mutace z 1000 bp[1] okolí variant z HGMD (*region model*). Křížová validace (varianta 10-fold) ukázala nejlepší výsledky pro les s náhodným vzorkem neutrálních mutací. Zřejmě proto, že takový model rovnoměrně pokrývá celý genom [41].

6.1.2 Testování

Validace natrénovaných lesů probíhala v několika fázích a na několika datových sadách. Nejdříve proběhlo testování na sadě škodlivých variant z databáze ClinVar [27] (po odstranění překryvu s HGMD) doplněné nejprve o neutrální mutace z ClinVar a poté o vzorek

[1] bp - počet párů bází (*base pairs*), tedy počet nukleotidových pozic.

neutrálních mutací z 1KG. Výsledné hodnoty AUC (*Area Under Curve*) jejich ROC (*Receiver Operating Characteristics*) charakteristik byly 0.84, respektive 0.75. Dále proběhla úspěšná testování na mutacích z GWAS studií [23] (*Genome-Wide Association Study*), z personální genomiky (mutace na chromozomu 22 jednoho jedince, z 1KG) a na mutacích spojených s rakovinou (databáze COSMIC) [41].

6.2 CADD

Nástroj CADD (*Combined Annotation Dependent Depletion*) není dostupný ke stažení, nicméně lze získat jeho výsledky, totiž předpočítaná skóre škodlivosti pro všechny možné substituce na lidském genomu (přibližně 8,6 miliard) a také nejčastější krátké inzerce a delece (do 50 bp). Při vývoji nástroje se vycházelo z podobných předpokladů jako u nástroje GWAVA a stejně tak se dospělo k řešení pomocí kombinace různých anotací, ovšem tato tematika byla rozebrána podrobněji. Vedle ENCODE byly použity i další zdroje anotací, např. Ensembl VEP [33] (*Variant Effect Predictor*) nebo webový portál UCSC Genome Browser, celkem pak bylo identifikováno 63 typů anotací. Ovšem, každá anotace má svou metriku a interpretace sady různých hodnot je tudíž problematická. Proto byla vytvořena numerická hodnota C-skóre, do které se svým způsobem promítají informace ze všech anotací [25].

Pro výběr vhodných anotací bylo použito následujícího experimentu: Zarovnáním lidského a šimpanzího genomu bylo odhaleno 15 milionů substitucí a 1,7 milionu krátkých inzercí a delecí. Na základě vícenásobného zarovnání genomů primátů bylo pak vygenerováno podobné množství nových mutací, celkem jich tedy bylo přibližně 30 milionů. Vznikla matice 30 milionů mutací krát 63 anotací. Úkolem bylo najít anotace, které co nejlépe odlišují mutace simulované od mutací derivovaných z lidského a šimpanzího genomu (dále označené jako derivované mutace). Tam, kde hodnoty anotací chyběly, byla nastavena nula nebo průměrná hodnota, jinak byla pro simulovanou mutaci vložena hodnota -1 a pro derivovanou 1 [25].

6.2.1 Trénování

S takto vzniklou maticí bylo natrénováno celkem deset SVM modelů s lineárním jádrem, každý na náhodném vzorku z trénovací množiny 30 milionů variant. Výstupem predikce je hodnota C-skóre, u které se předpokládala silná korelace se škodlivostí variant. Ovšem, experimenty potvrdily, že tato hodnota má odlišný význam v různých regionech genomu. Například, u synonymních kódujících variant je jiný práh mezi škodlivou a neutrální mutací než u mutací v regulačních úsecích. Pokud je tento fakt při testování zohledněn, dosahuje nástroj velmi dobrých výsledků. Hlavní výhodou trénování CADD je, že používá vlastní novou trénovací sadu, jejíž varianty pokrývají celý genom, což významně podporuje jeho důvěryhodnost [25].

6.2.2 Testování

K ověření korelace C-skóre a škodlivosti bylo provedeno testování, a to v pěti fázích. Validace probíhala na (i) kombinované sadě mutací z HGMD, GWAS a dalších, (ii) sadě mutací spojených s Kabuki syndromem, doplněné o neutrální mutace z ESP [52] (*Exome Sequencing Project*), (iii) sadě variant z okolí mutací způsobujících talasémii z databáze HbVar, (iv) sadě mutací z ClinVar, (v) sadě mutací, pozorovaných u dětí s autismem a intelektuálními disproporcemi. Korelace škodlivosti a C-skóre byla v těchto případech potvrzena [25].

6.3 MutationTaster2

Dalším z uvažovaných nástrojů je MutationTaster2. Jeho hlavní výhodou je dostupnost přes webové rozhraní, univerzálnost (predikuje všechny typy mutací, na celém genomu) a rychlost. Rychlost predikce je zaručena několika faktory:

- Nástroj má k dispozici předpočítané výsledky pro nejčastější mutace z 1KG, ClinVar a HGMD. Tato předpočítaná data však nelze stáhnout.
- Mutace, které se vyskytují více jak čtyřikrát v téže podobě v 1KG nebo databázi HapMap se označí jako neutrální, neboť jejich výskyt je příliš častý, než aby byly škodlivé.
- Jako rozhodovací model je použit Bayesovský klasifikátor, který se obecně rovněž vyznačuje vysokou rychlostí [45].

MutationTaster2 je dále proti své původní verzi vylepšen o zpracování mutací v místech posttranskripčního sestřihu, tedy hranic intronu a exonu, čímž se znatelně zvýšila jeho přesnost. Vstupem Bayesovského klasifikátoru jsou mimo jiné skóre konzervovanosti, které poskytují nástroje phyloP a phastCons a dále anotace z Ensembl [19].

6.3.1 Trénování

MutationTaster2 zahrnuje celkem tři natrénované Bayesovské klasifikátory. Neutrální trénovací varianty jsou převzaty z 1KG, přičemž se musí vyskytovat v téže podobě alespoň u dvaceti jedinců. Složka škodlivých mutací je vytvořena z patogenních mutací v databázi HGMD. První klasifikátor je pro mutace, které způsobí změnu jedné aminokyseliny. Je trénován na odpovídajícím vzorku trénovací sady (zahrnuje přibližně 170 000 variant). Druhý klasifikátor je použit, pokud kódující mutace mění více než jednu aminokyselinu, odpovídající trénovací vzorek obsahuje 125 000 variant. Třetí klasifikátor je pro mutace synonymní či v nekódujících částech, trénovací množina obsahuje přes 6 milionů záznamů. Výstupem klasifikátoru není reálná hodnota jako u jiných nástrojů (C-skóre u CADD, škodlivostní skóre u GWAVA), ale pouze binární vyjádření neutrální/škodlivá [45].

6.3.2 Testování

Nástroj je schopen sám určit, který z klasifikátorů použije. V rámci validace byla provedena 5-fold křížová validace, a to přesto, že Bayesovský klasifikátor je méně náchylný k přeučení než jiné modely. Samotná úspěšnost modelu byla testována ve dvou fázích. Nejprve došlo k porovnání predikční síly na kódujících mutacích (tedy prvních dvou klasifikátorů) s existujícími nástroji SIFT, PolyPhen-2 a PROVEAN. Na sadě 2200 mutací z 1KG byl MutationTaster2 nejúspěšnější. Rozhodovací síla pro nekódující varianty byla testována na vzorku dat z 1KG a HGMD, odlišném od trénovací sady [45].

6.4 FATHMM-MKL

FATHMM-MKL (*Functional Analysis through Hidden Markov Models - Multi Kernel Learning*), dále také jen zkráceně FATHMM, patří mezi nejnovější nástroje na poli predikce efektu mutací DNA (publikován v lednu 2015). Je schopen anotovat jednobodové substituce, a to jak v kódující, tak v nekódující části genomu. Stejně jako nástroj CADD, i FATHMM poskytuje předpočítaná skóre predikcí pro všechny mutace na celém genomu

GRCh37, stejně tak tedy platí, že při využití vhodného podpůrného softwaru je práce s ním velmi rychlá. Princip činnosti nástroje vychází z kombinace anotací z ENCODE a analýzy konzervovanosti na nukleotidové úrovni. Tyto znalosti jsou poté použity k natrénování SVM modelu, který bude provádět klasifikaci [47].

6.4.1 Trénování

Dataset pro trénování byl vybudován ze dvou hlavních zdrojů. Škodlivé mutace byly přejaty z databáze HGMD, kdy byly vybrány dědičné mutace v zárodečných buňkách. Je zřejmé, že v této množině jsou i mutace ve své podstatě neutrální, které se ovšem dědily vedle škodlivých. Složka škodlivých mutací je tak interpretována spíše jako množina mutací spojených s patogenicitou. Neutrální mutace byly převzaty z projektu 1KG. Těchto mutací je velké množství a i zde platí, že se mezi nimi mohou nacházet i mutace škodlivé. Proto byla tato množina ještě podrobena selekci, během které došlo k odstranění mutací, které se nacházely zároveň i ve škodlivé složce datasetu a dále ty, které se v rámci projektu 1KG neobjevovaly dostatečně často (je u nich riziko, že budou škodlivé) [47].

Na základě trénovací sady bylo z ENCODE vybráno 10 anotací, které nejlépe dokáží rozlišit mezi neutralitou a škodlivostí mutace. Byly to:

- Skóre konzervovanosti dané zarovnáním lidského genomu s genomy 46 obratlovců,
- lokalizace regulačních pozic dle ChIP-Seq analýzy,
- lokalizace míst navázání transkripčních faktorů dle PeakSeq analýzy,
- lokalizace regulačních úseků dle DNase I (deoxyribonukleáza I),
- skóre konzervovanosti dané zarovnáním lidského genomu s genomu 100 obratlovců,
- koncentrace CG párů,
- lokalizace míst navázání transkripčních faktorů metodou SPP [24],
- segmentace genomu,
- otisky DNA (*DNA footprints*).

Trénovací sada byla poté rozdělena na dvě složky – mutace v kódující a nekódující části genomu, na každé byl trénován jeden SVM model. Pro nekódující sadu byly nakonec uvažovány jen první čtyři anotace a z neutrální složky byly vybrány pouze ty, které se nacházely v 1000 bp okolí nějaké škodlivé mutace. V kódující trénovací sadě se k omezení nepřistoupilo, neboť sada by byla příliš malá.

6.4.2 Testování

Mimo 5-fold cross validace cílí FATHMM hlavně na srovnání s podobnými nástroji, a to zejména CADD a GWAVA. Testování probíhalo na nezávislých množinách mutací z ClinVar a HGMD (byl odstraněn překryv s trénovací sadou). V nekódující části vykazuje FATHMM vyšší úspěšnost než GWAVA a CADD, v kódující části je, co se přesnosti týče, srovnatelný s CADD, přičemž nástroj GWAVA není tak úspěšný.

6.5 SIFT DNA

SIFT DNA (*Separating Intolerable From Tolerable - DNA*), dále také zkráceně SIFT, je specifický nástroj predikce efektu mutací DNA. Jeho možnosti jsou omezené, dokáže zpracovávat pouze nesynonymní kódující mutace, tedy ty, které způsobují změnu aminokyseliny. Je to z toho důvodu, že se původně jedná o prediktor efektu aminokyselinových substitucí. Rozšířením možnosti vstupu lze ale nyní provést dotaz i na úrovni DNA.

V kódujících částech bývá hlavním rozhodovacím faktorem konzervovanost, neboť je nasnadě, že důležité úseky proteinů zůstávají během evoluce neměnné. Klasifikátory aminokyselinových substitucí proto často vycházejí ryze z informací o evolučně stabilních úsecích. V případě SIFT jsou konzervované úseky hledány na základě prohledávání databází. Pomocí programu PSI-BLAST jsou vyhledány nejpodobnější proteinové sekvence z některé z bioinformatických databází, dále se pak vytvoří jejich vzájemné zarovnání. Na každé pozici je vyčíslena pravděpodobnost výskytu každé z 20 aminokyselin, pravděpodobnosti jsou normalizovány dle té nejvyšší a uspořádány do matice substitucí. Pokud klesne pravděpodobnost výskytu nějaké aminokyseliny pod určitou hranici, je to známka, že se bude jednat o škodlivou substituci [26].

Nejvážnější limitací nástroje SIFT je, že neintegruje žádnou další znalost do procesu rozhodování. Jiné prediktory stejného typu využívají např. také informací o struktuře proteinu nebo berou v potaz fyzikálně-chemické vlastnosti aminokyselin. Tyto informace obecně přesnost klasifikace zvyšují. Výhodou nástroje je to, že jej lze aplikovat na všechny organismy, nejen na člověka, princip proteinové evoluce je totiž stejný. Mimo analýzy konkrétních substitucí může SIFT provádět analýzu i celé proteinové sekvence, zadané buď jako výpis řetězce aminokyselin, nebo jeho identifikátor.

6.5.1 Testování

SIFT nepatří mezi nástroje, které by byly trénovány některou z metod strojového učení. Vychází výhradně z databází a zarovnání. Jediným parametrem, který bylo nutné stanovit, byl práh pravděpodobnosti mezi škodlivou a neutrální mutací. Validace nástroje probíhala ve dvou fázích. Nejprve byla analyzována množina nesynonymních SNP zdravých a nemocných jedinců. SIFT označil 81 % z mutací u zdravých lidí jako neutrální a naopak 69 % mutací jako škodlivých u jedinců s chorobou. Ve druhé fázi bylo testování provedeno na vzorku mutací z databáze dbSNP [26].

6.6 Shrnutí

V následující tabulce je souhrnně uvedeno, které typy mutací dokáží nástroje klasifikovat. Tato informace je podstatná z pohledu celkového přehledu o možnostech nástrojů, velmi významnou roli pak má i v praktické části práce při budování trénovacího datasetu.

Nástroj	Mutace			
	Substituce			Inzerce a delece
	Kódující		Nekódující	
	Synonymní	Nesynonymní		
CADD	ANO	ANO	ANO	ANO
GWAVA	ANO	ANO	ANO	NE
FATHMM	ANO	ANO	ANO	NE
MutationTaster2	ANO	ANO	NE	ANO
SIFT	NE	ANO	NE	NE

Tabulka 6.2: Predikční schopnosti nástrojů.

Kapitola 7

Zdroje anotovaných mutací

V této kapitole budou rozebrány možnosti vytvoření trénovacího a testovacího datasetu mutací, který pak bude využit při implementaci prediktoru v praktické části. Nejprve budou zmíněny hlavní faktory, podle kterých jsou obecně trénovací sady vytvářeny. Dále bude uvedeno, které datové zdroje jsou v souvislosti s touto prací k dispozici a jaké varianty se v nich nacházejí.

7.1 Požadavky

Trénovací množina mutací slouží k nastavení parametrů rozhodovacího modelu tak, aby na dalších, nezávislých datech vykazoval co nejvyšší přesnost predikce. Z tohoto důvodu je potřeba na její vybudování klást patřičný důraz a nepodcenit důležité vlastnosti, které by měla splňovat. Hlavní faktory při vytváření datové sady jsou tyto:

- Dataset by měl obsahovat dostatečné množství jak škodlivých, tak neutrálních mutací. Informace o tom, jaký efekt mutace má, je samozřejmě pro trénování nezbytná. Některé zdroje uvádí, že složky škodlivých a neutrálních mutací by měly mít přibližně stejnou mohutnost. Tento fakt je ale přinejmenším diskutabilní, neboť dnešní rozhodovací algoritmy si s tímto problémem obvykle dokáží poradit, navíc někdy je nevyváženost počtu škodlivých a neutrálních mutací v jistém smyslu vyžadována. Naprostá většina nástrojů, které jsou uvedeny v minulé kapitole, má své trénovací sady velmi nevyvážené (co do mohutnosti škodlivé a neutrální složky), přitom jejich predikční síla je neoddiskutovatelná [41].

- Je třeba zajistit, aby byl trénovací množinou co nejlépe a nejrovnoměrněji pokryt celý lidský genom, resp. všechny jeho regiony (exony, introny, regulační oblasti, TSS a ostatní). Jinak by se mohlo stát, že výsledný prediktor bude vykazovat velké výkyvy přesnosti, případně, že některé mutace vůbec nedokáže vyhodnotit.

- V datových sadách by se neměly vyskytovat redundantní údaje, jednak z důvodu zbytečného zvyšování paměťových nároků, zejména však kvůli vlivu na trénování.

- Existují mechanismy, které odstraňují příliš podobné mutace nacházející se blízko sebe a ponechají jen jednu nebo zlomek jejich počtu. Je to způsobeno snahou minimalizovat problém přetrénování. Když se totiž v trénovací množině vyskytuje větší množství velice podobných mutací ve stejném místě, prediktor se je naučí predikovat velice přesně, ovšem na úkor přesnosti predikce na ostatních mutacích. Jedná se však spíše o problémy nástrojů, které se soustředí jen na konkrétní region, například

kódující nesynonymní mutace. V této práci je vytvářen prediktor pro celý genom, proto není třeba na tento aspekt klást takový důraz.

7.2 Zdroje DNA mutací

7.2.1 HGMD

HGMD je v současné době asi nejlepší databáze z hlediska rozsahu dostupných informací o mutacích nacházejících se na lidském genomu. Volně jsou však k dispozici pouze neúplné údaje k mutacím vloženým před více než třemi lety. Pro získání plných údajů je nutné mít zaplacený přístup [49].

7.2.2 OMIM

Databáze OMIM (*Online Mendelian Inheritace in Man*) se specializuje pouze na mutace způsobující Mendelovské choroby, nemá tedy dostatečný počet záznamů pro vytvoření celé trénovací sady. Navíc, díky propojení bioinformatických databází jsou všechna data z OMIM součástí dalších, obecnějších projektů (ClinVar, HGMD, UniProt) [22].

7.2.3 UniProt/SwissProt

Databáze UniProt slouží jako centrální místo pro shromažďování informací o funkci proteinů a obsahuje velké množství přesných anotací. Klade důraz zejména na obsáhlost informace a její relevantní zařazení. Každý záznam obsahuje mimo povinných položek (sekvence aminokyselin, název a popis proteinu, taxonomická data a citace) maximum dostupných anotací k danému proteinu [4]. Databáze Uniprot může i přes své zaměření na proteiny rovněž posloužit jako zdroj anotací DNA, v tomto směru však nedosahuje kvalit specifičtěji zaměřených databází (ClinVar, HGMD). Z tohoto důvodu nebude v této práci použita.

7.2.4 ClinVar

ClinVar je databáze, kterou provozuje vědecká instituce NCBI (*National Center for Biotechnology Information*). Databáze klade velký důraz na kvalitu informací a zajištění údržby. Projekt ClinVar vznikl v roce 2012 a od té doby významně rozšířil a stále rozšiřuje objem poskytovaných dat. Veškerá data jsou k dispozici ke stažení, zároveň jsou přístupná i přes webový portál [27].

Data této databáze jsou silně specifická, soustředí se zejména na zachycení vztahů z hlediska medicíny důležitých mutací a jejich fenotypů. Databáze je úzce propojena s dalšími zdroji, zejména databázemi dbSNP [46], dbVar a MedGen, které jsou rovněž provozovány NCBI. Příspěvky do ClinVar přibývají z více zdrojů, jedním z jejích úkolů je zajištění konzistence dat a zamezení ukládání duplikátních či konfliktních záznamů.

Každý nový záznam musí být opatřen identifikátorem autora, popisem varianty a jejího fenotypického projevu spolu s vysvětlením metody důkazu použité k ověření vykazovaného efektu. Záznamy o mutacích jsou pozičně mapovány do referenčního genomu GrCh37 nebo GrCh38, vyskytují se jak substituce, tak i inzerce a delece. ClinVar pozicím mutací nevytváří vlastní identifikátor, využívá takzvaného RS# čísla, které je uvedeno v databázi dbSNP. Při vložení nového záznamu se nejprve zkontroluje, zda má pozice v dbSNP již nějaký záznam, pokud ne, iniciuje se vložení i do databáze dbSNP a následně se odtud přejímá RS# číslo. Fenotyp může být vložen jako prostý text, ClinVar se ale snaží i tento údaj standardizovat pomocí existujících identifikátorů.

Nové záznamy jsou automaticky vkládány z databází OMIM, GeneReviews a dbSNP. Zároveň ale mohou nové informace vkládat i jednotliví uživatelé. ClinVar ověří formát a zda záznam nepopírá výsledek některého již existujícího. Schválené záznamy jsou integrovány a publikovány vždy s novou verzí [27].

7.2.5 VariSNP

Projekt VariSNP [44] vznikl v roce 2013 a o rok později byl oficiálně publikován. Jedná se o poměrně rozsáhlou databázi konstruovanou přímo k účelu vývoje nových predikčních metod pro stanovení vlivu mutací DNA na lidský organismus.

Hlavní výhodou VariSNP z hlediska této diplomové práce je její vysoce specifický účel přesně odpovídající tématu práce a možnost stažení všech jejích částí zdarma. Nevýhodou je, že obsahuje pouze neutrální mutace.

Primárním zdrojem datasetu VariSNP je databáze dbSNP, z níž jsou odfiltrovány všechny mutace spojené buď s nemocí, nebo škodlivostí z databází ClinVar, Swiss-Prot a Phen-Code [20]. Díky zmíněné vazbě na databázi dbSNP je každá mutace opatřena i RS# číslem. Po odfiltrování škodlivých mutací byly zbylé neutrální varianty (substituce, inzerce i delece) rozděleny do několika souborů podle typů na:

- jednobodové nesynonymní kódující substituce,
- inzerce a delece délky dělitelné třemi (nezpůsobí posun čtecího rámce),
- synonymní kódující mutace,
- intronové varianty,
- mutace v místech sestřihu,
- mutace způsobující vznik nebo zánik terminačního kodonu a další.

Nová verze databáze vychází vždy v reakci na novou verzi dbSNP, ze které čerpá [44].

Kapitola 8

Strojové učení

Tato kapitola se zabývá metodami strojového učení jako možného přístupu k vytváření klasifikačních modelů.

8.1 Úlohy strojového učení

Strojové učení je vědní disciplína, která se zabývá algoritmy, díky kterým se program může učit ze svých vstupních dat. Tyto algoritmy vytváří model, který je schopen díky znalostem ze vstupních dat provádět či podporovat rozhodování. Strojové učení má své využití v řadě matematických problémů, které mohou být rozděleny do tří základních kategorií: [43]

- **Učení s učitelem** (*supervised learning*) se vyznačuje tím, že všechny vzorky vstupních dat jsou doplněny o vzorový výstup.
- **Učení bez učitele** (*unsupervised learning*) zahrnuje problémy, ve kterých je úkolem modelu najít vzory podobností ve vstupních datech.
- **Posilované učení** (*reinforced learning*) je nejobecnější z těchto kategorií. V tomto případě je úkolem programu interakce s dynamickým prostředím, ve kterém má dosáhnout určitého cíle. Není ovšem přítomen učitel, který by explicitně oznamoval, zda se k cíli blíží. Jediné informace (posílení) dostává program právě z interakce s prostředím, např. uvíznutí ve slepé uličce.

Na základě požadovaného výstupu natrénovaného modelu jsou potom uvažovány tyto základní typy úloh: [43]

- **Klasifikací** rozumíme rozdělení vstupní množiny do konečného počtu tříd, její výstup je tedy diskrétní. Prvky trénovací množiny musí být doplněny o třídu, do které patří, jedná se tedy o učení s učitelem.
- **Regrese** je v principu velmi podobná klasifikaci, s tím rozdílem, že výstupem je spojitá hodnota (nekonečně mnoho tříd). Jedná se rovněž o učení s učitelem.
- **Shlukování** (*clustering*) je příkladem učení bez učitele. Úkolem modelu je na základě podobnosti rozřadit vstupní vzorky do konečného počtu shluků tak, aby prvky ve shluku byly vzájemně co nejpodobnější a zároveň, aby prvky v různých shlucích byly od sebe maximálně odlišné.

8.2 Metody klasifikace

Rozhodovací model, konstruovaný v rámci praktické části práce, je zaměřen na klasifikaci mutací, tedy jejich rozdělení na neutrální a škodlivé. K tomuto typu problému je využívána řada metod, které lze seskupit do několika tříd.

8.2.1 Stromové metody

Podstatou těchto metod je konstrukce rozhodovacích stromů z trénovací sady a posléze jejich aplikace na nové vstupy. V průběhu konstrukce stromu je nejdůležitější identifikace takových atributů trénovacích vzorků, které jsou mají velkou rozhodující sílu pro správné určení výsledné třídy. Z těchto znalostí si algoritmy budují kolekci pravidel, kterou uspořádají do stromové struktury. V uzlech stromů jsou umístěny podmínky pro hodnoty atributů, na základě výsledku jejich vyhodnocení se pokračuje některou z navazujících větví. Listy stromu jsou vyplněny identifikátory tříd.

Ve fázi klasifikace (po natrénování) už se strom nemění, na základě hodnot atributů se jím prochází od kořene směrem k listům a výsledek je zařazen do třídy listem určené.

Častou nevýhodou těchto metod je přetrénování. Metody se dobře naučí na trénovací data, nejsou však schopné správně zařazovat jiné obecné vzorky. Tomuto faktu se předchází např. prořezáváním (*prunning*) větví stromu. Jedna z nejpoužívanějších stromových metod, náhodný les, pak proti přetrénování aplikuje ještě další postup. Konstruuje větší množství stromů různě prořezaných a s různým pořadím rozhodovacích podmínek. Jako výsledek pak vydává majoritní volbu nad jejich výstupy.

8.2.2 Neuronové sítě

Neuronové sítě jsou klasifikátory vybudované podle předlohy v lidské nervové soustavě. Jejich základem je model umělého neuronu, tzv. perceptron. Perceptrony mají vektorový váhovaný vstup a číselný výstup (obvykle binární či bipolární) a lze je spojovat a vytvářet sítě.

Algoritmů pro učení perceptronu nebo neuronové perceptronové sítě je větší množství, s různou složitostí a přesností. Většinou jsou založené na opakovaném procházení trénovací sady a upravování vah vstupů perceptronů.

Neuronové sítě se kromě klasifikace používají rovněž při regresi a dále i v některých aplikacích učení bez učitele. Jejich hlavní výhodou je velká univerzálnost v řešení problémů. Navíc modifikací počtu vrstev a perceptronů v nich lze pro většinu problémů najít vhodnou a přesně pracující síť. Nevýhodou je naopak poměrně náročná fáze učení, kde zpravidla nestačí jeden průchod trénovacími daty.

Často používaným algoritmem pro učení samostatného perceptronu je Pocket algoritmus, pro perceptronové sítě se často používá algoritmus zpětného navracení (*Backpropagation*) [43].

8.2.3 Regresní metody

Jak už název napovídá, tyto metody slouží primárně k regresi. Podstatou metody zvané lineární regrese je snaha aproximovat soubor bodů v prostoru přímkou, případně funkcí, která je lineární kombinací nějakých dílčích funkcí, a to metodou nejmenších čtverců. Výstup této metody je numerická hodnota, proto je tedy používaná hlavně pro regresi.

Logistická regrese používá podobný přístup, její výstup ovšem spadá pouze do intervalu <0,1>. Lze ji tedy vnímat jako prostředek klasifikace do dvou tříd, reprezentovaných

krajními hodnotami intervalu. Trénování modelu logistické regrese představuje ladění koeficientů regresní křivky tak, aby co nejlépe oddělovala data jedné třídy od druhé.

Výhodou této metody je, že díky reálnému výstupu lze kromě výsledné třídy zařazení získat i pravděpodobnost správnosti klasifikace. Hodnoty blížící se krajním hodnotám intervalu budou do příslušné třídy zařazeny pravděpodobně správně, zatímco u výsledků kolem hodnoty 0,5 se mohou často vyskytnout chyby. [6, 34].

8.2.4 Bayesovské klasifikátory

Tyto klasifikátory patří mezi tzv. pravděpodobnostní, na základě vstupních parametrů zjistí pravděpodobnost příslušnosti vzorku do jednotlivých tříd. Nejpravděpodobnější výsledek pak považují za správný. Vychází z konceptu aplikace Bayesova teorému o podmíněné pravděpodobnosti dvou jevů. Ve fázi trénování se pro každou třídu vytvoří rozložení hodnot atributů trénovacích vzorků. Při aplikaci modelu se pak vypočítá zmíněná pravděpodobnost náležitosti do tříd podle odchylek atributů zařazovaného vzorku od průměrných hodnot dané třídy.

Naivní bayesovké klasifikátory předpokládají, že všechny atributy, z nichž se skládá trénovací vzorek, jsou vzájemně nezávislé. Obecně tomu však není, proto tyto klasifikátory často nevykazují dostatečnou přesnost. Protipólem k těmto metodám jsou Bayesovské sítě. Jsou to grafické modely, u nichž lze modelovat i závislost mezi atributy trénovacích vzorků a tím přidat důležitou znalost do rozhodovacícho procesu. To vše ale samozřejmě za cenu značného navýšení výpočetní náročnosti [43].

8.2.5 Neparametrické metody

Metody této třídy se vyznačují tím, že pro ně neexistují pevné parametry, které by přímo ovlivnily výsledky trénování a klasifikace. Patří sem například metoda K-nejbližších sousedů, jejíž princip je následující: Pro testovaný vzorek dat je vypočítána jeho vzdálenost od ostatních (trénovacích) vzorků. Po vybrání K nejbližších je zjištěno, která třída má v této množině sousedů nejvyšší zastoupení, ta je pak výstupem klasifikace. Tento model lze rovněž pojmout pravděpodobnostně, pokud k výsledné třídě připojíme ještě podíl počtu vzorků dané třídy a čísla K.

Výhodou těchto metod je, že se velmi rychle učí a pokud jsou atributy trénovacích dat správně zvoleny, vykazují i poměrně vysokou úspěšnost. Hlavním neduhem je zpracování různých typů parametrů a hlavně použití metrik pro vzdálenost. Zatímco pro numerické hodnoty lze ve většině případů efektivně použít euklidovskou vzdálenost, pro kategorické typy se vzdálenostní metrika definuje velmi složitě. V případě většího množství atributů v trénovacím vzorku je pak dalším problémem kombinace dílčích výsledků vzdáleností jednotlivých atributů [34].

8.2.6 Lineární metody s jádrem

Tato skupina metod se, podobně jako např. logistická regrese, snaží definovat lineární rozdělení prostoru trénovacích dat tak, aby vzorky jedné třídy byly odděleny od vzorků druhé třídy. Trénovací data jsou však často lineárně neseparovatelná, proto musí tyto metody obsahovat mechanismus, jak se s tímto vypořádat. Neuronové sítě využívají propojení více perceptronů a váhování jejich vstupů. Lineární metody s jádrem, jejichž nejvýznamnějším zástupcem jsou support vector machines (SVM), používají poněkud jiný přístup.

Cílem SVM je vytvořit mezi vzorky jedné a druhé třídy mezeru, a to co nejširší, aby bylo možné jednoznačně a co nejsprávněji určit, do které třídy daný vzorek patří. Nelinearita

rozložení je překonána namapováním trénovacích vzorků do nového prostoru, obvykle vyšší dimenze, kde už potom lze lineární separovanost identifikovat. Proces transformace se provádí pomocí tzv. jader (*kernels*) a v novém prostoru se pak provádí lineární klasifikace na základě výsledků funkcí těchto jader.

Metody SVM se vyznačují vysokou náročností výpočtů, ovšem díky jejich přesnosti a způsobu použití (integrací více atributů) jsou zejména v bioinformatice často využívány [6].

8.2.7 Konsenzuální přístup ke klasifikaci

V této práci je strojové učení využito k získání konsenzu nad výsledky klasifikace od jednotlivých nástrojů. Obecným cílem konsenzuálního přístupu (*ensemble approach*) je vhodně kombinovat dílčí výsledky za účelem vylepšení vlastností klasifikačního modelu.

V bioinformatice a v odvětví strojového učení se tohoto postupu využívá velmi často, a to na několika úrovních. Tato práce využívá metod strojového učení k nalezení nejvhodnějšího přístupu, jak zkombinovat výsledky dílčích nástrojů, popsaných v předchozích kapitolách. Zároveň se ale konsenzus používá již v rámci implementace některých metod. Například náhodný les (používá jej nástroj GWAVA) produkuje na výstup formu konsenzu výsledků klasifikací svých stromů.

Konsenzuální přístup může výrazně vylepšit celkovou přesnost, robustnost a spolehlivost klasifikace. I když toto není pravidlem, obecně lze říct, že kombinované výsledky z více zdrojů mají pozitivní dopad na kvalitu predikčních modelů [28]. Lze říci, že hlavním úkolem konsenzu je vyhladit výrazné nedostatky dílčích výsledků a naopak nechat eskalovat jejich přesnost.

8.3 Hodnocení přesnosti klasifikace

Praktická část této práce se věnuje klasifikaci mutací do třídy neutrálních, respektive škodlivých mutací. Jedná se tedy o binární klasifikaci. Přesnost klasifikačních metod se v tomto případě posuzuje nejčastěji podle následujících metrik: [28]

- **TPR** (*True Positive Rate*), označováno jako senzitivita, je poměr správně označených škodlivých mutací k počtu všech reálně škodlivých mutací.
- **TNR** (*True Negative Rate*), označováno jako specificita, je poměr správně klasifikovaných neutrálních mutací k počtu všech reálně neutrálních mutací.
- **FNR** (*False Negative Rate*) a **FPR** (*False Positive Rate*) jsou míry chybovosti, které se někdy uvádí místo TPR a TNR. Platí pro ně vztahy

$$FPR = 1 - TNR \tag{1}$$

$$FNR = 1 - TPR \tag{2}$$

- **Normovaná přesnost** klasifikace (*Normalized Accuracy*) je objektivní posouzení úspěšnosti, je definována jako průměr TNR a TPR:

$$ACC = \frac{TNR + TPR}{2} \tag{3}$$

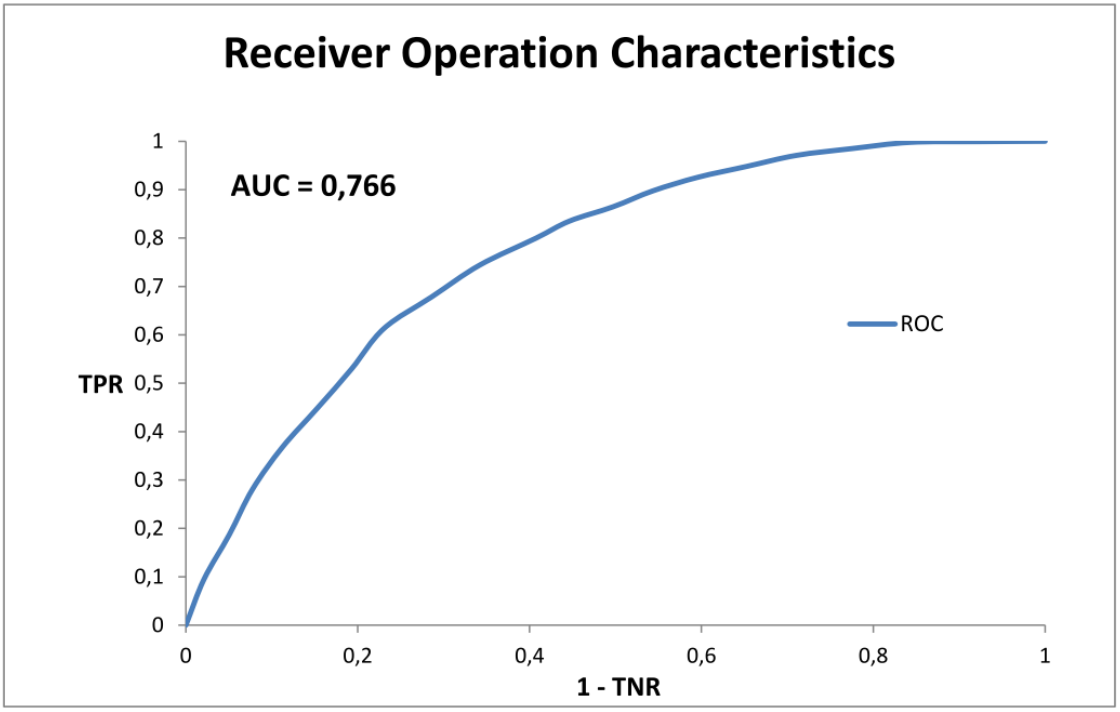
- **MCC** (Matthewsův korelační koeficient) rovněž udává přesnost klasifikátoru, může být i směrodatnější než normovaná přesnost, protože reflektuje i různé kardinality

množin škodlivých a neutrálních mutací. Lze jej vyjádřit pomocí vztahu [37]:

$$MCC = \frac{TPR \cdot TNR - FPR \cdot FNR}{\sqrt{(TPR + FPR) \cdot (TPR + FNR) \cdot (FPR + TNR) \cdot (FNR + TNR)}} \tag{4}$$

8.3.1 ROC charakteristika

Je zřejmé, že v ideálním případě by rozhodovací model měl vykazovat senzitivitu i specificitu rovnu jedné. V praxi tomu tak ale často není, proto se často využívá křivka ROC (*Receiver Operation Characteristics*), která zobrazuje závislost senzitivity (TPR) na opačné hodnotě specificity (1-TNR). Interpretovat ji lze tak, že ukazuje, jak úspěšně klasifikátor odděluje neutrální a škodlivou složku datové sady. Na obrázku 8.1 je příklad ROC charakteristiky pro binární klasifikační problém. Dle předchozích úvah platí, že čím blíže je křivka bodu [0,1], tím lépe klasifikátor odděluje obě třídy. Aby se míra oddělení složek datasetu klasifikátorem dala kvantifikovat a porovnat, často je ROC charaktieristika doplněna o metriku AUC (*Area Under Curve*), tedy plochu pod ROC křivkou [28]. Je zřejmé, že v ideálním případě se AUC blíží jedné. V dalších kapitolách bude uveden postup, jak ROC charakteristiku vybudovat.



Obrázek 8.1: Ukázka průběhu ROC charakteristiky.

8.3.2 K-fold křížová validace

Mimo testování klasifikace na nezávislé testovací sadě je cennou validační informací také křížová validace trénovací sady. Její výsledek má vypovídat o tom, jak bude natrénovaný model úspěšný na nezávislých datech. Validace se provádí následovně:

Trénovací množina je rozdělena do K vzájemně disjunktních množin (tzv. foldů). Poté následuje K učících a validačních kroků, kdy je postupně každý z foldů testovací sadou

a zbylých K-1 foldů tvoří trénovací sadu. Na konci každého validačního kroku jsou vyčísleny metriky TPR a TNR (z výsledků na testovacím foldu). Na závěr se tyto hodnoty průměrují a výsledkem je průměrná TPR, TNR, případně normovaná přesnost. Nejčastěji se (i v závislosti na velikosti trénovací sady) volí K rovno 5 nebo 10. Speciálním případem je tzv. *leave-one-out*, kdy je velikost každého z foldů rovna 1.

Cílem křížové validace je podat kvalitnější informace o trénování. Když není trénovací sada dostatečně obecná a nejsou z ní odstraněny příliš podobné záznamy, klasickým učením může klasifikátor dosáhnout vysoké přesnosti, která ovšem neodpovídá reálné hodnotě. Klasifikace na jiných, nezávislých datech totiž pravděpodobně nebude příliš úspěšná. Rozdělením sady do foldů a průměrováním výsledků učení je tento problém zčásti eliminován a informace o přesnosti tak mají vyšší vypovídací hodnotu.

Ovšem, dle [10] nelze křížovou validací nahradit testování modelu na nezávislé sadě. Výsledky křížové validace jsou totiž proti reálné přesnosti stále nadhodnoceny. Proto je vždy žádoucí využít testovací dataset k objektivnímu posouzení, samozřejmě v případě, že je k dispozici [10].

8.4 WEKA

WEKA (*Waikato Environment for Knowledge Analysis*) [21] je volně dostupný multiplatformní program, který slouží jako nadstavba nad metodami strojového učení. Byl vyvinut na University of Waikato na Novém Zélandu a publikován poprvé v roce 1993. Od té doby prošel řadou změn a verzí, praktická část této práce využívá verzi 3.6.12. Implementačním jazykem je Java.

WEKA poskytuje možnosti řešení široké škály problémů v oblasti získávání znalostí. Zahrnuje implementaci velkého počtu metod strojového učení. Lze jej využít pro klasifikaci, regresi, shlukování, dolování pravidel či výběr atributů. Velkou výhodou je také možnost přidání vlastní knihovny s metodami, které buď WEKA neimplementuje, nebo si uživatel z různých důvodů přeje danou metodu implementovat odlišně. Příkladem je knihovna LibSVM, která implementuje řadu metod založených na SVM.

Po stažení programu a spuštění má uživatel na výběr ze čtyř módů běhu. V souvislosti s aplikací metod strojového učení bude pravděpodobně nejpoužívanější **Explorer**. Dalšími módy jsou:

- **Experimenter**, který je používán k efektivnímu porovnání výsledků různých metod ve WEKA,
- **KnowledgeFlow**, kde si lze vlastní predikční model vytvořit v grafickém prostředí pomocí spojování jednotlivých modulů (předzpracování, trénování, testování),
- **Simple CLI**, což je jednoduché konzolové rozhraní pro ovládání WEKA.

8.4.1 WEKA Explorer

Tento mód spuštění je ideální k ručnímu provádění experimentů, na jeho použití budou nyní nastíněny možnosti WEKA zejména z hlediska klasifikace, která je předmětem praktické části práce.

První záložka Exploreru má název *Preprocess*, slouží k načtení dat a jejich předzpracování. Soubor s daty lze nahrát buď z lokálního úložiště, databáze, nebo z webu. Podporované jsou soubory ve formátu ARFF, který je velmi jednoduchý a bude podrobněji rozebrán dále. Hlavní funkční náplní předzpracování je ale možnost aplikace různých filtrů.

Tyto filtry mohou sloužit např. k výběru atributů ke klasifikaci, diskretizaci nebo dalším formám převodů. Po načtení a předzpracování dat si pak uživatel dle charakteru úlohy vybere jednu z dalších záložek.

Záložka *Classify* je určena, jak již napovídá její název, ke klasifikaci. Nejprve uživatel z nabídky vybere příslušný klasifikátor, který se má stát rozhodovacím modelem. Každý klasifikátor má své vlastní parametry, které mají nastaveny nějaké výchozí hodnoty. Lze je ale samozřejmě libovolně měnit. Poté už stačí spustit trénování. Po skončení procesu, jehož průběh lze sledovat ve stavovém řádku, jsou zobrazeny výsledky, jak se podařilo klasifikátor natrénovat. K hlavním výstupům patří informace o tom, kolik trénovacích vzorků bylo správně zařazeno, kolik z jednotlivých tříd a podobně. Zároveň s trénováním je možné zvolit i formu validace modelu, například použití křížové validace nebo přiložení vlastní testovací sady. V tabulce 8.1 jsou uvedeny klasifikační metody, které WEKA implementuje a rozděluje do tříd dle jejich charakteru.

Třída algoritmů	**Implementované metody strojového učení**
Bayes	BayesNet, NaiveBayes, NaiveBayesMultinomial
Functions	SMO, LibSVM_Linear / Polynomial, LogisticRegression, VotedPerceptron, MultilayerPerceptron
Lazy	IBK, KStar, LBR, LWL
Meta	Bagging, Stacking, Vote, AdaBoostM1
Rules	DecisionTable, ZeroR, PART, JRip, OneR
Trees	RandomForest, RandomTree, REPTree, J48, LMT

Tabulka 8.1: Příklady metod implementovaných programem WEKA.

Záložka *Cluster* slouží pro aplikaci metod shlukování. Ovládání je stejné jako u klasifikace, nelze ovšem samozřejmě provádět křížovou validaci. Dále je přítomna záložka *Associate* pro dolování pravidel, *Select Attributes* pro analýzu informačního přínosu atributů a *Visualize* pro vizualizaci některých aspektů vstupních dat.

ARFF formát pro klasifikaci

Vstupní soubory s daty pro klasifikaci musí být v souboru formátu ARFF. První řádek takového souboru definuje název výpočetní relace:

```
@relation NazevRelace
```

Na dalších řádcích jsou definovány atributy, a to jak vstupní, tak klasifikační, včetně datových typů. Pro klasifikaci na základě tří nástrojů je pokračování souboru:

```
@attribute cadd numeric
@attribute fathmm numeric
@attribute gwava numeric
@attribute effect {n,d}
```

Poslední atribut je vždy diskrétní (v příkladě atribut *effect*) a představuje množinu tříd, do níž budou data rozřazována. Ve vstupním souboru poté následuje řádek

```
@data
```

následovaný trénovacími daty. Každý trénovací vzorek je na jednom řádku, hodnoty atributů v daném pořadí jsou odděleny čárkou, tedy například:

```
18.57,0.97019,0.81,n
38.92,0.68298,0.24,d
```

8.4.2 Java API

Vzhledem k obsáhlosti použití WEKA (množství metod, více datasetů) je mnohdy použití GUI Explorer nevhodné a zdlouhavé. WEKA proto nabízí i Java rozhraní, díky kterému lze velmi efektivně a automaticky trénovat více klasifikátorů přímo v Java projektu a výsledky pak ukládat do souborů a porovnávat. Přístup k rozhraní je intuitivní a lze jej nastudovat dle přehledného návodu na webových stránkách WEKA. Pro účely trénování modelů v praktické části bylo využito jen těchto několik funkcí:

```
//načtení ARFF souboru
Instances data = new Instances(reader);

//nastaví výstupní atribut (poslední v řádku)
data.setClassIndex(data.numAttributes() - 1);

//vytvoří klasifikátor Random forest (náhodný les)
Classifier cls = new RandomForest();

//spustí trénování na vytvořeném datasetu
cls.buildClassifier(data);

//validační objekt
Evaluation eval = new Evaluation(data);
//cross validace - lze nastavit počet foldů a seed pro náhodné
//rozdělování do foldů
eval.crossValidateModel(cls, data, FoldCount, new Random(1))
```

Kapitola 9

Implementace

Tato kapitola je věnována samotné implementaci nového klasifikátoru mutací v lidské DNA. Při vytváření prediktoru jsem vycházel z teoretického základu, který je uveden v předešlých kapitolách. Implementační část byla rozdělena do několika fází, které budou dále rozebrány. Jsou to:

- extrakce dat z databází a vytvoření datasetu,
- instalace predikčních nástrojů a jejich aplikace na extrahované mutace,
- návrh výpočetních modelů,
- trénování klasifikátorů,
- implementace webového rozhraní.

Pro extrakci dat a přípravu vstupních souborů pro nástroje byl použit skriptovací jazyk Python. V některých úlohách bylo zapotřebí použít externí moduly, které nejsou součástí standardního balíčku jazyka. Informace o mutacích z trénovací sady včetně výsledků predikcí nástrojů jsem uchovával v MySQL databázi na webu, díky tomu bylo i velké množství dat snadno dotazovatelné. V jazyce Python také existuje podpora pro komunikaci s MySQL databází, čehož jsem hojně využíval, navíc byla data dostupná z více pracovišť, což bylo při vykonávání náročných výpočtů klíčové.

9.1 Sběr dat a konstrukce trénovací sady

Ze zdrojů, které byly popsány v kapitole 7.2, jsem nakonec vybral databáze ClinVar pro škodlivé a VariSNP pro neutrální mutace.

9.1.1 Škodlivé mutace

Obsah databáze ClinVar je dostupný zdarma z FTP serveru. V tabulce A.1 je uvedeno schéma této databáze včetně popisu sloupců. Algoritmus výběru mutací, které splňují podmínky pro zařazení do trénovací sady, se skládal z těchto klíčových kroků, které byly aplikovány zvlášť na každý řádek (jeden záznam o mutaci):

- Proběhla kontrola sloupce *Assembly*, v práci pracuji pouze s referenčním genomem GRCh37. Každá uložená mutace je v databázi totiž uložena dvakrát, jednou se souřadnicemi pro GRCh37, jednou pro GRCh38.

- Sloupec *ClinicalSignificance* musí obsahovat klíčové slovo *pathogenic* (škodlivý) a zároveň neobsahovat *benign* (neutrální). Některé záznamy totiž mají uloženo více anotací, proto ty, které nebyly jednoznačné, byly ihned vyřazeny.
- Poté už byly jen extrahovány informace o mutacích. Ze sloupců *Chromosome*, *Start*, *Stop* jsem získal souřadnice mutace, referenční a novou alelu ze záznamů *Name*. V tomto sloupci se totiž obvykle nacházel identifikátor mutace v databázi RefSeq (*NCBI Reference Sequence Database*), který vypadá například takto:

  ```
  NM_000410.3(HFE):c.175G>A (p.Val59Met) .
  ```

  Písmeno **M** v názvu značí, že se jedná o transkript DNA (tedy mediátorová RNA), číslo **3** za tečkou označuje chromozom číslo 3 a nejdůležitější částí pro extrakci je řetězec **G>A**, který říká, že tato mutace mění guanin na adenin. Do databáze jsem dále ukládal typ mutace (substituce, inzerce, delece), který jsem zjistil právě na základě referenční a nové alely, a identifikátor mutace v databázi dbSNP ze sloupce *RS#*. Tuto hodnotu bylo nutné zjistit pro maximální počet mutací, neboť je součástí vstupních dat pro nástroj GWAVA.

V případě, že došlo u některého záznamu k další nejednoznačnosti, kvůli které nebylo možné extrahovat přesné informace, mutaci jsem do databáze nevkládal. Z původních 266 296 záznamů (případně 133 148, pokud uvažuji pouze GRCh37) jsem nakonec vybral 27 140. Tento počet zohledňuje i ověření přítomnosti referenčních alel na uvedené pozici. Tato kontrola ale proběhla později při ohodnocení datasetu programem CADD.

9.1.2 Neutrální mutace

Databáze VariSNP poskytuje velké množství mutací, které jsou všechny označeny za neutrální. Z webových stránek je možné stáhnout celkem 13 souborů ve formátu CSV (*Comma-Separated Values*), přičemž na každém řádku se nachází alespoň jedna mutace. Schéma databáze je k dispozici v příloze v tabulce A.2. V každém souboru jsou mutace jiného typu (substituce, inzerce, delece) a genomického regionu (exon, intron, místa sestřihu a další). Pro zajištění konzistence se škodlivou složkou jsem typ mutace opět získal až z příslušných alel a region poté za pomoci externího programu, takže nebylo třeba se rozdělením do souborů zabývat. U VariSNP záznamů není uvedena verze referenčního genomu, dle nápovědy se ovšem jedná vždy o GRCh37. Zda odpovídají pozice a referenční alely mutací bylo ověřeno později.

Do výstupního souboru neutrálních mutací byly vloženy všechny, u kterých bylo možné získat potřebné informace (jejich formát vyhovoval) a zároveň se již nevyskytovaly v množině škodlivých mutací. Překryv obou složek by byl totiž zcela nežádoucí.

Při zpracovávání VariSNP záznamů bylo nejvýhodnější postupovat podle sloupce *HGVS names*. Každý řádek totiž může obsahovat více mutací a sloupec *HGVS names* obsahuje identifikátory všech. Na rozdíl od ClinVar jsou zde uvedeny identifikátory **NC**, které mají podobnou strukturu jako **NM**. Znak C značí, že se jedná o referenční, kompletně osekvenovaný genom. **NC** identifikátor substituce má tvar:

```
NC_000002.11:g.12881526T>G
```

Číslo **11** za tečkou značí chromozom, znak **g** značí, že se jedná o genomovou sekvenci, nikoli proteinovou, dále následuje pozice mutace a informace, že jde o záměnu thyminu za guanin. Inzerce je zapsána ve tvaru:

`NC_000007.13:g.99250393_99250394insA`

Rozdíl je pouze v tom, že jsou uvedeny dva údaje o pozici (mezi tyto dva nukleotidy je vložena sekvence), následované klíčovým řetězcem **ins** a poté vkládaným řetězcem nukleotidů. Delece a delece s inzercí mají identifikátory obdobné. Ze sloupce *dbSNP_id* lze dále získat RS# identifikátor mutace, ve sloupci *observed_alleles* jsou všechny alely, které se mohou na dané pozici vyskytovat (zde lze provést kontrolu s přítomnými **NC** identifikátory).

Z původních 5 886 724 mutací se podařilo získat 5 784 332 vyhovujících.

9.2 Ohodnocení datasetu nástroji

Při ohodnocení datasetu jsem vycházel z tabulky 6.2, dle které jsem každým nástrojem klasifikoval jen ty mutace, které ve své publikaci deklaruje jako přípustné. Tím došlo k významné úspoře výpočetního času. Výsledky nástrojů byly získány třemi způsoby.

9.2.1 Využití předpočítaných skóre

Nástroje CADD, FATHMM a částečně i GWAVA poskytují předpočítaná skóre pro mutace. Z jejich serverů lze stáhnout příslušné soubory ve formátu BED. V těchto souborech jsou na každém řádku informace k jedné mutaci, přičemž jednotlivé sloupce jsou vzájemně oddělené tabulátorem. Na obrázku 9.1 je vidět příklad mutace pro všechny tři nástroje. Vždy se vyskytuje číslo chromozomu, pozice, u CADD a FATHMM je mutace specifikována referenční a alternativní alelou, GWAVA pracuje s RS# identifikátory. Vzhledem k tomu, že CADD zvládá predikci všech možných substitucí na celém genomu, byl využit zároveň jako kontrola pro referenční alely mutací z trénovací sady, zda se opravdu nachází na příslušné pozici. Pokud alely nesouhlasily, záznam byl z trénovací sady vyřazen.

CADD

Chr	Pos	Ref	New	RawScore	ScaledScore
1	69428	T	G	0.84659	9.627

FATHMM

Chr	Pos	Ref	New	Coding	Groups	Noncoding	Groups
1	91645	A	C	0,18343	ALL	0,01973	ABCDE

GWAVA

Chr	Start	End	RS	regionScore	tss	unmatched
1	74225	74226	rs46486	0.19	0.45	0.42

Obrázek 9.1: Příklady mutací včetně skóre v BED souborech.

Výstup nástroje CADD je *ScaledScore*, které nabývá hodnot <0,99>. Dle publikace k nástroji CADD ale neexistuje jednotný dělící práh tohoto skóre, který by odděloval neutrální od škodlivých mutací. Pravděpodobně by jej šlo přibližně stanovit pro podmnožiny mutací ve stejném genomickém regionu.

FATHMM poskytuje pro každou mutaci skóre od obou svých klasifikátorů (pro mutace v kódující a nekódující části), do databáze jsem uložil tedy obě hodnoty a až posléze na základě zjištěného regionu jsem vybral správnou hodnotu.

GWAVA rovněž poskytuje výsledky všech svých predikčních modelů, opět jsem ukládal všechny tři výsledky. Na základě úspěšnosti na trénovací sadě jsem pak pracoval s modelem *region*. GWAVA i FATHMM produkují výsledné skóre z intervalu <0,1>, kde hodnoty menší než 0,5 označují neutrální mutace a hodnoty větší nebo rovno 0,5 naopak škodlivé.

Tabix

K rychlému vyhledávání ohodnocených mutací v rozsáhlých BED souborech jsem použil program tabix [29], který lze importovat jako modul do skriptu v jazyce Python. Tabix si vytvoří indexovací strom nad čísly pozic a chromozomů v seřazeném BED souboru a pro zadané rozpětí pozic vrátí velmi rychle všechny nalezené řádky, které vyhovují.

9.2.2 Instalované nástroje

Nástroje GWAVA a SIFT jsou dostupné ke stažení a lze je nainstalovat do lokálního úložiště a poté spustit nad množinou mutací. Při zprovoznění je nutné postupovat dle návodu, obvykle je nutné jako prerekvizitu nainstalovat doplňující balíčky. Například GWAVA pro svůj běh vyžaduje knihovny Pythonu *tabix*, *numpy*, *scipy*, *pandas*, *scikit-learn* a *pybedtools*.

Vstup pro GWAVA je TSV (*Tab-Separated Values*) soubor se čtyřmi sloupci, kde každá mutace je na jednom řádku, který obsahuje číslo chromozomu, počáteční a koncovou pozici a RS# identifikátor mutace, např.

```
chr1 889158 889158 rs13303056
```

Na výstupu je potom rovněž TSV soubor, který má stejné řádky jako vstupní soubor, ovšem doplněné o výsledky tří klasifikačních modelů nástroje GWAVA.

Vstupní formát programu SIFT je CSV soubor, kde na řádku je vždy číslo chromozomu, počáteční a koncová pozice mutace, orientace vlákna a referenční a alternativní alela. Mutace na přímém vlákně vypadá takto:

```
22,30163532,30163533,1,A/C
```

Na reverzním vlákně by byla odpovídající hodnota rovna -1. Výstup nástroje SIFT je TSV soubor, ve kterém je poměrně hodně sloupců. Podstatný je sloupec *Score*, ve kterém je číselná hodnota z intervalu <0,1>, přičemž hodnoty menší než nebo rovny 0,05 znamenají škodlivou mutaci, zatímco hodnoty vyšší než 0,05 mutaci neutrální.

9.2.3 Webové prediktory

Nástroje CADD a MutationTaster2 zprostředkovávají své výstupy mimo jiné přes webový formulář. Zatímco u nástroje CADD jsem si vystačil s předpočítanými výsledky, nástroj MutationTaster2 bylo nutné dotázat přes jeho rozhraní na webovém serveru. Rychlost je přijatelná pro několik málo mutací, pro větší množství je časově již velmi náročná. Proto jsem z trénovací množiny vybral jen ty mutace, které leží v kódující části DNA, případně v místech sestřihu. MutationTaster2 totiž mutace v ostatních regionech neumí predikovat.

Vstupem pro klasifikační webovou službu nástroje je soubor s mutacemi ve formátu VCF [16](*Variant Call Format*), což je specifický devítisloupcový TSV formát používaný výhradně pro zápis mutací. Jeho popis je rozsáhlý, nicméně pro zisk výsledku klasifikace MutationTaster2 stačí vyplnit prvních pět sloupce podle následujícího příkladu:

```
11 126145284 rs267606829 C T . . . . .
```

Je nutné uvést číslo chromozomu, pozici, RS# identifikátor, referenční a alternativní alelu. Výstupem je stejně jako u SIFT TSV soubor s větším množstvím sloupců. Sloupec *pred_index* obsahuje buď řetězec *polymorphism* pro neutrální, nebo *disease_causing* pro škodlivé mutace, Sloupec *probability* pak udává pravděpodobnost správnosti klasifikace.

9.3 Návrh klasifikačního modelu

Tato podkapitola je věnována designu výpočetního modelu, který bude provádět klasifikaci. Z tabulky 6.2 je zřejmé, že nelze předpokládat plné pokrytí celé trénovací sady všemi nástroji. Některé totiž predikují efekt jen pro určité typy mutací, a to ještě jen z určitých genomických regionů. Bylo proto nutné dodat k mutacím informace o těchto regionech a znovu analyzovat výsledky.

9.3.1 ANNOVAR

Program ANNOVAR [54] je pro studijní a výzkumné účely volně dostupný nástroj, který se specializuje na různé druhy anotací míst a mutací v DNA. Je k dispozici ke stažení a lokální instalaci, zároveň je možné využít i jeho webového rozhraní wANNOVAR.

V této práci jsem ANNOVAR použil pro identifikaci genomických regionů trénovacích mutací. Vzhledem k náročnosti výpočtu jsem se rozhodl pro instalaci nástroje na lokální počítač, tento přístup pak bude rovněž použit u výsledného klasifikátoru. K danému úkolu je v rámci ANNOVAR k dispozici skript *table_annovar*. Ten ke svému běhu vyžaduje stažení několika databází anotací, které poté přiřazuje k mutacím ve vstupním souboru.

Vstup programu je TSV soubor s mutacemi, kde každý řádek obsahuje opět chromozom, pozici, referenční a alternativní alelu:

```
13  20763686  20763686  G  -
```

Výstup programu ANNOVAR je tabulkový CSV soubor, který vstupní data doplňuje o anotace z databází zvolených při spuštění. Pro zařazení mutace do regionu je klíčový sloupec *Func* z databáze *refGene* a dále *ExonicFunc* z téže databáze. Z dalších databází lze rovněž získat cenné informace, například z databáze *snp138* lze doplnit chybějící RS# identifikátory, databáze *ljb26* pak poskytuje k některým mutacím skóre nástrojů MutationTaster2 a CADD.

9.3.2 Genomické regiony a jejich pokrytí nástroji

V následujících tabulkách je číselně vyjádřeno, kolik trénovacích mutací z daného regionu který nástroj dokázal ohodnotit. Tabulka na obrázku 9.4 ukazuje souhrnné pokrytí celého datasetu na různých typech mutací. Obrázek 9.2 zobrazuje počty mutací pro různé regiony. V tabulce na obrázku 9.3 se nachází pouze ty mutace, které jsou z regionu **exonic**. Tyto mutace jsou dále rozděleny podle sloupce *ExonicFunc*.

Func	typ	efekt	celkem	GWAVA	CADD	MT2	FATHMM	SIFT
UTR3	SNV	d	17	17	17	17	17	0
		n	105597	105417	105597	95843	105581	466
	deletion	d	4	0	4	4	0	0
		n	296	0	296	278	0	0
	indel	n	3	0	3	2	0	0
	insertion	n	10	0	10	7	0	0
UTR5	SNV	d	36	33	36	36	36	0
		n	9663	9663	9663	9128	9663	94
	deletion	d	4	0	4	2	0	0
		n	34	0	34	31	0	0
UTR5;UTR3	SNV	d	1	0	1	1	1	0
		n	56	56	56	54	56	0
	deletion	n	2	0	2	2	0	0
downstream	SNV	d	1	1	1	1	1	1
		n	42508	42508	42508	4762	42500	31
	deletion	n	88	0	88	9	0	0
	insertion	n	1	0	1	0	0	0
intergenic	SNV	d	31	17	12	2	12	0
		n	218667	215094	217627	76679	217598	1956
	deletion	d	3	0	3	0	0	0
		n	244	0	244	77	0	0
	indel	n	1	0	1	0	0	0
intronic	SNV	d	415	339	413	7	413	8
		n	4713016	4706935	4712967	2876	4712198	3216
	deletion	d	24	0	24	0	0	0
		n	6049	0	6049	0	0	0
	indel	d	1	0	1	0	0	0
		n	83	0	74	0	0	0
	insertion	n	110	0	110	0	0	0
ncRNA	SNV	d	39	29	32	2	32	0
		n	442162	441591	442159	14074	441889	1243
	deletion	d	10	0	10	0	0	0
		n	529	0	529	25	0	0
	indel	d	2	0	2	0	0	0
		n	9	0	8	0	0	0
	insertion	n	12	0	12	1	0	0
splicing	SNV	d	1470	1215	1470	1467	1470	22
		n	573	573	573	560	573	17
	deletion	d	52	0	52	51	0	0
	indel	d	6	0	6	6	0	0
upstream	SNV	d	37	37	37	22	37	0
		n	70573	70573	70572	7643	70565	81
	deletion	d	2	0	2	0	0	0
		n	167	1	167	25	0	0
	indel	d	1	0	1	0	0	0
		n	4	0	3	1	0	0
	insertion	n	5	0	5	0	0	0
upstream downstream	SNV	n	3030	3030	3030	308	3030	5
	deletion	n	10	0	10	1	0	0

Obrázek 9.2: Pokrytí jednotlivých genomických regionů (zkratky: MT2 - MutationTaster2, efekt „d“ - deleterious /škodlivý efekt, efekt „n“ - neutral /neutrální efekt).

exonic	ex. func	typ	efekt	celkem	GWAVA	CADD	MT2	FATHMM	SIFT
exonic	frameshift deletion	deletion	d	3511	1	3511	3468	0	0
			n	5	0	5	5	0	0
	frameshift substitution	indel	d	144	0	142	131	0	0
	nonframeshift deletion	deletion	d	364	0	364	325	0	0
			n	12	0	12	11	0	0
	nonframeshift insertion	insertion	d	1	0	1	1	0	0
	nonframeshift substitution	indel	d	126	0	126	121	0	0
	nonsynonymous SNV	SNV	d	15811	14859	15796	15773	15796	15494
			n	78605	78354	78605	78060	78600	74716
	stopgain	SNV	d	4696	3630	4693	4679	4693	4689
			n	1253	1253	1253	1236	1253	1217
		deletion	d	84	0	84	84	0	0
	stoploss	SNV	d	55	51	54	54	54	2
			n	94	94	94	93	94	0
		deletion	d	1	0	1	1	0	0
	synonymous SNV	SNV	d	128	114	126	126	126	3
			n	89977	89627	89977	89440	89965	237
	unknown	SNV	d	58	33	58	41	58	51
			n	881	877	881	459	881	260
		deletion	d	5	0	5	1	0	0
			n	1	0	1	1	0	0

Obrázek 9.3: Pokrytí různých exonických kategorií.

typ	efekt	celkem	GWAVA	CADD	MT2	FATHMM	SIFT
SNV	d	22795	20375	22746	22228	22746	20270
	n	5776657	5765646	5775564	381215	5774448	83540
deletion	d	4064	1	4064	3936	0	3940
	n	7437	1	7437	465	0	25
indel	d	280	0	278	258	0	261
	n	100	0	89	3	0	0
insertion	d	1	0	1	1	0	1
	n	138	0	138	8	0	0

Obrázek 9.4: Pokrytí trénovacího datasetu nástroji.

Jednotlivé regiony jsou programem ANNOVAR přiřazovány v tomto pořadí a jejich význam je následující:

- **Exonic** – mutace zasahuje do exonu,
- **splicing** – mutace zasahuje do 2 bp okolí místa sestřihu,
- **ncRNA** – mutace zasahuje do transkriptu bez kódující anotace,
- **UTR5** – mutace je v 5' oblasti nepodléhající translaci[1],
- **UTR3** – mutace je v 3' oblasti nepodléhající translaci[2],
- **intronic** – mutace zasahuje do intronu,
- **upstream** – mutace zasahuje do oblasti 1000 bp před TSS,

[1] Jedná se o oblast bezprostředně před iniciačním kodonem, která je významná především pro regulaci translace již vytvořeného transkriptu.

[2] Jedná se o oblast hned za terminačním kodonem, obsahuje regulační sekvence pro genovou expresi.

- **downstream** – mutace zasahuje do oblasti 1000 bp za TES (*Transcription End Site*),
- **intergenic** – mutace je v mezigenovém prostoru.

Význam hodnot sloupce *ExonicFunc*, který rozděluje exonické mutace na další kategorie, je potom následující:

- **Frameshift deletion** – delece v protein–kódující části genomu způsobující posun čtecího rámce,
- **frameshift substitution** – delece následovaná inzercí v protein–kódující části genomu způsobující posun čtecího rámce,
- **nonframeshift deletion** – delece v protein–kódující části genomu nezpůsobující posun čtecího rámce,
- **nonframeshift insertion** – inzerce v protein–kódující části genomu nezpůsobující posun čtecího rámce,
- **nonframeshift substitution** – delece následovaná inzercí v protein–kódující části genomu nezpůsobující posun čtecího rámce,
- **nonsynonymous SNV** – nesynonymní jednobodová substituce,
- **stopgain** – nesynonymní jednobodová mutace způsobující předčasné ukončení translace (vznikne terminační kodon),
- **stoploss** – nesynonymní jednobodová mutace způsobující ztrátu terminačního kodonu,
- **synonymous SNV** – synonymní jednobodová substituce,
- **unknown** – ANNOVAR nedokázal určit.

Na základě informací o pokrytí a znalostí o regionech jsem kategorie slučoval a vytvořil sedm skupin, kdy ke každé bude poté vytvořen nezávislý klasifikační model. Tento způsob jsem zvolil proto, že mutace v odlišných regionech mají různý charakter a typy anotací, takže i predikční nástroje tvořící konsenzus k nim přistupují odlišně. Pokud bych vytvořil pouze jeden klasifikátor pro všechny typy mutací, ztratila by se část predikční síly nástrojů, které rovněž pro rozdílné regiony pracují různě přesně. V tabulce 9.1 je uvedeno všech sedm klasifikačních modelů včetně genomických regionů a typů mutací, na kterých budou natrénovány. Z tabulky lze dále vyčíst, které nástroje budou součástí konsenzu pro daný model. Pro každý z modelů musí platit, že bude sloužit pro predikci efektu mutací, které se vzájemně nebudou příliš lišit a zároveň jich bude dostatek pro trénovací fázi.

Modely (a jejich trénovací datasety) byly navrhovány tak, aby byl vždy do konsenzu zařazen maximální počet nástrojů, ale zároveň tak, aby žádný z těchto nástrojů příliš nesnižoval velikost trénovací množiny (tím, že nedokáže ohodnotit dostatek mutací). Dle tabulky na obrázku 9.3 je patrné, že součástí konsenzu v modelu 7 mohl být i nástroj SIFT, neboť ohodnotí také téměř všechny příslušné substituce. Po analýze výsledků bylo ale zjištěno, že SIFT označuje všechny mutace měnící délku vznikajícího proteinu rovnou za škodlivé, proto byl z modelu 7 dodatečně vyřazen.

Model	Region	Typ	Nástroje
1	intronic	SNV	CADD, GWAVA, FATHMM
2	splicing	SNV	CADD, GWAVA, FATHMM, MT2
3	nonexonic[3]	SNV	CADD, GWAVA, FATHMM
4	nonexonic	Indel	CADD
5	exonic nonsynonymous	SNV	CADD, GWAVA, FATHMM, MT2, SIFT
6	exonic synonymous	SNV	CADD, GWAVA, FATHMM, MT2
7	stopgain, stoploss	all	CADD, GWAVA, FATHMM, MT2

Tabulka 9.1: Rozdělení trénovacího datasetu a zastoupení nástrojů.

9.3.3 Finální trénovací datasety

Finální trénovací množiny pro jednotlivé predikční modely se skládaly vždy pouze z těch mutací, které patřily do příslušného regionu a byly k nim k dispozici predikční skóre všech zúčastněných nástrojů. Tabulka 9.2 ukazuje velikosti datasetů (jejich škodlivé a neutrální složky).

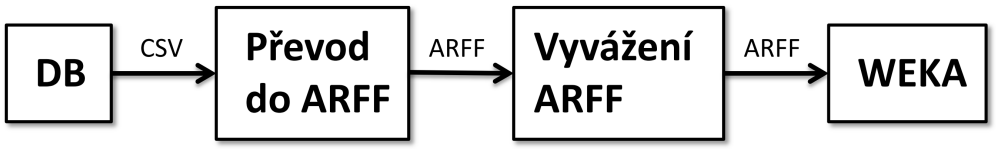
Dataset	Velikost	Škodlivé	Neutrální
1	4 712 303	322	4 711 981
2	1 705	1 145	560
3	5 604 639	1 227	5 603 412
4	7 755	109	7 646
5	88 217	14 126	74 091
6	88 762	109	88 653
7	5 461	4 133	1 328

Tabulka 9.2: Trénovací datasety a jejich velikost.

9.4 Trénování modelů

Jak již bylo uvedeno v předchozích kapitolách, k natrénování klasifikačních modelů byly využity metody strojového učení implementované programem WEKA. Vzhledem k množství těchto metod a jejich parametrů jsem se rozhodl využít funkce Java API, které byly popsány v kapitole 8.4.2. Vstupem programu WEKA jsou datasety ve formátu ARFF. Transformace do tohoto formátu proběhla ve dvou fázích. Nejprve byly převedeny CSV soubory s datasety získané dotazem do databáze do formátu ARFF, poté bylo provedeno vyrovnání velikostí neutrálních a škodlivých složek datasetů. Některé metody se totiž s různými kardinalitami těchto složek dokáží vyrovnat, jiné ovšem svou predikční sílu ztrácí. Balancování obou podmnožin probíhalo tak, že složka menší kardinality zůstala nezměněna a z té druhé byl náhodně vybrán počet vzorků odpovídající rozdílu velikosti složek. Na závěr bylo nutné ještě řádky s daty v ARFF náhodně zamíchat, protože výsledky některých metod strojového učení jsou značně ovlivněny tím, když jim nejprve přichází pouze vzorky jedné třídy a až poté vzorky třídy druhé. Pro přehlednost je celý proces naznačen na obrázku 9.5.

[3]Mutace ze všech regionů mimo exonic



Obrázek 9.5: Vytvoření ARFF souborů pro trénování.

Připravené datasety už bylo možné použít pro trénování. Pro tyto účely byl vytvořen Java projekt, do kterého byly přiloženy knihovny pro program WEKA a dále knihovna LibSVM, která v prostředí WEKA zpřístupňuje metody SVM různých druhů. Níže uvádím pseudokód postupu při trénování s Java API pro WEKA. Tato implementační část byla výpočetně poměrně náročná, nicméně s využitím služeb externího výpočetního střediska MetaCentrum[4] se podařilo vyzkoušet velké množství metod, z nichž jsem následně vybíral ty nejúspěšnější.

9.4.1 Trénovací procedura

```
1. for dataset in datasety:
2.   data = načti_arff(dataset)
3.   for metoda in metody:
4.     vytvoř_klasifikátor(metoda)
5.     natrénuj_klasifikátor(data)
6.     ulož_klasifikátor()
7.     cross_validace_natrénovaného_klasifikátoru()
8.     vytvoř_ROC_křivku_klasifikátoru()
9.     zapiš_výsledky()
```

9.4.2 Konstrukce ROC

V kapitole 8.3.1 byla uvedena ROC charakteristika a plocha pod ní jako jedny z hlavních metrik pro vizualizaci a kvantifikaci výsledků klasifikace. ROC křivka zobrazuje závislost senzitivity na opačné hodnotě specificity. Výstup programu WEKA (cross validace) však udává pouze konečné hodnoty metrik TP, TN, FP, FN, z nichž lze získat jen jednu hodnotu TPR, respektive FPR, tedy jeden bod ROC křivky. Pro vytvoření celé křivky bylo tedy nutné provést ruční 10–fold křížovou validaci a výsledky si průběžně ukládat.

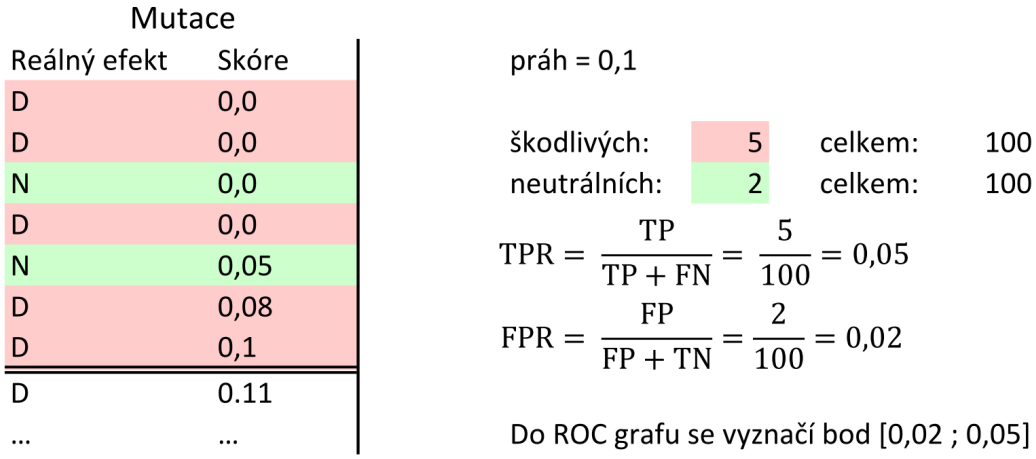
K problému jsem přistoupil tak, že jsem si nejdříve pro každý dataset vytvořil příslušné trénovací a testovací množiny (složené z foldů). Krok č. 8 z pseudokódu trénovací procedury v kapitole 9.4.1 jsem pak provedl takto:

[4]MetaCentrum je organizace poskytující paralelní výpočetní prostředí pro náročné úlohy. Jedná se výhradně o stroje s unixovými platformami a lze se na ně připojit vzdáleně přes porotokol SSH. Pro studijní a nekomerční účely jsou služby MetaCentra zdarma.

```
1. for i = 1..10: //10 foldů
2.   data_train = načti_data(dataset_bez_foldu_i)
3.   data_test = načti_data(dataset_fold_i)
4.   vytvoř klasifikátor(metoda)
5.   cls = natrénuj_klasifikátor(data_train)
6.   for mutace in data_test:
7.     cls.Klasifikuj(mutace)
8.     ulož_mutaci_a_skóre() //skóre na škále <0,1>
9. vytvoř_ROC()
```

Výstup cyklu je seznam všech mutací trénovacího datasetu (každá mutace byla dle principu křížové validace v některém testovacím foldu) včetně jejich **reálného** efektu a skóre, které jí přiřadil natrénovaný klasifikátor. Seznam mutací se seřadí podle tohoto skóre a potom se pro různé hodnoty rozhodovacího prahu počítají metriky TPR a FPR, které se vynesou do ROC grafu.

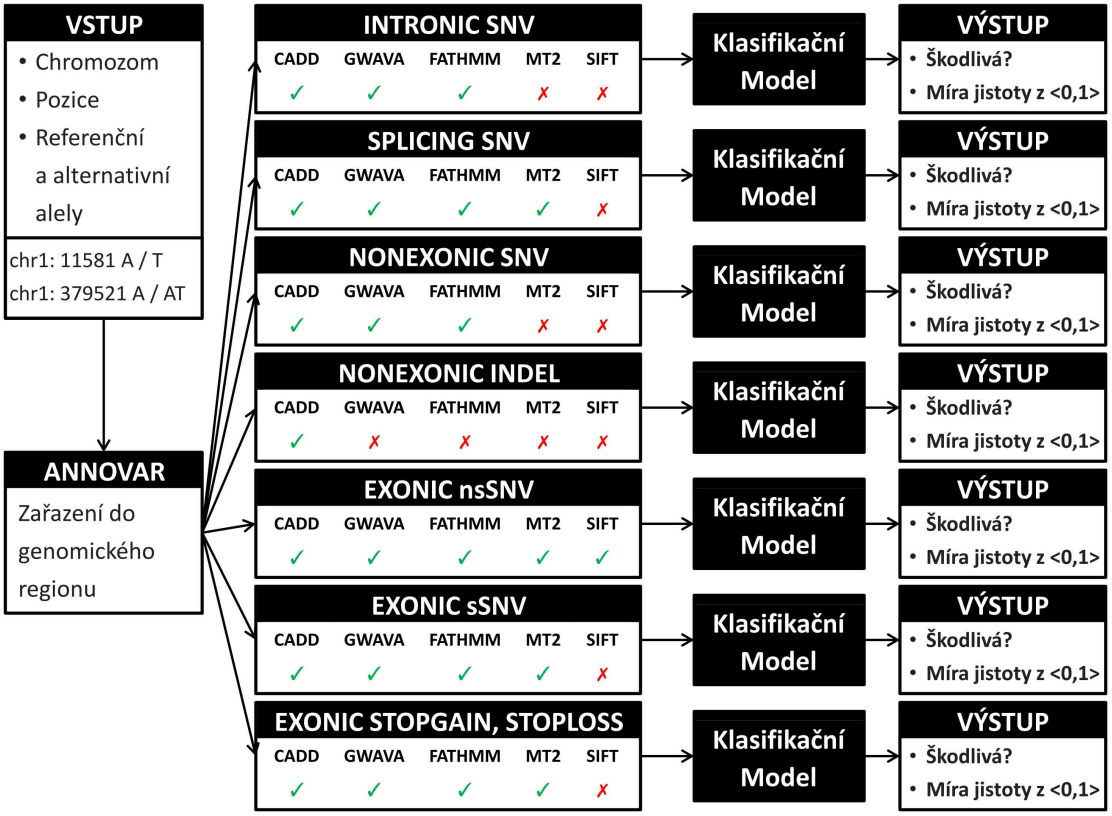
Výpočet metrik pro práh 0,1 je naznačen na modelovém obrázku 9.6. Je třeba vzít v úvahu, že WEKA na základě pořadí tříd v atributu effect v ARFF souboru přiřazuje skóre blíže nule pro první třídu (vezměme nyní škodlivé - **d**) a skóre blíže jedné pro třídu druhou (neutrální - **n**). Dále uvažujme, že se celý dataset skládá z 100 neutrálních a 100 škodlivých mutací.



Obrázek 9.6: Výpočet bodu ROC křivky.

9.5 Webové rozhraní

Součástí zadání práce byla také implementace webového rozhraní pro nový klasifikátor. Na obrázku 9.7 je celkové schéma klasifikátoru, které demonstruje posloupnost jeho činností. Pro vstupní množinu mutací se nejprve zjistí jejich genomický region. Mutace budou dle



Obrázek 9.7: Schéma implementovaného klasifikátoru.

regionu ohodnoceny jednotlivými nástroji a poté i příslušným metanástrojem. Výsledkem bude třída klasifikace a míra důvěry modelu ve správnost klasifikace.

K implementaci rozhraní jsem použil prostředí Eclipse a softwarový balíček Google Web Toolkit[5] (dále GWT) v kombinaci s jeho nadstavbou SmartGWT[6] .Tato práce je pokračováním projektu PredictSNP, což je webový nástroj pro analýzu vlivu aminokyselinových mutací. Z toho důvodu jsem z tohoto projektu převzal vizuální koncept a svůj klasifikátor nazval PredictSNP2.

Balíček GWT zajišťuje rozdělení projektu na klientskou část, která je přeložena do JavaScriptu a dostupná v prohlížeči, a část serverovou. Úkolem bylo zajistit komunikaci těchto stran na principu *request-response*, realizovat výpočetní serverovou část a zobrazit výsledky v klientském prohlížeči. Hlavní body správného běhu jsou následující:

- Uživatel otevře webovou stránku se vstupním formulářem.
- Uživatel zadá množinu mutací ve specifikovaném formátu a vydá požadavek k ohodnocení.
- Je odeslán požadavek na server, který registruje novou klasifikační úlohu, vytvoří pro ni unikátní identifikátor a dočasnou složku. Jako odpověď zasílá zpět vygenerovaný unikátní identifikátor.
- Prohlížeč přijme identifikátor úlohy a přesměruje uživatele na stránku s výsledky, kde čeká na dokončení výpočtu.
- Server provede klasifikaci mutací dle schématu 9.7. Rozdělí tedy mutace podle regionů, ohodnotí je nástroji a případně i natrénovaným modelem, pokud je pro danou gemonickou kategorii k dispozici.
- Výsledky pak zasílá zpět na klientskou část, kde jsou zobrazeny v tabulce (obrázek 9.9).
- Uživatel má možnost stažení výsledků klasifikace v CSV souboru.

9.5.1 Průběh klasifikace

Při realizaci klasifikace jsem zčásti využil skriptů, které byly použity pro ohodnocení trénovací sady mutací a trénování. Výsledkem klasifikace je hodnota míry důvěry modelu ve správnost predikce na škále <-1,1>, kde záporné hodnoty znamenají škodlivou predikci, kladné neutrální a nula byla vložena v případě, že nástroj nedovedl určit efekt mutace. Tato procentuální hodnota důvěry byla získána z trénovacích datasetů, podle toho, jak přesně pro dané skóre klasifikují (dále rozvedeno v kapitole 10.3). Z toho plyne, že pro mutace mimo regiony[7], nad kterými probíhalo trénování, není tato hodnota dostupná. Pokud se taková mutace vyskytla, byla míra důvěry stanovena na základě toho, jak moc je predikční skóre nástroje vzdáleno od nástrojem deklarovaného rozhodovacího prahu[8].

[5] http://www.gwtproject.org/

[6] http://www.smartclient.com/smartgwt/

[7] Dle obrázku 9.3 se jedná o exonické kategorie mutací, které způsobují posun čtecího rámce.

[8] Nástroj CADD rozhodovací práh neposkytuje, u trénovacích sad byl experimentálně určen, u ostatních mutací je v tabulce s výsledky zobrazeno pouze jeho skóre.

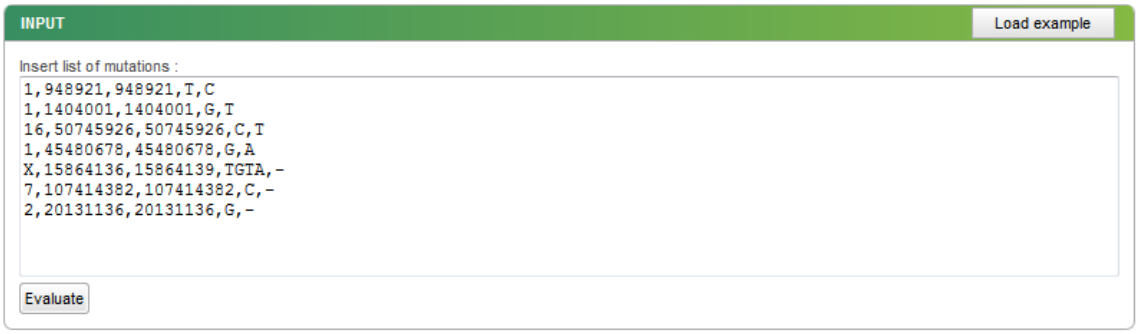
9.5.2 Popis rozhraní

Jak již bylo zmíněno, vizuální styly webového rozhraní byly převzaty z projektu PredictSNP. V příloze C.1 je k dispozici snímek domovské stránky. V horní části je menu, napravo kontaktní informace a použité zdroje. Samotná funkčnost rozhraní je jednoduchá a intuitivní. Na úvodní stránce se nachází vstupní formulář (obrázek 9.8), do kterého uživatel zadá množinu mutací oddělených novým řádkem ve tvaru:

```
chr,start,end,ref_allele,alternative_allele
```

Pro detailnější popis vstupního formátu může uživatel využít odkaz v menu, případně nahrát vzorový příklad. Po stisknutí tlačítka *Evaluate* a dokončení výpočtu se zobrazí tabulka s výsledky (obrázek 9.9). Pro identifikaci mutace je uveden její vstupní řetězec, dále následuje region a exonická funkce, kterou přiřadil ANNOVAR, a klasifikace jednotlivých nástrojů. Zeleně podbarvené políčko značí, že nástroj zařazuje mutaci mezi neutrální, červené políčko je pro mutace škodlivé. Hodnota v něm udává míru důvěry ve výsledek klasifikace. U nástroje CADD je přípustné i modré políčko, které místo míry důvěry obsahuje CADD skóre. Vyskytuje se u mutací, které spadají do regionu bez natrénovaného modelu a není tedy k dispozici experimentálně stanovený rozhodovací práh.

Vedle zobrazení výsledků klasifikace je možné rovněž stažení výstupní tabulky ve formátu CSV. Tento CSV soubor obsahuje hlavičku a poté na každém řádku identifikaci jedné mutace. Pro ni je k dispozici vždy anotace programem ANNOVAR a dále skóre a hodnoty míry důvěry všech integrovaných nástrojů včetně konsenzuálního metanástroje. V případě chybějících hodnot je vložen znak **-**.



Obrázek 9.8: Vstupní formulář webového rozhraní.

RESULTS

neutral deleterious score XX % confidence

Mutation	Region	Exonic Function	PredictSNP2	CADD	GWAVA	FATHMM	MutationTaster2	SIFT
1:84875173-84875173 C/T	intronic	.	95 %	72 %	82 %	78 %	-	-
1:5935162-5935162 A/T	splicing	.	100 %	35 %	41 %	12 %	94 %	-
1:948921-948921 T/C	UTR5	.	96 %	97 %	59 %	99 %	-	-
1:1404001-1404001 G/T	UTR3	.	96 %	99 %	63 %	100 %	-	-
X:15864136-15864139 TGTA/-	splicing	.	-	100 %	-	-	-	-
1:67705958-67705958 G/A	exonic	nonsynonymous SNV	55 %	54 %	55 %	42 %	44 %	65 %
16:50745926-50745926 C/T	exonic	nonsynonymous SNV	74 %	54 %	51 %	84 %	44 %	54 %
1:45480678-45480678 G/A	exonic	synonymous SNV	77 %	91 %	39 %	92 %	-	-
7:107414382-107414382 C/-	exonic	stopgain	68 %	90 %	-	-	-	-
2:20131136-20131136 G/-	exonic	frameshift deletion	-	36	-	-	-	-

DOWNLOAD

CSV file

Obrázek 9.9: Výstupní formulář s výsledky klasifikace.

Kapitola 10

Experimenty

Při práci s metodami strojového učení jsem vyzkoušel celkem 22 různých metod strojového učení. Podle jejich úspěšnosti na trénovací sadě jsem poté vybral 8 nejúspěšnějších, se kterými jsem experimentoval dále. Snažil jsem se přitom vybírat metody různých tříd dle tabulky 8.1. V tabulce 10.1 jsou vybrané metody včetně nastavení parametrů, které jsem u nich nakonec použil. Parametry jsou ve formátu pro spuštění přes Java API, jejich význam lze dohledat v nápovědě WEKA [21].

Metoda	Parametry
IBk	-K 1 -W 0 -A „Linear“
NaiveBayes	nejsou
LogisticRegression	-R 1.0E-8 -M -1
VotedPerceptron	-I 1 -E 1.0 -S 1 -M 10000
MultiLayerPerceptron	-L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
LibSVM_linear	-S 0 -K 0 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1
LibSVM_polynomial	-S 0 -K 1 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1
RandomForest	-I 100 -K 0 -S 1

Tabulka 10.1: Zúžený výběr metod strojového učení pro hledání konsenzu.

10.1 Výsledky trénování

Vzhledem k tomu, že na datasetu a modelu číslo 4 je zastoupen pouze nástroj CADD, postrádá smysl provádět nad jeho výsledky rozsáhlé experimenty. Metody strojového učení totiž v podstatě jen hledají ideální práh C-skóre mezi neutrálními a škodlivými mutacemi. Pro kategorii inzercí a delecí v intronické části (dataset 4) proto bude uvedena pouze analýza výsledků CADD ve formě příslušných grafů. Výstupem klasifikace bude, stejně jako u kategorií mimo trénovací datasety, pouze výsledek predikce dílčích nástrojů (v tomto případě jen CADD).

V příloze B.1 jsou k dispozici kompletní tabulky výsledků trénování pro všechny trénovací datasety. Uvedené hodnoty TP, TN, FP a FN jsou výsledkem křížové validace na trénovacích sadách. Podle tabulek lze provést komplexní srovnání výkonnosti na základě statistických metrik. V tabulce 10.1 jsou shrnuty nejdůležitější z nich, tedy Matthewsův korelační koeficient (MCC), Normovaná přesnost (ACC) a plocha pod ROC křivkou (AUC).

		Nástroje					Konsensus
		MT	SIFT	FATHMM	CADD	GWAVA	Logistická regrese
1	ACC	--	--	0,817	0,807	0,679	**0,834**
	MCC	--	--	0,662	0,620	0,434	**0,678**
	AUC	--	--	0,919	0,856	0,765	**0,911**
		MT	SIFT	FATHMM	CADD	GWAVA	Logistická regrese
2	ACC	0,637	--	0,646	0,679	0,567	**0,737**
	MCC	0,396	--	0,410	0,390	0,139	**0,513**
	AUC	0,330	--	0,796	0,739	0,595	**0,795**
		MT	SIFT	FATHMM	CADD	GWAVA	Logistická regrese
3	ACC	--	--	0,933	0,952	0,616	**0,954**
	MCC	--	--	0,870	0,904	0,246	**0,908**
	AUC	--	--	0,980	0,987	0,629	**0,988**
		MT	SIFT	FATHMM	CADD	GWAVA	Náhodný les
5	ACC	0,760	0,526	0,705	0,695	0,540	**0,788**
	MCC	0,563	0,143	0,463	0,402	0,090	**0,577**
	AUC	0,620	0,629	0,811	0,766	0,557	**0,855**
		MT	SIFT	FATHMM	CADD	GWAVA	Naivní Bayes
6	ACC	0,624	--	0,693	0,706	0,573	**0,766**
	MCC	0,346	--	0,388	0,415	0,193	**0,542**
	AUC	0,468	--	0,808	0,742	0,598	**0,828**
		MT	SIFT	FATHMM	CADD	GWAVA	Naivní Bayes
7	ACC	0,570	--	0,696	0,642	0,587	**0,727**
	MCC	0,231	--	0,401	0,301	0,212	**0,465**
	AUC	0,267	--	0,766	0,687	0,677	**0,791**

Obrázek 10.1: Porovnání úspěšnosti nástrojů a nejlepší z metod strojového učení.

Pro každý dataset uvádím výsledky nástrojů a pro srovnání výsledky konsenzuálního klasifikátoru, který je reprezentován metodou strojového učení vykazující nejlepší výsledky.

Z tabulky 10.1 lze vyčíst hned několik důležitých věcí. Jak je vidět, tak nástroje klasifikují mutace v různých regionech různě přesně. Zatímco na datasetu 1 (intronické substituce) je dle přesnosti nejúspěšnější nástroj FATHMM, na datasetu 2 (substituce v místě sestřihu) je dle normované přesnosti úspěšnější CADD. V dalších datasetech a metrikách je těchto situací mnoho. To jen dokládá prvotní domněnku, že žádný z nástrojů globálně nepřevyšuje všechny ostatní, proto má kombinace jejich výsledků význam.

Dále, v tabulce 10.1 ve sloupci zcela vpravo jsou uvedeny výsledky nejlepší z metod strojového učení, která reprezentuje konsenzuální klasifikátor. Aplikace konsenzuálního přístupu v každém z datasetů vylepšila alespoň dvě ze tří klíčových metrik. V některých případech se jednalo o zlepšení spíše zanedbatelné (dataset 3), jindy však poměrně významné (dataset 2).

Z tabulky 10.1 a odpovídajících tabulek v příloze B.1 je dále patrné, že experimentování s metodami strojového učení je vhodné věnovat dostatečnou pozornost. Jak je vidět, tak neexistuje univerzální způsob hledání konsenzu, neboť pro každou kategorii může být úspěšná jiná metoda. Jedny z nejlepších výsledků vykazoval obecně klasifikátor založený

na logistické regresi, nebyl však nejúspěšnější ve všech kategoriích. Konsenzuální nástroje byly pro každý dataset realizovány pomoci nejúspěšnější z metod na daném datasetu, přesně podle tabulky 10.1.

10.1.1 ROC charakteristiky

Na obrázku 10.2 a 10.3 jsou zobrazeny ROC charakteristiky nástrojů a metod strojového učení na jednotlivých trénovacích sadách. Jak již bylo zmíněno v předchozích kapitolách, slouží hlavně pro vizuální porovnání úspěšnosti, pro číselné porovnání se používají hodnoty obsahů ploch pod těmito křivkami. Tyto křivky a obsahy ploch byly získány v rámci křížové validace jednotlivých modelů a jsou hlavním výstupem provedených experimentů.

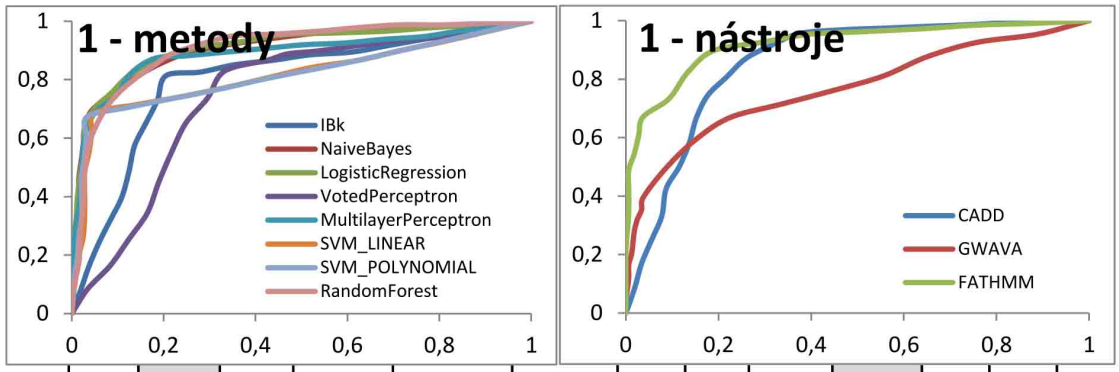
Na obrázku je v řádku vždy graf s ROC křivkami metod strojového učení (různých forem konsenzů) a vedle pro srovnání ROC charakteristiky dílčích nástrojů. Pod těmito grafy se nachází tabulka AUC hodnot. Nejvyšší hodnota, která označuje nejúspěšnější klasifikátor, je podbarvena šedou barvou.

Z ROC charakteristik lze vyvodit také několik závěrů. Až na intronické substituce v datasetu 1 je vždy konsenzuální nástroj dle metriky AUC minimálně srovnatelný s dílčími nástroji, často je však převyšuje. Tím lze opět potvrdit závěry prezentované výše, tedy že konsenzuální přístup obecně zvyšuje kvalitu predikce.

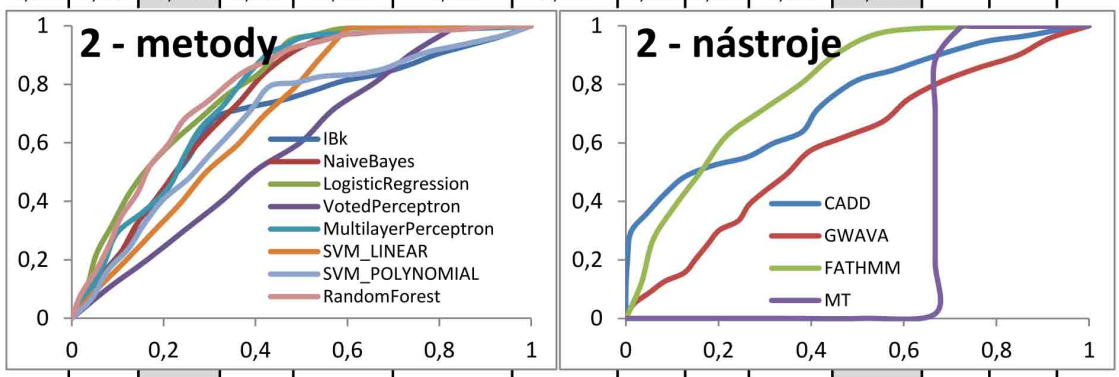
Výsledky pro obecné neexonické substituce (dataset 3) na obrázku 10.2 ukazují, že tato kategorie je nástroji zpracovávána velmi dobře. ROC charakteristiky se zde blíží ideálnímu tvaru a obsah plochy pod křivkami je blízko jedné. Naopak třeba mutace v místě sestřihu (dataset 2) jsou kategorií ke klasifikaci obtížnější, protože hodnoty AUC i u nejlepších metod se pohybují pod hranicí 0,8. Je třeba říci, že budování konsenzu nad dílčími výsledky do jisté míry sdílí všechny jejich nedostatky a nepřesnosti (popsáno v kapitole 8.2.7). Nelze proto předpokládat, že nad výsledky méně úspěšných nástrojů vznikne konsenzus značně přesnější. Tento fakt lze ověřit i na ROC křivkách, Kde je vidět, že metody strojového učení vždy kopírují či mírně vylepšují nejlepší z dílčích nástrojů.

Dále, nástroj MutationTaster2 má diskrétní výstup, kterým pouze říká, do jaké třídy mutace patří. Proto má jeho ROC charakteristika schodový tvar a pro porovnání je lepší využít třeba normovanou přesnost. MutationTaster2 sice poskytuje i hodnoty pravděpodobnosti své klasifikace, ovšem v naprosté většině případů jsou tyto pravděpodobnosti rovny 1.

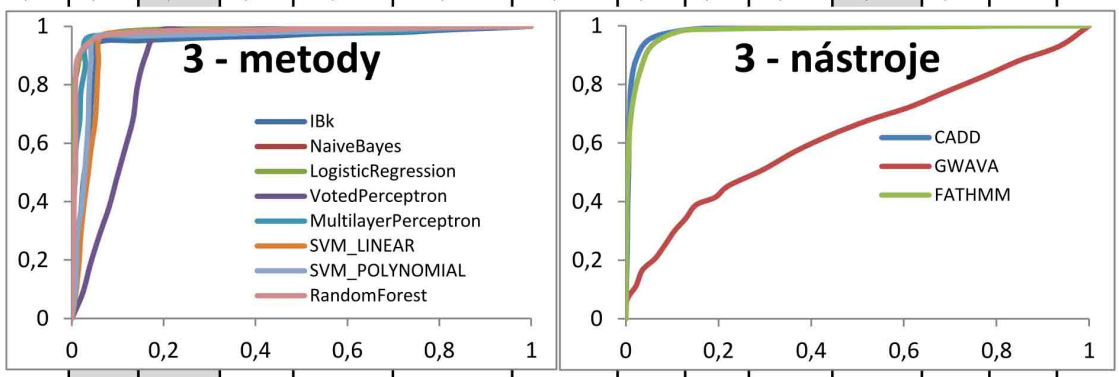
A konečně, přestože se některé nástroje jeví jako méně úspěšné, jejich vstup do konsenzu je cenný. Je zřejmé, že metody strojového učení se staly úspěšnějšími právě kvůli kombinaci vstupů, které by samy o sobě takovou úspěšnost neměly. Dílčí nástroj nemusí vykazovat vysokou úspěšnost, ale jeho výsledky se mohou stát vyvažujícím elementem, který nakonec klasifikaci ovlivní kladným směrem.



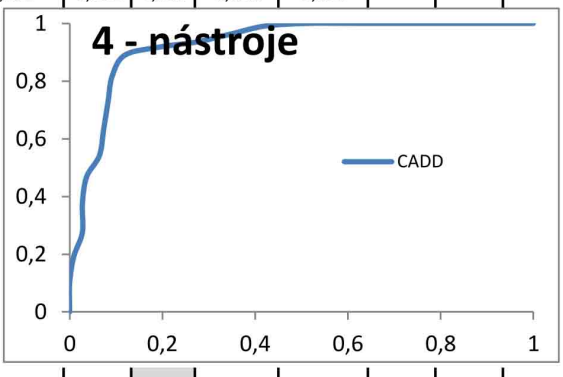
IBk	Bayes	Logistic	Voted	Multilayer	SVM_LINEAR	SVM_POLY	Forest	CADD	GWAVA	FATHMM	MT	SIFT
0,801	0,909	0,911	0,748	0,888	0,814	0,817	0,908	0,856	0,765	0,920	--	--



IBk	Bayes	Logistic	Voted	Multilayer	SVM_LINEAR	SVM_POLY	Forest	CADD	GWAVA	FATHMM	MT	SIFT
0,682	0,754	0,795	0,593	0,766	0,696	0,678	0,790	0,739	0,595	0,796	0,330	--

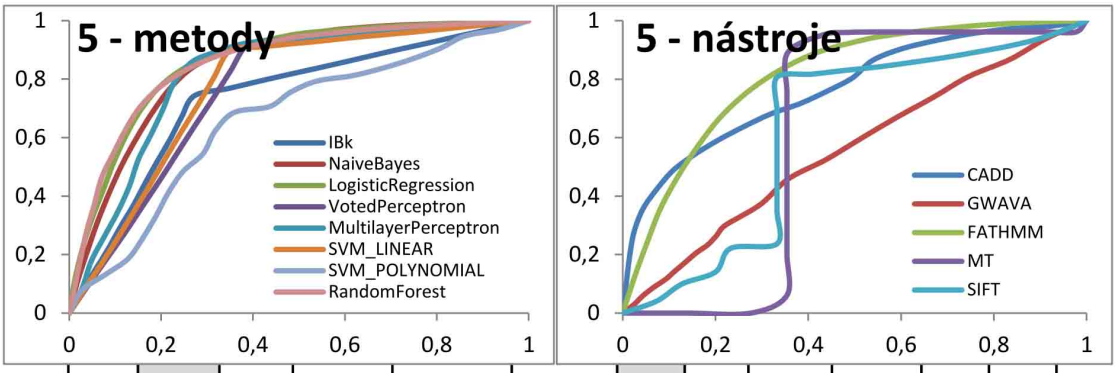


IBk	Bayes	Logistic	Voted	Multilayer	SVM_LINEAR	SVM_POLY	Forest	CADD	GWAVA	FATHMM	MT	SIFT
0,945	0,988	0,988	0,898	0,970	0,951	0,958	0,982	0,987	0,629	0,980	--	--

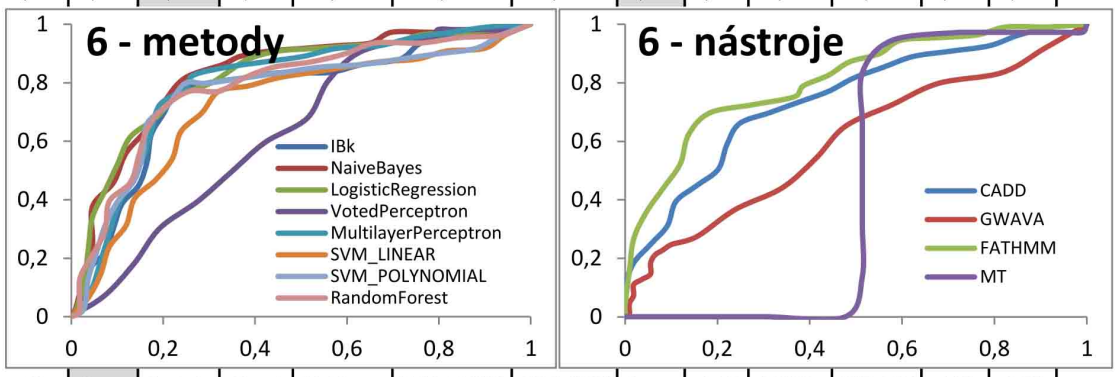


IBk	Bayes	Logistic	Voted	Multilayer	SVM_LINEAR	SVM_POLY	Forest	CADD	GWAVA	FATHMM	MT	SIFT
--	--	--	--	--	--	--	--	0,927	--	--	--	--

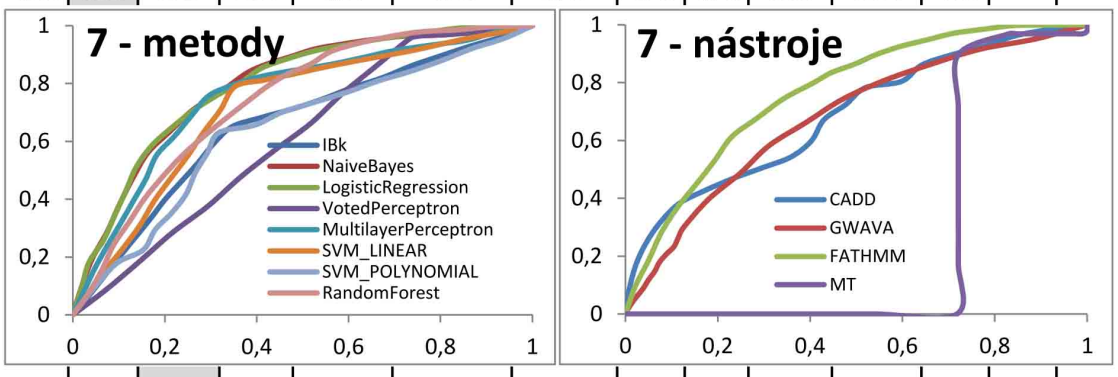
Obrázek 10.2: ROC charakteristiky klasifikátorů na datasetech 1-4.



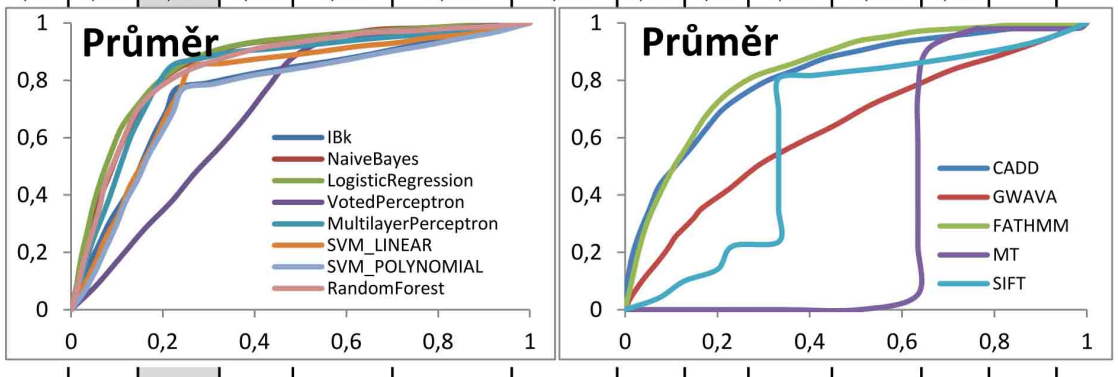
IBk	Bayes	Logistic	Voted	Multilayer	SVM_LINEAR	SVM_POLY	Forest	CADD	GWAVA	FATHMM	MT	SIFT
0,734	0,841	0,856	0,763	0,820	0,772	0,660	0,855	0,766	0,557	0,811	0,620	0,629



IBk	Bayes	Logistic	Voted	Multilayer	SVM_LINEAR	SVM_POLY	Forest	CADD	GWAVA	FATHMM	MT	SIFT
0,764	0,828	0,820	0,635	0,797	0,721	0,765	0,786	0,742	0,598	0,808	0,468	--



IBk	Bayes	Logistic	Voted	Multilayer	SVM_LINEAR	SVM_POLY	Forestt	CADD	GWAVA	FATHMM	MT	SIFT
0,654	0,791	0,792	0,609	0,753	0,717	0,639	0,740	0,687	0,677	0,766	0,267	--



IBk	Bayes	Logistic	Voted	Multilayer	SVM_LINEAR	SVM_POLY	Forest	CADD	GWAVA	FATHMM	MT	SIFT
0,771	0,857	0,867	0,705	0,836	0,789	0,756	0,847	0,815	0,635	0,833	0,355	0,629

Obrázek 10.3: ROC charakteristiky klasifikátorů na datasetech 5-7 a průměrné.

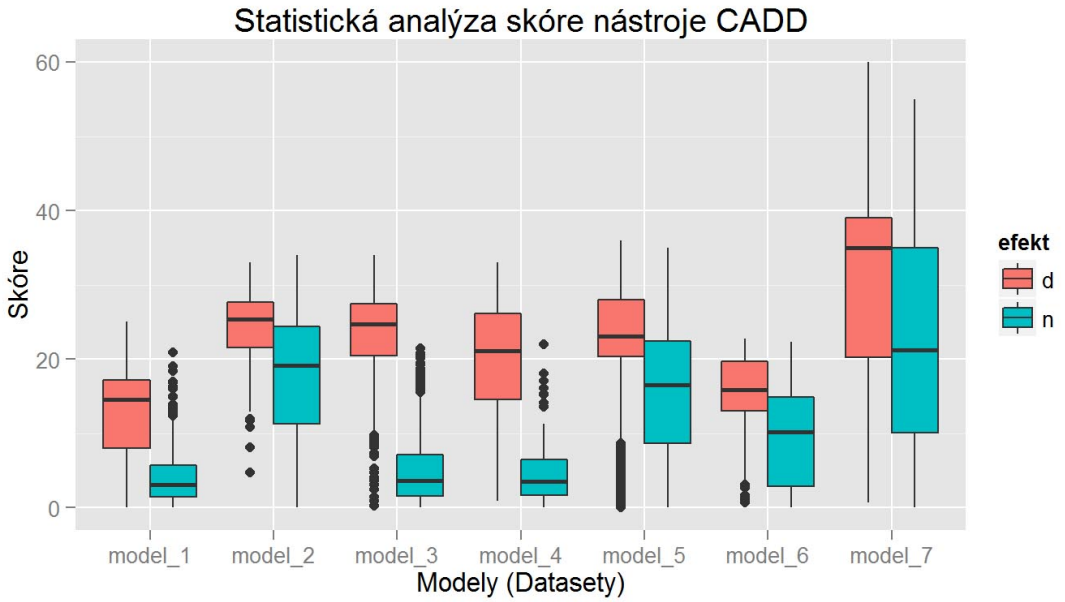
10.2 Analýzy distribuce a přesnosti

V kapitole 8.2.7 je řečeno, že hlavní výhodou použití konsenzu je zvýšení přesnosti a robustnosti klasifikace. Zlepšení přesnosti bylo ověřeno v předchozí části pomocí různých statistických metrik. Robustnost klasifikátoru je dána tím, jak širokou škálu vstupů je prediktor schopen zpracovat. Dle tabulky 6.2 je zřejmé, že nástroje obecně nezvládají predikci všech typů mutací na celém genomu. Nový metanástroj je však díky kombinaci jejich vstupů schopen klasifikovat jakoukoli mutaci, ke které existuje výstup alespoň jednoho dílčího nástroje, což značně zvyšuje jeho robustnost a univerzálnost jeho použití.

Kvůli této jisté závislosti konsenzuálního klasifikátoru na dílčích výsledcích jsem v rámci experimentů provedl také analýzu úspěšnosti dílčích nástrojů v porovnání s nad nimi postaveným metanástrojem. Cílem bylo zjistit hlubší vztah mezi hodnotami skóre na výstupu nástrojů a přesností klasifikace. Tuto analýzu jsem prováděl odděleně na všech trénovacích datasetech a využil jsem k tomu dvou typů grafů, a to krabicové grafy (*boxplots*) a houslové grafy (*violin plots*).

10.2.1 Statistická analýza skóre - krabicové grafy

V příloze B.2 jsou k dispozici krabicové grafy pro všechny nástroje včetně konsenzuálního metanástroje. Na vodorovné ose jsou vyznačeny jednotlivé trénovací datasety, respektive odděleně jejich škodlivé a neutrální složky. Svislá osa má rozsah podle predikčního skóre, které nástroje na sadách produkují. Krabicový graf zřetelně ukazuje, jak jsou hodnoty skóre rozloženy pro neutrální a škodlivé mutace. Na obrázku 10.4 jsou jako příklad krabicové grafy pro nástroj CADD. Červené grafy jsou pro škodlivé mutace, modré pro neutrální. Na první pohled je patrné, že distribuce CADD skóre je různá pro odlišné regiony. Proto také nástroj neuvádí jednotný rozhodovací práh mezi neutrální a škodlivou mutací a je u něj žádoucí zkoumat různé genomické regiony zvlášť.



Obrázek 10.4: Krabicové grafy nástroje CADD.

Samotný krabicový graf zobrazuje ale samozřejmě mnohem více informací. Obdélníkový útvar je ohraničen tzv. prvním a třetím kvartilem. Jedná se o takové hodnoty skóre,

pro které platí následující:

- První kvartil, který je značen Q1, je hodnota skóre taková, že 25 % ze všech hodnot v dané kategorii (v témže krabicovém grafu) je menší nebo rovno této hodnotě a naopak 75 % hodnot je větší než nebo rovno Q1. V krabicovém grafu je první kvartil na spodní hranici obdélníka.
- Třetí kvartil, který se značí Q3, je hodnota skóre taková, že 25 % ze všech hodnot v dané kategorii (v témže krabicovém grafu) je větších nebo rovno této hodnotě a naopak 75 % hodnot je menší než nebo rovno Q3. V krabicovém grafu je třetí kvartil na horní hranici obdélníka.
- Vzdálenost mezi kvartily Q1 a Q3, tedy výška obdélníku je tzv. mezikvartilová vzdálenost, pro kterou se užívá zkratka IQR. Vodorovná čára uvnitř obdélníku pak označuje medián skóre.
- Vodorovná čára uprostřed obdélníka udává medián hodnot skóre.

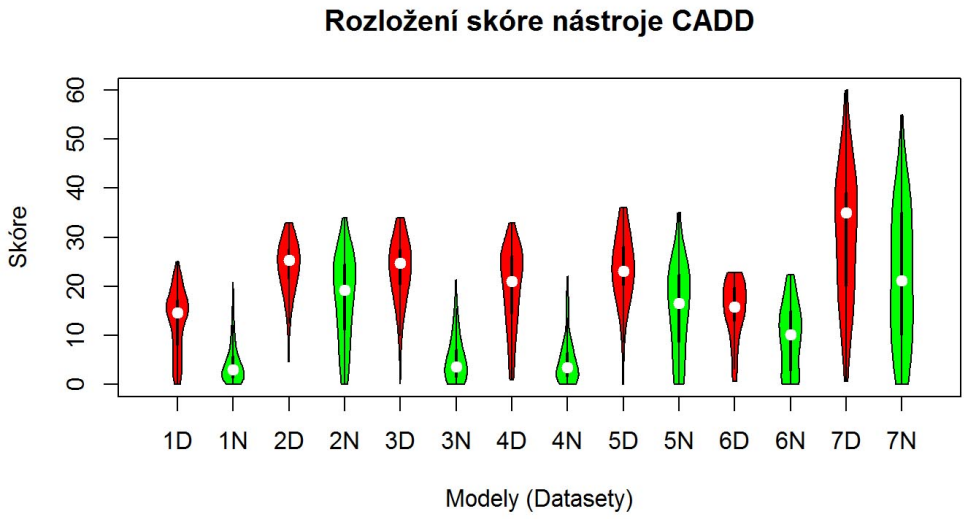
Z hlediska analýzy rozložení skóre klasifikátorů je ale cenná hlavně vizuální informace, kterou je překryv rozložení predikčního skóre neutrálních a škodlivých mutací v trénovací sadě. U dobrých klasifikátorů by nemělo docházet k vzájemnému překrytí krabicového grafu pro škodlivou a neutrální složku datasetu. Na obrázku 10.4 je vidět, že pro CADD je toto nejlépe splněno v modelech 3 a 4, což také koresponduje s vysokými hodnotami metriky AUC na těchto modelech. V grafu lze také pozorovat výskyt odlehlých hodnot, které jsou reprezentovány vyplněnými puntíky. Statisticky se jedná o hodnoty vzdálené více než 1,5 násobek IQR od Q1 nebo Q3 a obvykle jsou považovány za šum. V případě rozložení skóre klasifikace mají tyto hodnoty dvojí význam. Podle grafů se nejčastěji se jedná o špatně klasifikované prvky, neboť se vyskytují v místě, kde dominuje krabicový graf opačné třídy klasifikace. Druhou variantou jsou naopak zcela správně klasifikované mutace, neboť se nachází na správné straně od rozhodovacího prahu, jsou od něj však více vzdáleny. Lze je pozorovat na grafu pro složku škodlivých mutací modelu 3 u nástroje GWAVA na obrázku B.9. Jako odlehlé jsou označeny proto, že krabicový graf pro danou třídu nemá dostatečně velkou IQR.

Vedle krabicových grafů pro dílčí nástroje je v příloze B.2 přiložen rovněž obrázek s krabicovými grafy metanástroje. Je patrné, že k vzájemnému překryvu grafů pro škodlivé a neutrální mutace dochází v mnohem menší míře než u dílčích nástrojů. Metanástroj tedy dokáže lépe rozeznat třídu mutace. Na modelu číslo 3 lze pozorovat velké množství odlehlých hodnot. I k tomu může u krabicových grafů dojít, tento fakt však poté velmi zkresluje pohled na rozložení skóre. Proto jsem v následující kapitole analyzoval ještě další grafy, které nedostatky krabicových grafů (týkající se zobrazení rozložení skóre) kompenzují.

10.2.2 Rozložení skóre - houslové grafy

Krabicové grafy ukazují mnoho metrik, které se týkají statistického rozložení skóre na trénovacích sadách. Jedná se však obvykle o číselné údaje, pro lepší vizualizaci rozložení se používají houslové grafy. Tyto grafy jsou k dispozici v příloze B.3 a na obrázku 10.5 je jako příklad opět graf pro nástroj CADD. Význam os je stejný jako u krabicových grafů, místo obdélníku, kvartilů a odlehlých hodnot je však zobrazena křivka funkce hustoty pravděpodobnosti pro hodnoty skóre. Na rozdíl od krabicového grafu tak lze na první pohled zjistit nejen informace o hustotě rozložení, ale i získat lepší přehled o překryvu skóre neutrálních a škodlivých mutací. Opět by mělo platit, že houslový graf pro neutrální a škodlivou složku datasetu by měl být na svislé ose oddělitelný. Na houslovém grafu

nástroje CADD lze pozorovat, že pro datasety 3 a 4 je překryv nejmenší a největší naopak pro datasety 6 a 7. Výsledek tedy koresponduje s AUC metrikami pro tento nástroj.



Obrázek 10.5: Houslové grafy nástroje CADD.

Pro srovnání jsou přiloženy i grafy pro metanástroj na obrázku B.13. Výsledky houslových grafů potvrzují, že pro datasety 2 a 7 je obecně obtížné najít rozdělovací práh mezi třídami škodlivých a neutrálních mutací, a to jak u dílčích, tak u konsenzuálního nástroje. Celkově nižší hodnoty metriky AUC pro tyto datasety na obrázku 10.1 to dokládají.

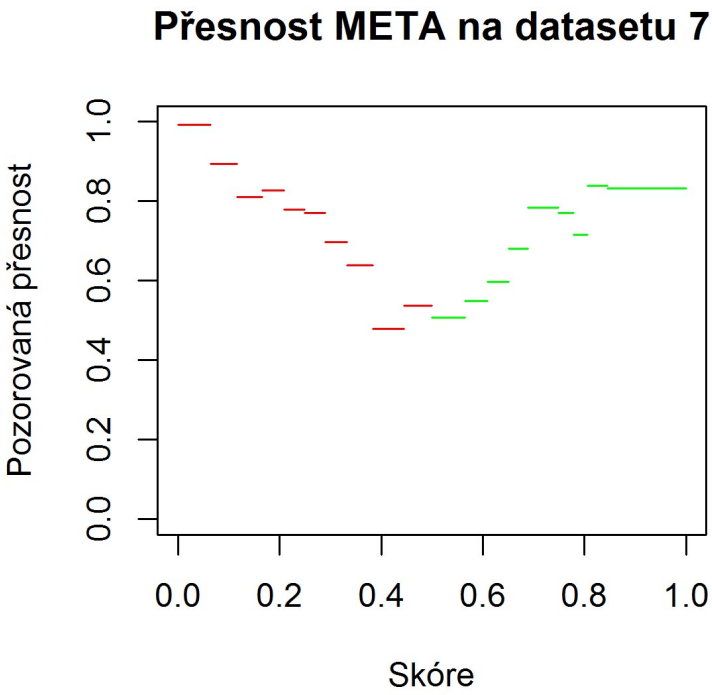
Z dílčích nástrojů vykazují průměrně nejmenší překryvy FATHMM (B.16) a CADD (B.14), naopak nástroj GWAVA (B.15) neodděluje složky škodlivých a neutrálních dostatečně dobře ani na jednom datasetu. Konsensuální nástroj (B.13) podle očekávání zvýraznil dobré výsledky ostatních nástrojů.

10.3 Přesnost klasifikátorů dle skóre

V příloze B.4 jsou k dispozici grafy, které ukazují, s jakou přesností jednotlivé nástroje pracují pro různé intervaly skóre. Na vodorovné ose je škála skóre, kterou pro daný trénovací dataset nástroj poskytuje, na svislé ose je vynesena přesnost. Červené úseky označují hodnoty skóre škodlivé mutace, zelené pak neutrální. Předpokládaný tvar po částech spojitého grafu by měl být podle obrázku 10.6, tedy že blízko krajních bodů vodorovné osy je přesnost nejvyšší a blízko rozhodovacího prahu naopak nižší, neboť zde může samozřejmě častěji docházet k chybám.

Jak je ale vidět z dalších grafů v příloze B.4, tak obvyklé průběhy lze pozorovat pouze u několika nástrojů (např. CADD na datasetu 3, konsenzuální metanástroj na datasetu 7) a ne na všech datasetech. Pravděpodobně je to způsobeno nižší úspěšností nástrojů na mutacích v různých genomických regionech. Příčinou může být také nedostatečná velikost trénovací sady a tím způsobená ztráta obecnosti.

Nástroj CADD neudává hodnotu oddělovacího prahu pro své výstupy, pro následující analýzu však byly nutné. Jejich přibližné hodnoty jsem nalezl experimentálně a odděleně na všech datasetech tak, aby byla úspěšnost CADD maximální. Pro ilustraci uvádím nalezené prahy v tabulce 10.2.



Obrázek 10.6: Závislost přesnosti metanástroje na intervalu výstupního skóre na datasetu 7.

Dataset / Model	Rozhodovací práh CADD
1	8,6
2	18,7
3	14,0
4	9,6
5	18,3
6	13,0
7	35,1

Tabulka 10.2: Experimentálně stanovené prahy CADD skóre pro jednotlivé datasety.

Data z grafů B.4 mohou být dále využita na výstupu klasifikátoru v podobě míry důvěry ve správnost výsledku, jak je naznačeno na schématu 9.7. Podle výsledného skóre, které konsenzuální klasifikátor přiřadí dané mutaci bude v grafu dohledána přesnost, s jakou pro tyto hodnoty skóre pracuje. Uživatel tak bude mít k dispozici indikátor spolehlivosti predikce, který mu může pomoci při prioritizaci mutací k dalšímu experimentování.

10.4 Srovnání s aminokyselinovými prediktory

V rámci experimentů byla ještě provedena analýza přesnosti nukleotidových nástrojů v porovnání s klasifikátory aminokyselinových substitucí. Pro tyto účely byly z trénovací množiny vyextrahovány pouze nesynonymní mutace, ke kterým byla dohledána způsobená aminokyselinová substituce. Po odstranění mutací, které by nebylo možné aminokyselinovými klasifikátory zpracovat (příliš krátká nebo dlouhá referenční proteinová sekvence) a po odstranění překryvů s trénovacími sadami nástrojů vznikl dataset 88 955 mutací (73 506 neutrálních a 15 449 škodlivých).

Aminokyselinové klasifikátory		**Nukleotidové klasifikátory**	
Nástroj	**Normovaná přesnost**	**Nástroj**	**Normovaná přesnost**
MAPP	0,669	MT2	0,769
PhD-SNP	0,787	SIFT	0,740
PPH-1	0,715	FATHMM	0,712
PPH-2	0,721	CADD	0,702
SNAP	0,704	GWAVA	0,547

Tabulka 10.3: Porovnání přesnosti aminokyselinových a nukleotidových klasifikátorů.

Úspěšnost nástrojů je shrnuta v tabulce 10.3. Na základě těchto údajů lze říci, že klasifikátory obou tříd mají srovnatelnou přesnost. A to i přesto, že klasifikátory použité v rámci této práce nejsou na nesynonymní substituce obvykle nijak specializované. Pro komplexnější výzkum je tedy vhodnější použít nukleotidový klasifikátor, neboť je schopen stanovit efekt širší škály mutací.

Do budoucna by bylo možné vybudovat konsenzus i napříč těmito dvěma třídami klasifikátorů a porovnat přesnosti. Díky rozdílnosti přístupů ke stanovení funkčního efektu mutací je pravděpodobné, že by se přesnost predikce dále zvyšovala. Na druhou stranu by pak ale bylo nutné převádět aminokyselinový a nukleotidový zápis mutací dle referenčního genomu, což zvýší paměťovou a časovou náročnost výpočtu.

Kapitola 11

Závěr

Cílem diplomové práce byl návrh a implementace nového metaklasifikátoru efektu nukleotidového polymorfismu v lidské DNA a jeho zpřístupnění přes webové rozhraní. Tento prediktor kombinuje výsledky pěti již existujících nástrojů pomocí metod strojového učení.

Hledání vhodné formy této kombinace zahrnovalo jednak vybudování trénovacího datasetu a jeho rozdělení na kategorie příbuzných mutací, dále pak experimentování s metodami strojového učení nad těmito podmnožinami. Ukázalo se, že zejména výběru metod je nutné věnovat zvýšenou pozornost, neboť tyto metody mají různé vlastnosti a úspěšnosti na klasifikačních úlohách. Parametry pro jejich spuštění byly voleny rovněž experimentálně. Na základě úspěšnosti metod strojového učení na trénovacích datasetech byla pro každý z těchto datasetů vybrána ta nejúspěšnější, což je shrnuto na obrázku 10.1. Obecně byly v každé kategorii úspěšnější jiné metody, což jen podtrhuje význam opakování pokusů a hledání parametrů těchto metod.

Jedním z hlavních cílů experimentální části bylo ověření, že konsenzuální přístup bude mít kladný vliv na výslednou přesnost a robustnost klasifikace. Z grafů a údajů na obrázcích 10.2 a 10.3 je patrné, že formy konsenzu reprezentované metaklasifikátory, které byly natrénovány metodami strojového učení a validované 10-fold křížovou validací, jsou obecně úspěšnější než kterýkoli z dílčích nástrojů. Podle metrik AUC se celkovou přesnost podařilo vylepšit ve všech kategoriích kromě intronických nukleotidových substitucí (dataset číslo 1). V průměru vykazuje nejlepší z metod strojového učení o **3,4** vyšší hodnotu AUC než nejúspěšnější dílčí nástroj. Vezmeme-li pak v úvahu tři nejdůležitější metriky pro porovnání, které jsou shrnuty v tabulce 10.1, pak užití konsenzu vedlo vždy k vylepšení alespoň dvou z nich. Normovanou přesnost se přitom podařilo zvýšit až o **7 %** (dataset 6 - synonymní exonické mutace). V příloze B.2 a B.3 jsou k dispozici také krabicové a houslové grafy pro všechny integrované nástroje na všech datasetech. Tyto grafy korespondují s výsledky statistických metrik, lze na nich také pozorovat rozdíl v rozložení skóre a možnostech oddělení škodlivých a neutrálních mutací. Metaklasifikátor pracuje podle očekávání tak, že vychází převážně z úspěšnějších nástrojů na příslušném datasetu. Jedná se tedy opět o pozitivní dopad konsenzu, který zvýrazňuje výhody a dobré výsledky nástrojů, nad nimiž je vybudován, a zároveň potlačuje jejich nedostatky.

Srovnání různých typů klasifikátorů mutací v tabulce 10.3 ukazuje, že nukleotidové prediktory jsou na množině nesynonymních substitucí podobně úspěšné jako prediktory aminokyselinových substitucí. A to i přesto, že na tento typ mutací nejsou nijak specializované. Tento fakt skýtá mnohé rozšiřující možnosti do budoucna, kdy by bylo možné budovat konsenzus napříč oběma třídami a ještě tak zvýšit celkovou úspěšnost klasifikace.

Robustnost a univerzálnost použití klasifikátoru je jedním ze základních znaků konsenzuálního přístupu. Nový klasifikátor je schopen ohodnotit jakoukoli mutaci, kterou

klasifikuje alespoň jeden dílčí nástroj a která spadá do regionu, nad nímž je natrénován některý z modelů. Další výhodou metaklasifikátoru je možnost analýzy jakéhokoliv typu mutace na celém lidském genomu. Díky implementaci webového rozhraní je také možné všechny výsledky přehledně zobrazit.

Literatura

[1] Abecasis, G. R.; Auton, A.; Brooks, L. D.; aj.: An integrated map of genetic variation from 1,092 human genomes. *Nature*, ročník 491, č. 7422, 2012: s. 56–65.

[2] Adzhubei, I. A.; Schmidt, S.; Peshkin, L.; aj.: A method and server for predicting damaging missense mutations. *Nature Methods*, ročník 7, 2010: s. 248–249.

[3] Alberts, B.; Bray, D.; Johnson, A.; aj.: *Základy buněčné biologie*. Garland Publishing, 1998, iSBN 80-902906-2-0.

[4] Apweiler, R.; Martin, M. J.; O'onovan, C.; aj.: Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, ročník 40, 2012: s. 71–75.

[5] Asthana, S.; Roytberg, M.; Stamatoyannopoulos, J.; aj.: Analysis of sequence conservation at nucleotide resolution. *PLoS Comput. Biol.*, ročník 3, č. 12, 2007: str. 254.

[6] Baldi, P.: *Bioinformatics : the machine learning approach*. Cambridge, Mass: MIT Press, 2001, ISBN 0-262-02506-X.

[7] Bromberg, Y.; Rost, B.: SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acid Research*, ročník 35, 2007: s. 3823–3835.

[8] Cappriotti, E.; Fariselli, P.; Calabrese, R.; aj.: Predicting protein stability changes from sequences using support vector machines. *Nucleic Acid Research*, ročník 21, 2005: s. 54–58.

[9] Capriotti, E.; Nehrt, N. L.; Kann, M. G.; aj.: Bioinformatics for personal genome interpretation. *Brief. Bioinformatics*, ročník 13, č. 4, 2012: s. 495–512.

[10] Castaldi, P. J.; Dahabreh, I. J.; Ioannidis, J. P. A.: An empirical assessment of validation practices for molecular classifiers. *Briefings in Bioinformatics*, ročník 12, č. 3, feb 2011: s. 189–202.

[11] Church, D. M.; Schneider, V. A.; Graves, T.; aj.: Modernizing reference genome assemblies. *PLoS Biol.*, ročník 9, č. 7, 2011.

[12] Clancy, S.: DNA damage and repair: mechanisms for maintaining DNA integrity. *Nature Education*, ročník 1, č. 1, 2008: str. 103.

[13] Clancy, S.: DNA transcription. *Nature Education*, ročník 1, č. 1, 2008: str. 41.

[14] Clancy, S.; Brown, W.: Translation: DNA to mRNA to protein. *Nature Education*, ročník 1, č. 1, 2008: str. 101.

[15] Cooper, G. M.; Shendure, J.: Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.*, ročník 12, č. 9, 2011: s. 628–640.

[16] Danecek, P.; Auton, A.; Abecasis, G.; aj.: The variant call format and VCFtools. *Bioinformatics*, ročník 27, č. 15, 2011: s. 2156–2158.

[17] Davydov, E. V.; Goode, D. L.; Sirota, M.; aj.: Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.*, ročník 6, č. 12, 2010: s. 100–125.

[18] Dunham, I.; Kundaje, A.; aj.: An integrated encyclopedia of DNA elements in the human genome. *Nature*, ročník 489, č. 7414, 2012: s. 57–74.

[19] Flicek, P.; aj.: Ensembl 2013. *Nucleic Acids Res.*, ročník 41, č. Database issue, 2013: s. 48–55.

[20] Giardine, B.; Riemer, C.; Hefferon, T.; aj.: PhenCode: connecting ENCODE data with mutations and phenotype. *Hum. Mutat.*, ročník 28, č. 6, 2007: s. 554–562.

[21] Hall, M.; Frank, E.; Holmes, G.; aj.: The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, ročník 11, č. 1, nov 2009: str. 10.

[22] Hamosh, A.; Scott, A. F.; Amberger, J. S.; aj.: Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, ročník 33, 2005: s. 514–517.

[23] Hindorff, L. A.; Sethupathy, P.; Junkins, H. A.; aj.: Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.*, ročník 106, č. 23, 2009: s. 9362–9367.

[24] Kharchenko, P. V.; Tolstorukov, M. Y.; Park, P. J.: Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.*, ročník 26, č. 12, nov 2008: s. 1351–1359.

[25] Kircher, M.; Witten, D. M.; Jain, P.; aj.: A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, ročník 46, č. 3, 2014: s. 310–315.

[26] Kumar, P.; Henikoff, S.; Ng, P. C.: Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, ročník 4, č. 7, 2009: s. 1073–1081.

[27] Landrum, M. J.; Lee, J. M.; Riley, G. R.; aj.: ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, ročník 42, č. Database issue, 2014: s. D980–985.

[28] Larranaga, P.: Machine learning in bioinformatics. *Briefings in Bioinformatics*, ročník 7, č. 1, feb 2006: s. 86–112.

[29] Li, H.: Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*, ročník 27, č. 5, 2011: s. 718–719.

[30] Loewe, L.: Genetic mutation. *Nature Education*, ročník 1, č. 1, 2008: str. 113.

[31] MacArthur, D. G.; Tyler-Smith, C.: Loss-of-function variants in the genomes of healthy humans. *Hum. Mol. Genet.*, ročník 19, č. R2, 2010: s. 125–130.

[32] Margulies, E. H.; Blanchette, M.; Haussler, D.; aj.: Identification and characterization of multi-species conserved sequences. *Genome Res.*, ročník 13, č. 12, 2003: s. 2507–2518.

[33] McLaren, W.; Pritchard, B.; Rios, D.; aj.: Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, ročník 26, č. 16, 2010: s. 2069–2070.

[34] Murphy, K.: *Machine learning a probabilistic perspective*. Cambridge, Mass: MIT Press, 2012, ISBN 978-0-262-01802-9.

[35] Nečas, O.; aj.: *Obecná biologie pro lékařské fakulty*. H&H, 2000, ISBN 80-86022-46-3.

[36] Pollard, K. S.; Hubisz, M. J.; Rosenbloom, K. R.; aj.: Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, ročník 20, č. 1, 2010: s. 110–121.

[37] Powers, D. M. W.: Evaluation: from precision, recall and f-measure to ROC, informedness, markedness and correlation. *Machine Learning Technologies*, ročník 2, 2011: s. 37–63.

[38] Pray, L.: Discovery of DNA structure and function: Watson and Crick. *Nature Education*, ročník 1, č. 1, 2008: str. 100.

[39] Pray, L.: DNA replication and causes of mutation. *Nature Education*, ročník 1, č. 1, 2008: str. 214.

[40] Pray, L.: What is a gene? Colinearity and transcription units. *Nature Education*, ročník 1, č. 1, 2008: str. 97.

[41] Ritchie, G. R.; Dunham, I.; Zeggini, E.; aj.: Functional annotation of noncoding sequence variants. *Nat. Methods*, ročník 11, č. 3, 2014: s. 294–296.

[42] Ritchie, G. R.; Flicek, P.: Computational approaches to interpreting genomic sequence variation. *Genome Med.*, ročník 6, č. 10, 2014: str. 87.

[43] Russell, S.: *Artificial intelligence : a modern approach*. Upper Saddle River, N.J: Prentice Hall/Pearson Education, 2003, ISBN 0-13-790395-2.

[44] Schaafsma, G. C.; Vihinen, M.: Varisnp, a benchmark database for variations from dbSNP. *Hum. Mutat.*, 2014.

[45] Schwarz, J. M.; Cooper, D. N.; Schuelke, M.; aj.: MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods*, ročník 11, č. 4, 2014: s. 361–362.

[46] Sherry, S. T.; Ward, M. H.; Kholodov, M.; aj.: dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, ročník 29, č. 1, 2001: s. 308–311.

[47] Shihab, H. A.; Rogers, M. F.; Gough, J.; aj.: An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*, 2015.

[48] Siepel, A.; Bejerano, G.; Pedersen, J. S.; aj.: Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, ročník 15, č. 8, 2005: s. 1034–1050.

[49] Stenson, P. D.; Mort, M.; Ball, E. V.; aj.: The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.*, ročník 133, č. 1, 2014: s. 1–9.

[50] Stone, E. A.; Sidow, A.: Physicochemical constraint violation b missense substitutions mediates impairment of protein function and disease severity. *Genome Research*, ročník 15, 2005: s. 978–986.

[51] Studer, R. A.; Dessailly, B. H.; Orengo, C. A.: Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. *Biochem. J.*, ročník 449, č. 3, 2013: s. 581–594.

[52] Tennessen, J. A.; Bigham, A. W.; O'Connor, T. D.; aj.: Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, ročník 337, č. 6090, 2012: s. 64–69.

[53] Thomas, P. D.; Campbell, M. J.; Kejariwal, A.; aj.: PANTHER: A library of protein families and subfamilies indexed by function. *Genome Research*, ročník 13, 2003: s. 2129–2141.

[54] Wang, K.; Li, M.; Hakonarson, H.: ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, ročník 38, č. 16, 2010: str. 164.

[55] Watson, J. D.; Crick, F. H.: Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, ročník 171, č. 4356, 1953: s. 737–738.

Dodatek A

Schéma databází

ClinVar		
Sloupec	**Datový typ**	**Popis**
AlleleID	integer	Identifikátor ClinVar
Type	string	Typ mutace
Name	string	Název, NM identifikátor
GeneID	integer	ID genu v NCBI databázi
GeneSymbol	string	Seznam GeneID genů, do kterých mutace zasahuje
ClinicalSignificance	string	Seznam prokázaných efektů mutace
RS#	integer	RS# identifikátor v dbSNP
nsv	string	NSV identifikátor v dbVar
RCVaccession	string	Seznam RCV, které reportují mutaci
TestedInGTR	string	Y/N, pokud existuje specifický záznam v GTR
PhenotypeIDs	string	Seznam DB jmen a identifikátorů fenotypů
Origin	string	Seznam všech alel na dané pozici
Assembly	string	Název referenčního genomu
Chromosome	string	ID chromozomu
Start	integer	Počáteční pozice mutace
Stop	integer	Koncová pozice mutace
Cytogenetic	string	ISCN band
ReviewStatus	string	Nejvyšší dosažený status ověření
HGVS(c.)	string	HGVS identifikátor (na nukleotidové úrovni)
HGVS(p.)	string	HGVS identifikátor (na proteinové úrovni)
NumberSubmitters	integer	Počet příspěvků, ve kterých se mutace vyskytla
LastEvaluated	datetime	Datum posledního vložení
Guidelines	string	Vazba na ACMG
OtherIDs	string	Seznam dalších identifikátorů
VariantID	integer	Hodnota pro dotazování na mutaci v NCBI

Tabulka A.1: Schéma databáze ClinVar.

VariSNP		
Sloupec	**Datový typ**	**Popis**
dbSNP_id	integer	RS# identifikátor z dbSNP
heterozygosity	float	Odhad heterozygozity podle frekvencí alel
heterozygosity_standard_error	float	Střední odchylka odhadu
creation_date	datetime	Datum vytvoření
creation_build	integer	Přiřazené číslo při vytvoření
update_date	datetime	Datum poslední úpravy
update_build	integer	Poslední přiřazené číslo
observed_alleles	string	Seznam pozorovaných alel
asn_from	integer	Počáteční pozice mutace na contigu
asn_to	integer	Koncová pozice
reference_allele	string	Referenční alela, ev. '-' v případě inzerce
orientation	string	Orientace vlákna k contigu
minor_allele_frequency	float	Globální MAF
minor_allele	string	Alela, k níž se vztahuje MAF
sample_size	integer	Počet chromozomů ve vzorové populaci
validation	string	Metoda validace
hgvs_names	string	HGVS identifikátory
allele_origin	string	Genetický původ
clinical_significance	string	Efekt mutace
functional_class	string	Genomický region výskytu
ncbi_gi	string	NCBI GI identifikátor
ncbi_accession	string	NCBI identifikátor sekvence
gene_symbol	string	Genový identifikátor dle HGNC
refseq_start_description	string	Popis počátku transkripce ref. sekvence
coding_dna_description	string	Popis kódující varianty dle HGVS - nukleotidový
protein_description	string	Popis kódující varianty dle HGVS - proteinový
coding_reference	string	NCBI číslo vydání a verze - mRNA
protein_reference	string	NCBI číslo vydání a verze - protein
predicted_RNA_variation	string	HGVS popis predikované RNA mutace
DNA_annotation	string	VariO anotace na úrovni DNA
RNA_annotation	string	VariO anotace na úrovni RNA
protein_annotation	string	VariO anotace na úrovni proteinu

Tabulka A.2: Schéma databáze VariSNP

Dodatek B

Tabulky a grafy

B.1 Statistické výsledky trénování

Dataset 1: nástroje FATHMM,GWAVA,CADD														
	NÁSTROJE					KONSENZUS - METODY STROJOVÉHO UČENÍ								
	MT	SIFT	FATHMM	CADD	GWAVA	IBk	NaiveBayes	LogisticRegression	Voted	Multilayer	SVM_LINEAR	SVM_POLYNOMIAL	RandomForest	
TP	--	--	216	239	127	265	233	241	276	287	222	215	277	
FN	--	--	106	83	195	57	89	81	46	35	100	107	45	
TN	--	--	310	281	310	255	299	296	190	266	307	310	265	
FP	--	--	12	41	12	67	23	26	132	56	15	12	57	
TP norm	--	--	0,671	0,742	0,394	0,823	0,724	0,748	0,857	0,891	0,689	0,668	0,860	
FN norm	--	--	0,329	0,258	0,606	0,177	0,276	0,252	0,143	0,109	0,311	0,332	0,140	
TN norm	--	--	0,963	0,873	0,963	0,792	0,929	0,919	0,590	0,826	0,953	0,963	0,823	
FP norm	--	--	0,037	0,127	0,037	0,208	0,071	0,081	0,410	0,174	0,047	0,037	0,177	
TPR / sensitivity	--	--	0,671	0,742	0,394	0,823	0,724	0,748	0,857	0,891	0,689	0,668	0,860	
FNR	--	--	0,329	0,258	0,606	0,177	0,276	0,252	0,143	0,109	0,311	0,332	0,140	
TNR / specificity	--	--	0,963	0,873	0,963	0,792	0,929	0,919	0,590	0,826	0,953	0,963	0,823	
FPR	--	--	0,037	0,127	0,037	0,208	0,071	0,081	0,410	0,174	0,047	0,037	0,177	
Cases +	--	--	322	322	322	322	322	322	322	322	322	322	322	
Cases -	--	--	322	322	322	322	322	322	322	322	322	322	322	
Total	--	--	644	644	644	644	644	644	644	644	644	644	644	
Accuracy	--	--	0,817	0,807	0,679	0,807	0,826	0,834	0,724	0,859	0,821	0,815	0,842	
Accuracy norm	--	--	**0,817**	**0,807**	**0,679**	**0,807**	**0,826**	**0,834**	**0,724**	**0,859**	**0,821**	**0,815**	**0,842**	
Precision	--	--	0,947	0,854	0,914	0,798	0,910	0,903	0,676	0,837	0,937	0,947	0,829	
Precision norm	--	--	0,947	0,854	0,914	0,798	0,910	0,903	0,676	0,837	0,937	0,947	0,829	
NPV	--	--	0,745	0,772	0,614	0,817	0,771	0,785	0,805	0,884	0,754	0,743	0,855	
NPV norm	--	--	0,745	0,772	0,614	0,817	0,771	0,785	0,805	0,884	0,754	0,743	0,855	
F-measure	--	--	0,785	0,794	0,551	0,810	0,806	0,818	0,756	0,863	0,794	0,783	0,845	
AUC	--	--	**0,919**	**0,856**	**0,765**	**0,801**	**0,909**	**0,911**	**0,748**	**0,888**	**0,814**	**0,817**	**0,908**	
MCC	--	--	0,662	0,620	0,434	0,615	0,666	0,678	0,464	0,719	0,666	0,660	0,684	
MCC norm	--	--	**0,662**	**0,620**	**0,434**	**0,615**	**0,666**	**0,678**	**0,464**	**0,719**	**0,666**	**0,660**	**0,684**	

Obrázek B.1: Statistické metriky pro model (dataset) 1.

Dataset 2: nástroje FATHMM,GWAVA,CADD,MT													
	NÁSTROJE					KONSENZUS - METODY STROJOVÉHO UČENÍ							
	MT	SIFT	FATHMM	CADD	GWAVA	IBk	NaiveBayes	LogisticRegression	Voted	Multilayer	SVM_LINEAR	SVM_POLYNOMIAL	RandomForest
TP	559	--	558	490	240	389	553	521	558	512	556	440	455
FN	1	--	2	70	320	171	7	39	2	48	4	120	105
TN	154	--	166	271	395	374	228	304	98	306	222	310	371
FP	406	--	394	289	165	186	332	256	462	254	338	250	189
TP norm	0,998	--	0,996	0,875	0,429	0,695	0,988	0,930	0,996	0,914	0,993	0,786	0,813
FN norm	0,002	--	0,004	0,125	0,571	0,305	0,013	0,070	0,004	0,086	0,007	0,214	0,188
TN norm	0,275	--	0,296	0,484	0,705	0,668	0,407	0,543	0,175	0,546	0,396	0,554	0,663
FP norm	0,725	--	0,704	0,516	0,295	0,332	0,593	0,457	0,825	0,454	0,604	0,446	0,338
TPR / sensitivity	0,998	--	0,996	0,875	0,429	0,695	0,988	0,930	0,996	0,914	0,993	0,786	0,813
FNR	0,002	--	0,004	0,125	0,571	0,305	0,013	0,070	0,004	0,086	0,007	0,214	0,188
TNR / specificity	0,275	--	0,296	0,484	0,705	0,668	0,407	0,543	0,175	0,546	0,396	0,554	0,663
FPR	0,725	--	0,704	0,516	0,295	0,332	0,593	0,457	0,825	0,454	0,604	0,446	0,338
Cases +	560	--	560	560	560	560	560	560	560	560	560	560	560
Cases -	560	--	560	560	560	560	560	560	560	560	560	560	560
Total	1 120	--	1 120	1 120	1 120	1 120	1 120	1 120	1 120	1 120	1 120	1 120	1 120
Accuracy	0,637	--	0,646	0,679	0,567	0,681	0,697	0,737	0,586	0,730	0,695	0,670	0,738
Accuracy norm	**0,637**	**--**	**0,646**	**0,679**	**0,567**	**0,681**	**0,697**	**0,737**	**0,586**	**0,730**	**0,695**	**0,670**	**0,738**
Precision	0,579	--	0,586	0,629	0,593	0,677	0,625	0,671	0,547	0,668	0,622	0,638	0,707
Precision norm	0,579	--	0,586	0,629	0,593	0,677	0,625	0,671	0,547	0,668	0,622	0,638	0,707
NPV	0,994	--	0,988	0,795	0,552	0,686	0,970	0,886	0,980	0,864	0,982	0,721	0,779
NPV norm	0,994	--	0,988	0,795	0,552	0,686	0,970	0,886	0,980	0,864	0,982	0,721	0,779
F-measure	0,733	--	0,738	0,732	0,497	0,685	0,765	0,779	0,706	0,772	0,765	0,704	0,756
AUC	**0,330**	**--**	**0,796**	**0,739**	**0,595**	**0,682**	**0,754**	**0,795**	**0,593**	**0,766**	**0,696**	**0,678**	**0,790**
MCC	0,396	--	0,410	0,390	0,139	0,363	0,485	0,513	0,301	0,495	0,485	0,349	0,480
MCC norm	**0,396**	**--**	**0,410**	**0,390**	**0,139**	**0,363**	**0,485**	**0,513**	**0,301**	**0,495**	**0,485**	**0,349**	**0,480**

Obrázek B.2: Statistické metriky pro model (dataset) 2.

Dataset 3: nástroje FATHMM,GWAVA,CADD													
	NÁSTROJE					**KONSENZUS - METODY STROJOVÉHO UČENÍ**							
	MT	SIFT	FATHMM	CADD	GWAVA	IBk	NaiveBayes	LogisticRegression	Voted	Multilayer	SVM_LINEAR	SVM_POLYNOMIAL	RandomForest
TP	--	--	1204	1164	557	1159	1199	1183	1205	1174	1190	1183	1173
FN	--	--	23	63	670	68	28	44	22	53	37	44	54
TN	--	--	1085	1172	955	1155	1115	1158	1011	1190	1146	1169	1166
FP	--	--	142	55	272	72	112	69	216	37	81	58	61
TP norm	--	--	0,981	0,949	0,454	0,945	0,977	0,964	0,982	0,957	0,970	0,964	0,956
FN norm	--	--	0,019	0,051	0,546	0,055	0,023	0,036	0,018	0,043	0,030	0,036	0,044
TN norm	--	--	0,884	0,955	0,778	0,941	0,909	0,944	0,824	0,970	0,934	0,953	0,950
FP norm	--	--	0,116	0,045	0,222	0,059	0,091	0,056	0,176	0,030	0,066	0,047	0,050
TPR / sensitivity	--	--	0,981	0,949	0,454	0,945	0,977	0,964	0,982	0,957	0,970	0,964	0,956
FNR	--	--	0,019	0,051	0,546	0,055	0,023	0,036	0,018	0,043	0,030	0,036	0,044
TNR / specificity	--	--	0,884	0,955	0,778	0,941	0,909	0,944	0,824	0,970	0,934	0,953	0,950
FPR	--	--	0,116	0,045	0,222	0,059	0,091	0,056	0,176	0,030	0,066	0,047	0,050
Cases +	--	--	1 227	1 227	1 227	1 227	1 227	1 227	1 227	1 227	1 227	1 227	1 227
Cases -	--	--	1 227	1 227	1 227	1 227	1 227	1 227	1 227	1 227	1 227	1 227	1 227
Total	--	--	2 454	2 454	2 454	2 454	2 454	2 454	2 454	2 454	2 454	2 454	2 454
Accuracy	--	--	0,933	0,952	0,616	0,943	0,943	0,954	0,903	0,963	0,952	0,958	0,953
Accuracy norm	**--**	**--**	**0,933**	**0,952**	**0,616**	**0,943**	**0,943**	**0,954**	**0,903**	**0,963**	**0,952**	**0,958**	**0,953**
Precision	--	--	0,895	0,955	0,672	0,942	0,915	0,945	0,848	0,969	0,936	0,953	0,951
Precision norm	--	--	0,895	0,955	0,672	0,942	0,915	0,945	0,848	0,969	0,936	0,953	0,951
NPV	--	--	0,979	0,949	0,588	0,944	0,976	0,963	0,979	0,957	0,969	0,964	0,956
NPV norm	--	--	0,979	0,949	0,588	0,944	0,976	0,963	0,979	0,957	0,969	0,964	0,956
F-measure	--	--	0,936	0,952	0,542	0,943	0,945	0,954	0,910	0,963	0,953	0,959	0,953
AUC	**--**	**--**	**0,980**	**0,987**	**0,629**	**0,945**	**0,988**	**0,988**	**0,898**	**0,970**	**0,951**	**0,958**	**0,982**
MCC	--	--	0,870	0,904	0,246	0,886	0,888	0,908	0,816	0,927	0,904	0,917	0,906
MCC norm	**--**	**--**	**0,870**	**0,904**	**0,246**	**0,886**	**0,888**	**0,908**	**0,816**	**0,927**	**0,904**	**0,917**	**0,906**

Obrázek B.3: Statistické metriky pro model (dataset) 3.

Dataset 5: nástroje FATHMM,GWAVA,CADD,MT,SIFT													
	NÁSTROJE					KONSENZUS - METODY STROJOVÉHO UČENÍ							
	MT	SIFT	FATHMM	CADD	GWAVA	IBk	NaiveBayes	LogisticRegression	Voted	Multilayer	SVM_LINEAR	SVM_POLYNOMIAL	RandomForest
TP	13445	14004	13246	11562	4229	10522	12657	12580	12972	11479	12730	9049	11656
FN	681	122	880	2564	9897	3604	1469	1546	1154	2647	1396	5077	2470
TN	8019	866	6666	8062	11015	10252	9201	9484	8643	11135	9117	8989	10604
FP	6107	13260	7460	6064	3111	3874	4925	4642	5483	2991	5009	5137	3522
TP norm	0,952	0,991	0,938	0,818	0,299	0,745	0,896	0,891	0,918	0,813	0,901	0,641	0,825
FN norm	0,048	0,009	0,062	0,182	0,701	0,255	0,104	0,109	0,082	0,187	0,099	0,359	0,175
TN norm	0,568	0,061	0,472	0,571	0,780	0,726	0,651	0,671	0,612	0,788	0,645	0,636	0,751
FP norm	0,432	0,939	0,528	0,429	0,220	0,274	0,349	0,329	0,388	0,212	0,355	0,364	0,249
TPR / sensitivity	0,952	0,991	0,938	0,818	0,299	0,745	0,896	0,891	0,918	0,813	0,901	0,641	0,825
FNR	0,048	0,009	0,062	0,182	0,701	0,255	0,104	0,109	0,082	0,187	0,099	0,359	0,175
TNR / specificity	0,568	0,061	0,472	0,571	0,780	0,726	0,651	0,671	0,612	0,788	0,645	0,636	0,751
FPR	0,432	0,939	0,528	0,429	0,220	0,274	0,349	0,329	0,388	0,212	0,355	0,364	0,249
Cases +	14 126	14 126	14 126	14 126	14 126	14 126	14 126	14 126	14 126	14 126	14 126	14 126	14 126
Cases -	14 126	14 126	14 126	14 126	14 126	14 126	14 126	14 126	14 126	14 126	14 126	14 126	14 126
Total	28 252	28 252	28 252	28 252	28 252	28 252	28 252	28 252	28 252	28 252	28 252	28 252	28 252
Accuracy	0,760	0,526	0,705	0,695	0,540	0,735	0,774	0,781	0,765	0,800	0,773	0,638	0,788
Accuracy norm	**0,760**	**0,526**	**0,705**	**0,695**	**0,540**	**0,735**	**0,774**	**0,781**	**0,765**	**0,800**	**0,773**	**0,638**	**0,788**
Precision	0,688	0,514	0,640	0,656	0,576	0,731	0,720	0,730	0,703	0,793	0,718	0,638	0,768
Precision norm	0,688	0,514	0,640	0,656	0,576	0,731	0,720	0,730	0,703	0,793	0,718	0,638	0,768
NPV	0,922	0,877	0,883	0,759	0,527	0,740	0,862	0,860	0,882	0,808	0,867	0,639	0,811
NPV norm	0,922	0,877	0,883	0,759	0,527	0,740	0,862	0,860	0,882	0,808	0,867	0,639	0,811
F-measure	0,798	0,677	0,761	0,728	0,394	0,738	0,798	0,803	0,796	0,803	0,799	0,639	0,796
AUC	**0,620**	**0,629**	**0,811**	**0,766**	**0,557**	**0,734**	**0,841**	**0,856**	**0,763**	**0,820**	**0,772**	**0,660**	**0,855**
MCC	0,563	0,143	0,463	0,402	0,090	0,471	0,565	0,576	0,557	0,601	0,565	0,277	0,577
MCC norm	**0,563**	**0,143**	**0,463**	**0,402**	**0,090**	**0,471**	**0,565**	**0,576**	**0,557**	**0,601**	**0,565**	**0,277**	**0,577**

Obrázek B.4: Statistické metriky pro model (dataset) 5.

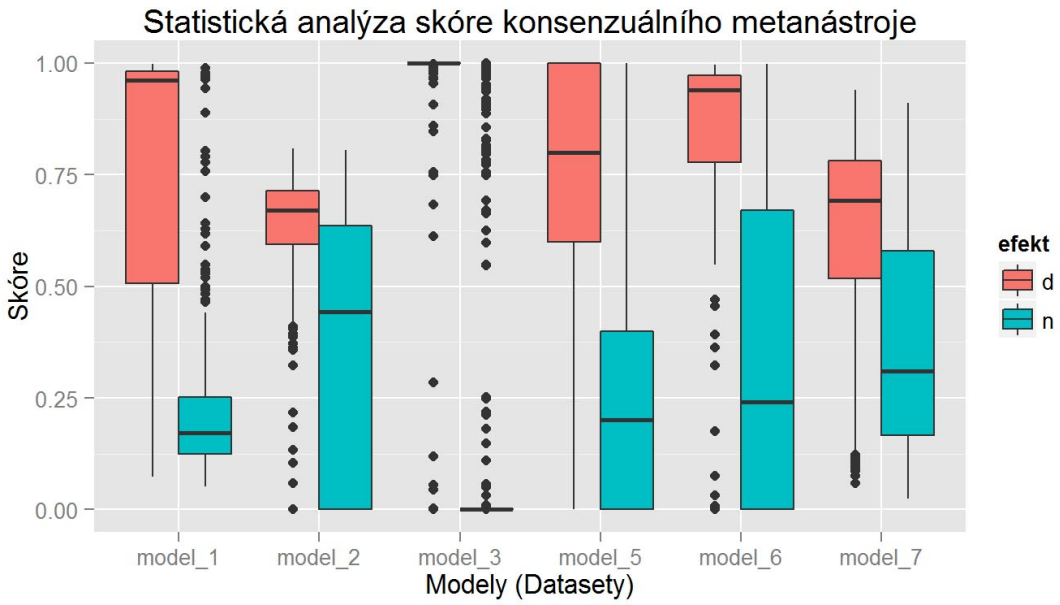
Dataset 6: nástroje FATHMM,GWAVA,CADD,MT													
	NÁSTROJE					**KONZENZUS - METODY STROJOVÉHO UČENÍ**							
	MT	SIFT	FATHMM	CADD	GWAVA	IBk	NaiveBayes	LogisticRegression	Voted	Multilayer	SVM_LINEAR	SVM_POLYNOMIAL	RandomForest
TP	106	--	82	82	27	83	94	86	104	78	84	87	77
FN	3	--	27	27	82	26	15	23	5	31	25	22	32
TN	30	--	69	72	98	79	73	73	23	79	72	81	72
FP	79	--	40	37	11	30	36	36	86	30	37	28	37
TP norm	0,972	--	0,752	0,752	0,248	0,761	0,862	0,789	0,954	0,716	0,771	0,798	0,706
FN norm	0,028	--	0,248	0,248	0,752	0,239	0,138	0,211	0,046	0,284	0,229	0,202	0,294
TN norm	0,275	--	0,633	0,661	0,899	0,725	0,670	0,670	0,211	0,725	0,661	0,743	0,661
FP norm	0,725	--	0,367	0,339	0,101	0,275	0,330	0,330	0,789	0,275	0,339	0,257	0,339
TPR / sensitivity	0,972	--	0,752	0,752	0,248	0,761	0,862	0,789	0,954	0,716	0,771	0,798	0,706
FNR	0,028	--	0,248	0,248	0,752	0,239	0,138	0,211	0,046	0,284	0,229	0,202	0,294
TNR / specificity	0,275	--	0,633	0,661	0,899	0,725	0,670	0,670	0,211	0,725	0,661	0,743	0,661
FPR	0,725	--	0,367	0,339	0,101	0,275	0,330	0,330	0,789	0,275	0,339	0,257	0,339
Cases +	109	--	109	109	109	109	109	109	109	109	109	109	109
Cases -	109	--	109	109	109	109	109	109	109	109	109	109	109
Total	218	--	218	218	218	218	218	218	218	218	218	218	218
Accuracy	0,624	--	0,693	0,706	0,573	0,743	0,766	0,729	0,583	0,720	0,716	0,771	0,683
Accuracy norm	**0,624**	**--**	**0,693**	**0,706**	**0,573**	**0,743**	**0,766**	**0,729**	**0,583**	**0,720**	**0,716**	**0,771**	**0,683**
Precision	0,573	--	0,672	0,689	0,711	0,735	0,723	0,705	0,547	0,722	0,694	0,757	0,675
Precision norm	0,573	--	0,672	0,689	0,711	0,735	0,723	0,705	0,547	0,722	0,694	0,757	0,675
NPV	0,909	--	0,719	0,727	0,544	0,752	0,830	0,760	0,821	0,718	0,742	0,786	0,692
NPV norm	0,909	--	0,719	0,727	0,544	0,752	0,830	0,760	0,821	0,718	0,742	0,786	0,692
F-measure	0,721	--	0,710	0,719	0,367	0,748	0,787	0,745	0,696	0,719	0,730	0,777	0,691
AUC	**0,468**	**--**	**0,808**	**0,742**	**0,598**	**0,764**	**0,828**	**0,820**	**0,635**	**0,797**	**0,721**	**0,765**	**0,786**
MCC	0,346	--	0,388	0,415	0,193	0,487	0,542	0,462	0,247	0,440	0,434	0,542	0,367
MCC norm	**0,346**	**--**	**0,388**	**0,415**	**0,193**	**0,487**	**0,542**	**0,462**	**0,247**	**0,440**	**0,434**	**0,542**	**0,367**

Obrázek B.5: Statistické metriky pro model (dataset) 6.

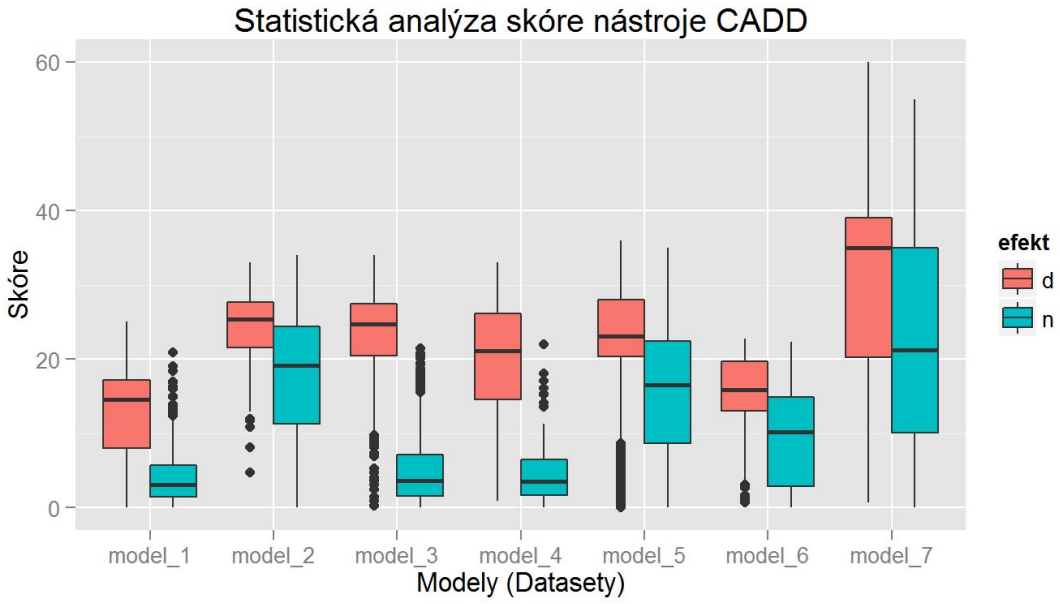
Dataset 7: nástroje FATHMM,GWAVA,CADD,MT													
	NÁSTROJE					KONSENZUS - METODY STROJOVÉHO UČENÍ							
	MT	SIFT	FATHMM	CADD	GWAVA	IBk	NaiveBayes	LogisticRegression	Voted	Multilayer	SVM_LINEAR	SVM_POLYNOMIAL	RandomForest
TP	1286	0,000	1062	638	397	863	1104	1015	1268	957	1049	750	968
FN	42	0,000	266	690	931	465	224	313	60	371	279	578	360
TN	227	0,000	787	1068	1161	873	828	906	345	932	857	914	818
FP	1101	0,000	541	260	167	455	500	422	983	396	471	414	510
TP norm	0,968	--	0,800	0,480	0,299	0,650	0,831	0,764	0,955	0,721	0,790	0,565	0,729
FN norm	0,032	--	0,200	0,520	0,701	0,350	0,169	0,236	0,045	0,279	0,210	0,435	0,271
TN norm	0,171	--	0,593	0,804	0,874	0,657	0,623	0,682	0,260	0,702	0,645	0,688	0,616
FP norm	0,829	--	0,407	0,196	0,126	0,343	0,377	0,318	0,740	0,298	0,355	0,312	0,384
TPR / sensitivity	0,968	--	0,800	0,480	0,299	0,650	0,831	0,764	0,955	0,721	0,790	0,565	0,729
FNR	0,032	--	0,200	0,520	0,701	0,350	0,169	0,236	0,045	0,279	0,210	0,435	0,271
TNR / specificity	0,171	--	0,593	0,804	0,874	0,657	0,623	0,682	0,260	0,702	0,645	0,688	0,616
FPR	0,829	--	0,407	0,196	0,126	0,343	0,377	0,318	0,740	0,298	0,355	0,312	0,384
Cases +	1 328	--	1 328	1 328	1 328	1 328	1 328	1 328	1 328	1 328	1 328	1 328	1 328
Cases -	1 328	--	1 328	1 328	1 328	1 328	1 328	1 328	1 328	1 328	1 328	1 328	1 328
Total	2 656	--	2 656	2 656	2 656	2 656	2 656	2 656	2 656	2 656	2 656	2 656	2 656
Accuracy	0,570	--	0,696	0,642	0,587	0,654	0,727	0,723	0,607	0,711	0,718	0,627	0,672
Accuracy norm	**0,570**	**--**	**0,696**	**0,642**	**0,587**	**0,654**	**0,727**	**0,723**	**0,607**	**0,711**	**0,718**	**0,627**	**0,672**
Precision	0,539	--	0,663	0,710	0,704	0,655	0,688	0,706	0,563	0,707	0,690	0,644	0,655
Precision norm	0,539	--	0,663	0,710	0,704	0,655	0,688	0,706	0,563	0,707	0,690	0,644	0,655
NPV	0,844	--	0,747	0,608	0,555	0,652	0,787	0,743	0,852	0,715	0,754	0,613	0,694
NPV norm	0,844	--	0,747	0,608	0,555	0,652	0,787	0,743	0,852	0,715	0,754	0,613	0,694
F-measure	0,692	--	0,725	0,573	0,420	0,652	0,753	0,734	0,709	0,714	0,737	0,602	0,690
AUC	**0,267**	**--**	**0,766**	**0,687**	**0,677**	**0,654**	**0,791**	**0,792**	**0,609**	**0,753**	**0,717**	**0,639**	**0,740**
MCC	0,231	--	0,401	0,301	0,212	0,307	0,465	0,448	0,298	0,423	0,440	0,255	0,347
MCC norm	**0,231**	**--**	**0,401**	**0,301**	**0,212**	**0,307**	**0,465**	**0,448**	**0,298**	**0,423**	**0,440**	**0,255**	**0,347**

Obrázek B.6: Statistické metriky pro model (dataset) 7.

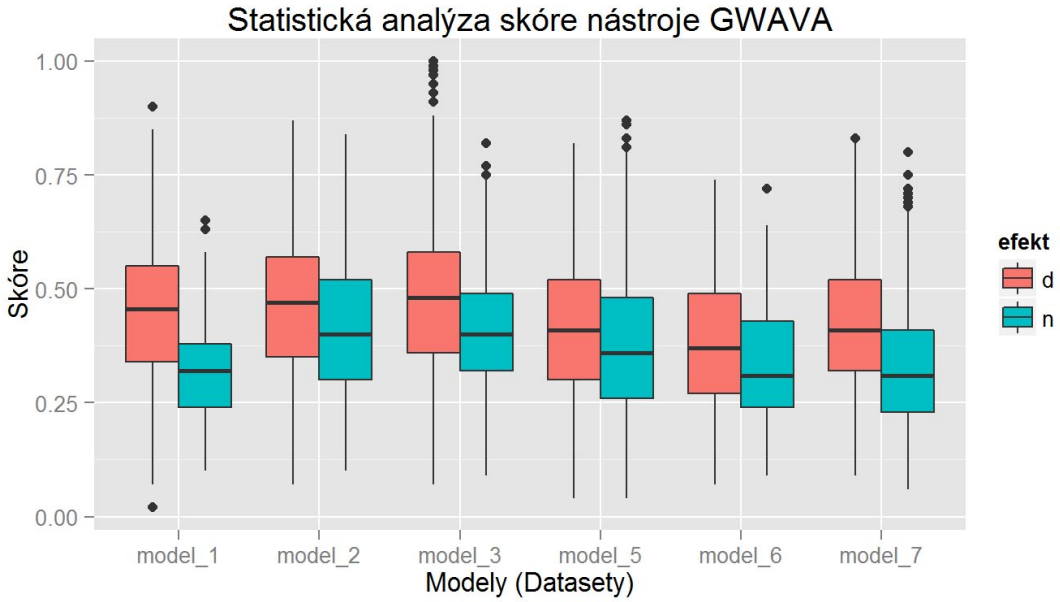
B.2 Krabicové grafy



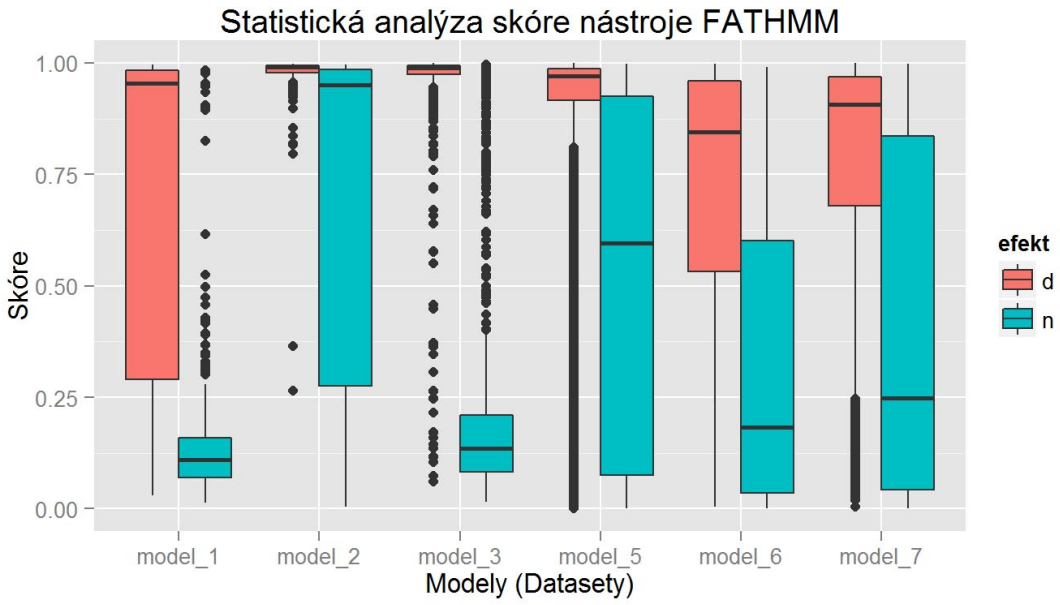
Obrázek B.7: Krabicové grafy pro nový metanástroj.



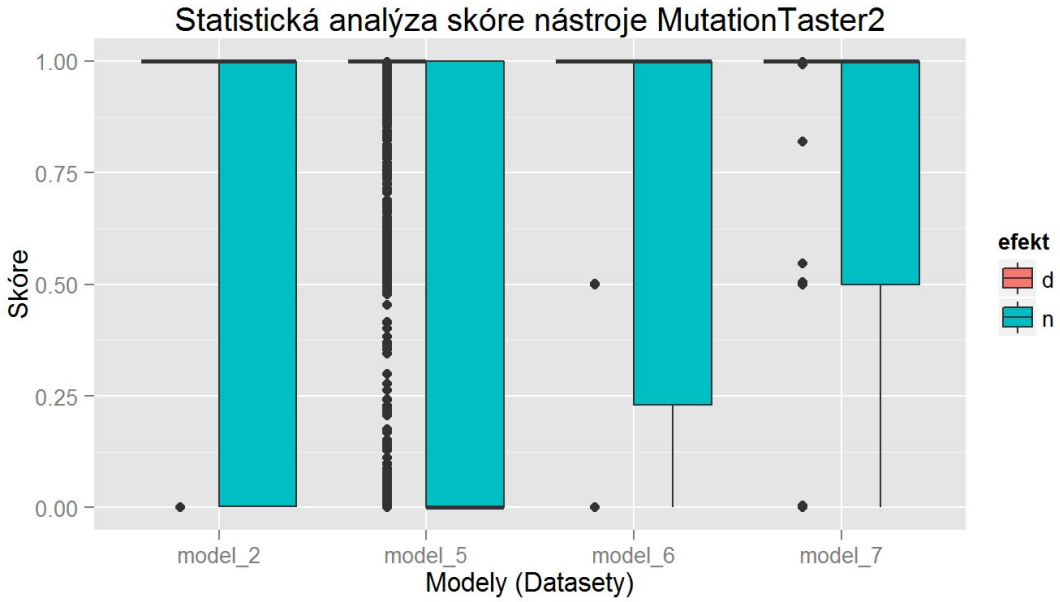
Obrázek B.8: Krabicové grafy pro nástroj CADD.



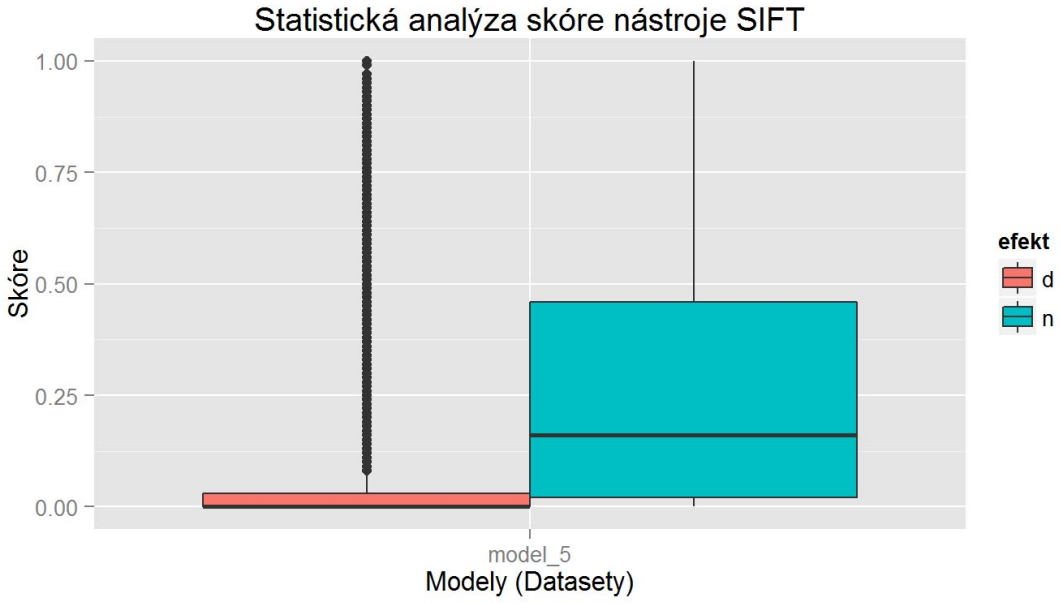
Obrázek B.9: Krabicové grafy pro nástroj GWAVA.



Obrázek B.10: Krabicové grafy pro nástroj FATHMM.

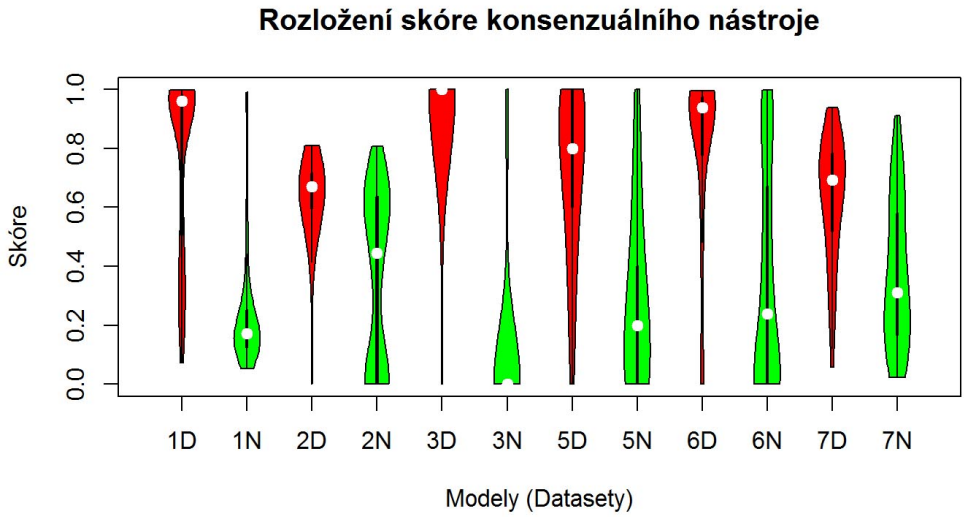


Obrázek B.11: Krabicové grafy pro nástroj MutationTaster2.

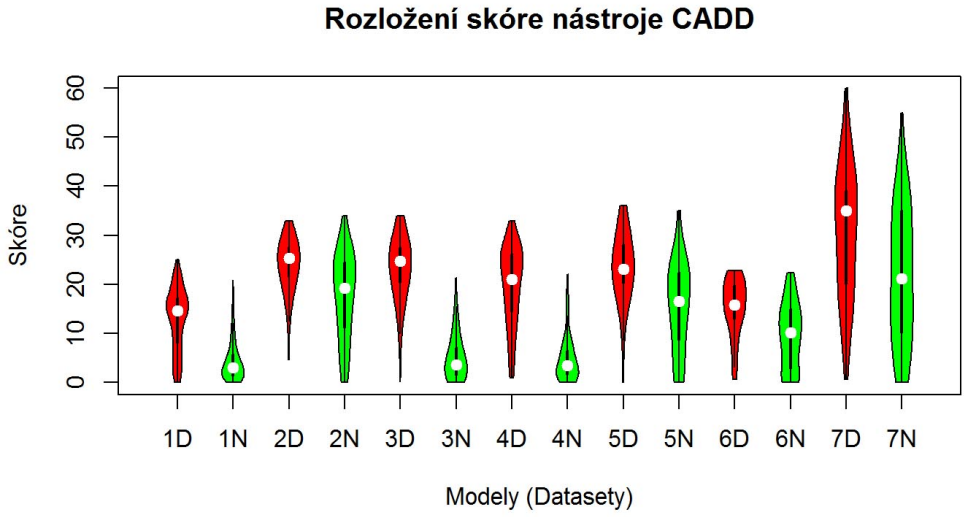


Obrázek B.12: Krabicové grafy pro nástroj SIFT.

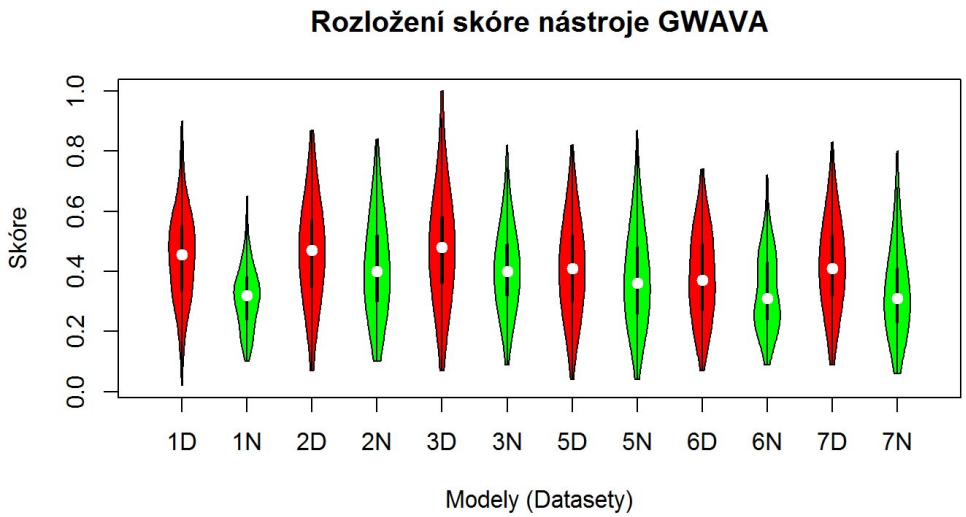
B.3 Houslové grafy



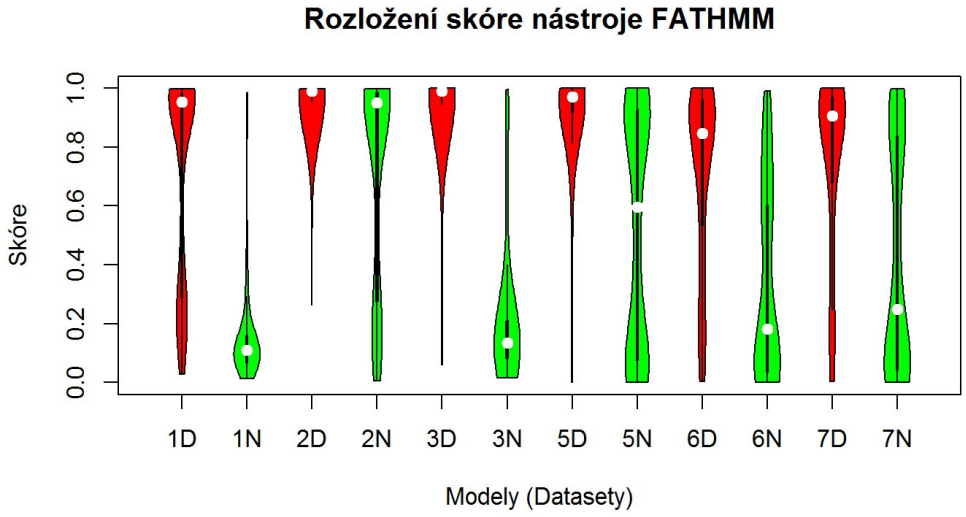
Obrázek B.13: Houslové grafy pro nový metanástroj.



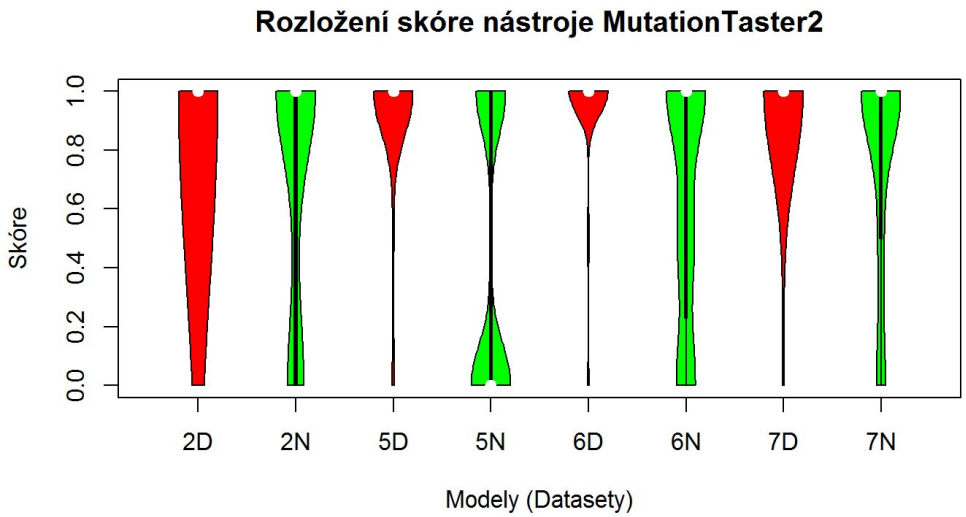
Obrázek B.14: Houslové grafy pro nástroj CADD.



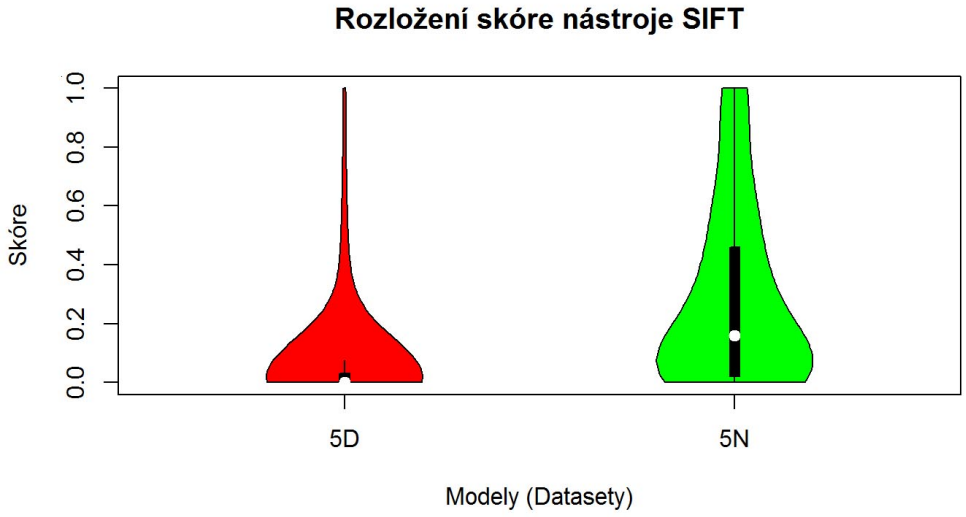
Obrázek B.15: Houslové grafy pro nástroj GWAVA.



Obrázek B.16: Houslové grafy pro nástroj FATHMM.

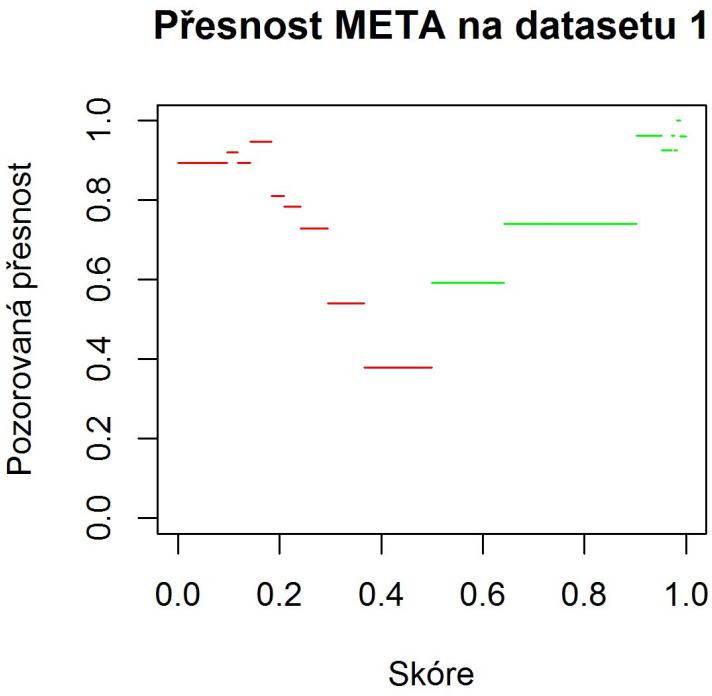


Obrázek B.17: Houslové grafy pro nástroj MutationTaster2.

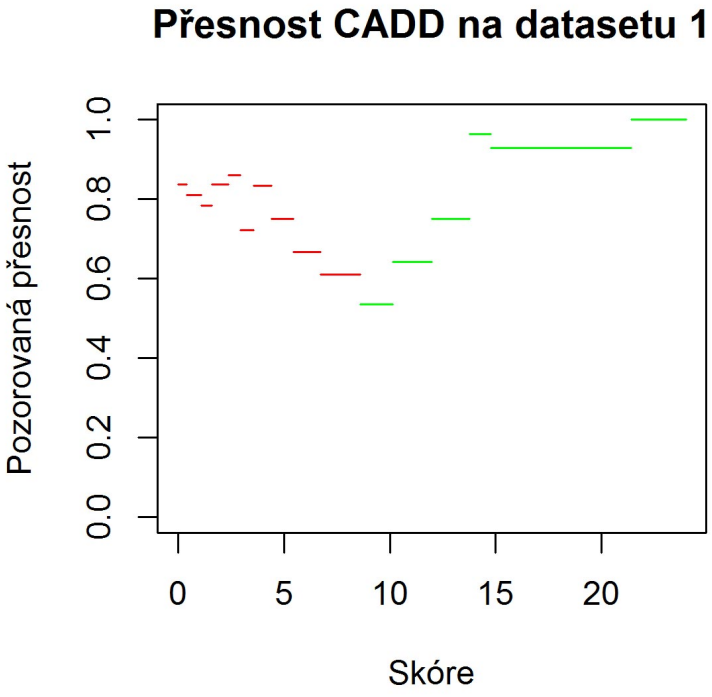


Obrázek B.18: Houslové grafy pro nástroj SIFT.

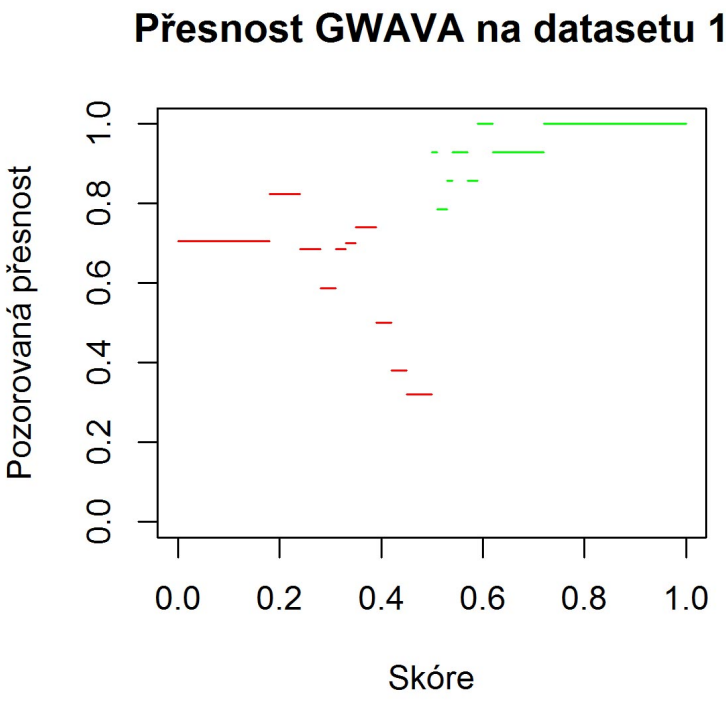
B.4 Grafy přesnosti podle skóre



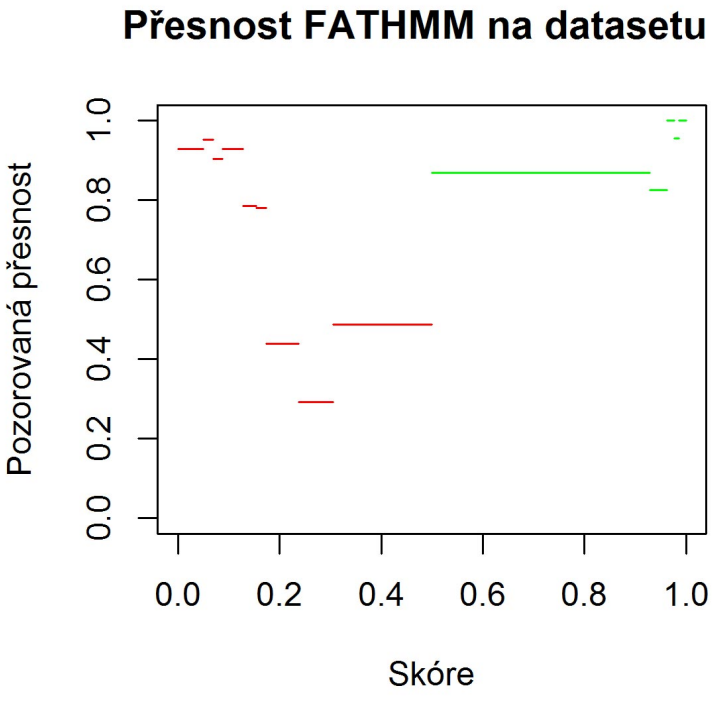
Obrázek B.19: Závislost přesnosti metanástroje na intervalu výstupního skóre na datasetu 1.



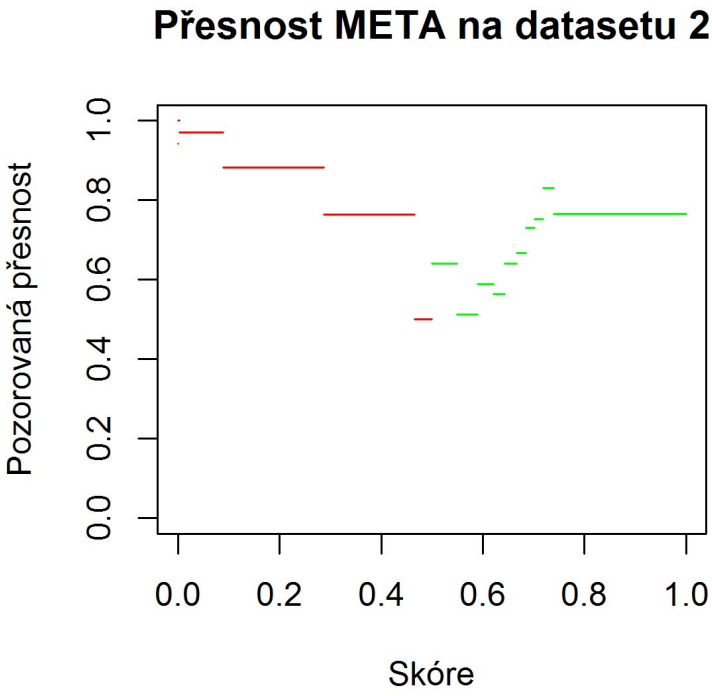
Obrázek B.20: Závislost přesnosti CADD na intervalu jeho skóre na datasetu 1.



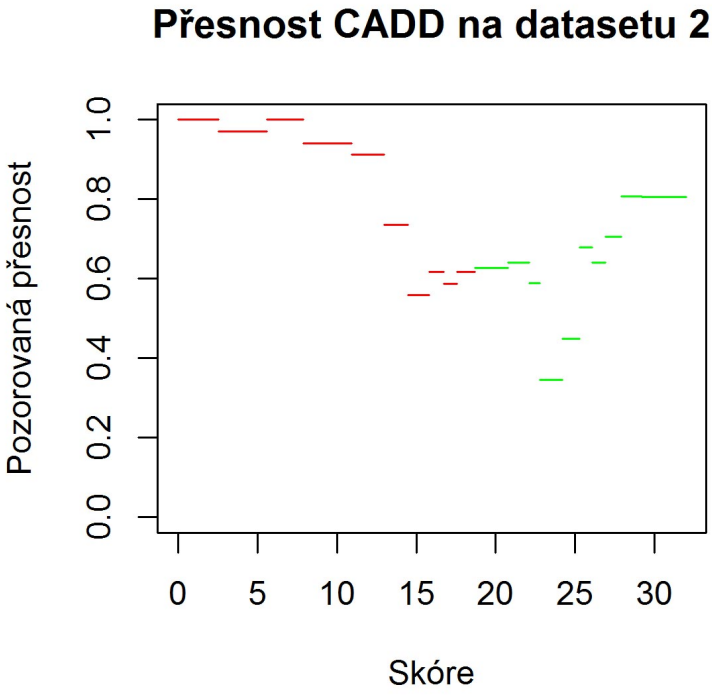
Obrázek B.21: Závislost přesnosti GWAVA na intervalu jeho skóre na datasetu 1.



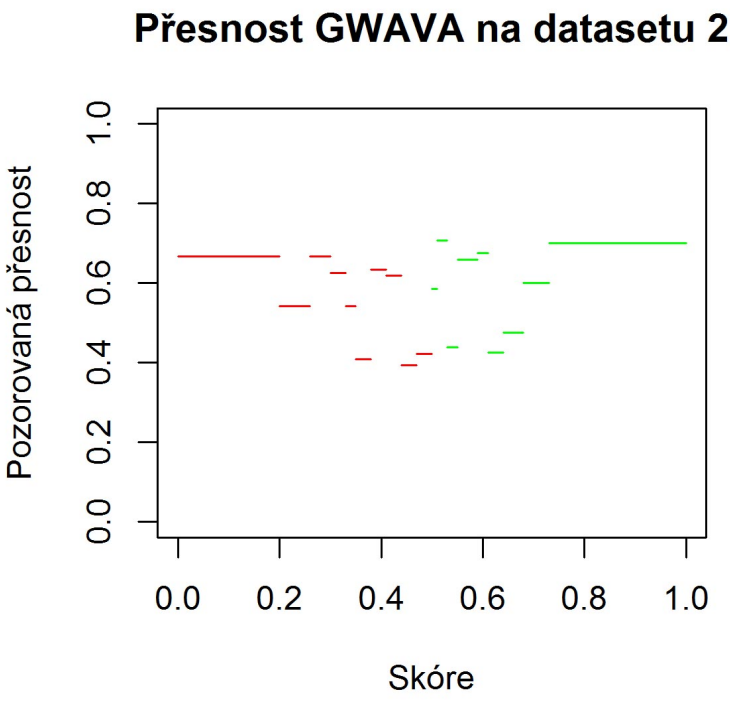
Obrázek B.22: Závislost přesnosti FATHMM na intervalu jeho skóre na datasetu 1.



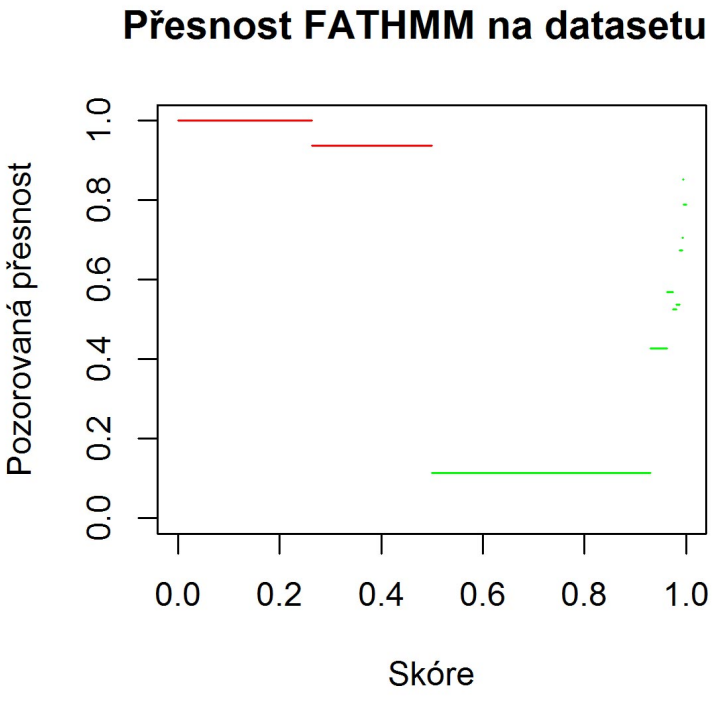
Obrázek B.23: Závislost přesnosti metanástroje na intervalu výstupního skóre na datasetu 2.



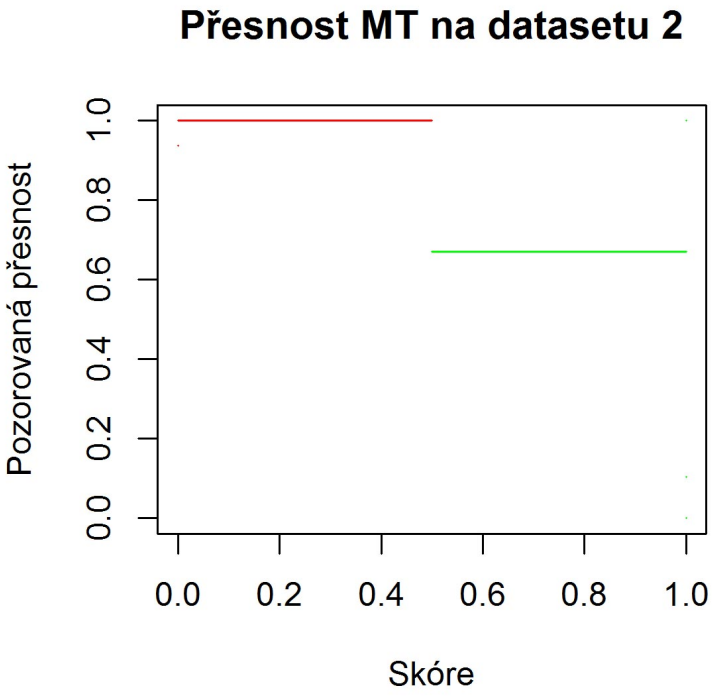
Obrázek B.24: Závislost přesnosti CADD na intervalu jeho skóre na datasetu 2.



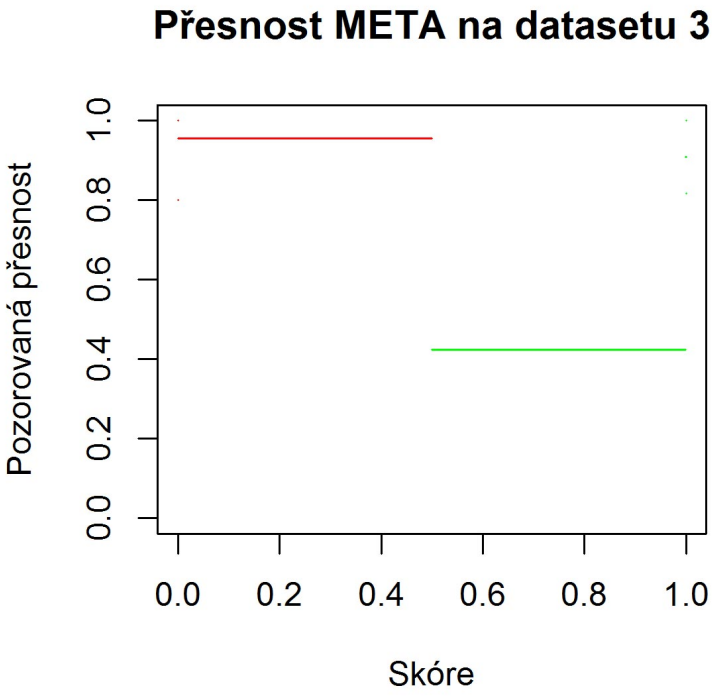
Obrázek B.25: Závislost přesnosti GWAVA na intervalu jeho skóre na datasetu 2.



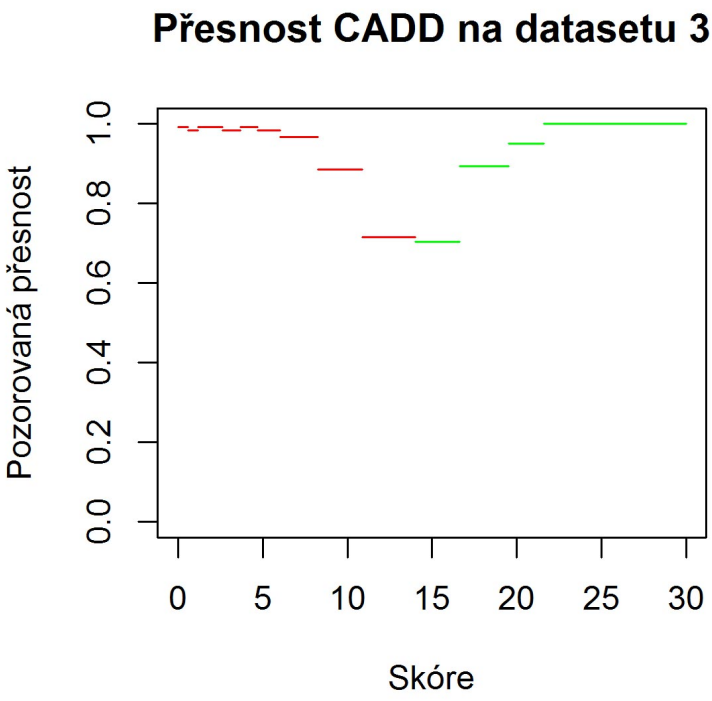
Obrázek B.26: Závislost přesnosti FATHMM na intervalu jeho skóre na datasetu 2.



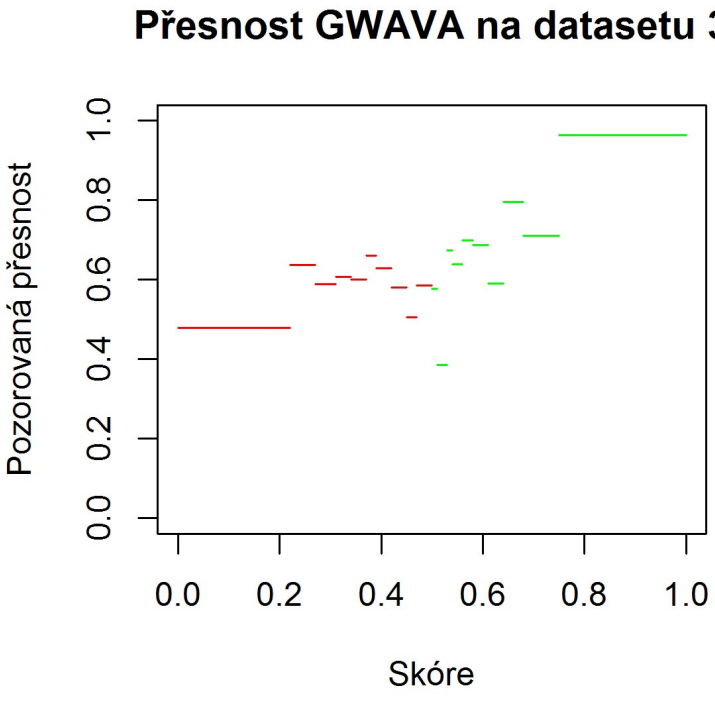
Obrázek B.27: Závislost přesnosti MutationTaster2 na intervalu jeho skóre na datasetu 2.



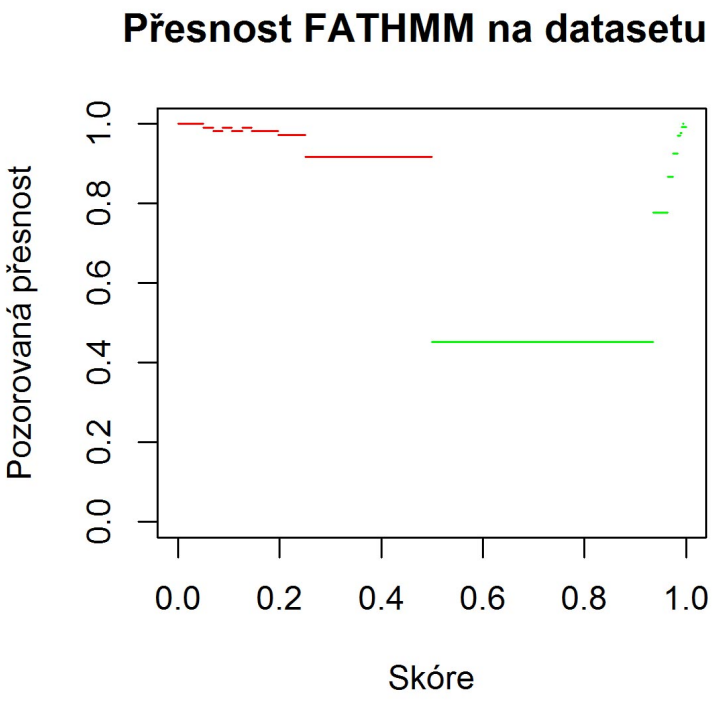
Obrázek B.28: Závislost přesnosti metanástroje na intervalu výstupního skóre na datasetu 3.



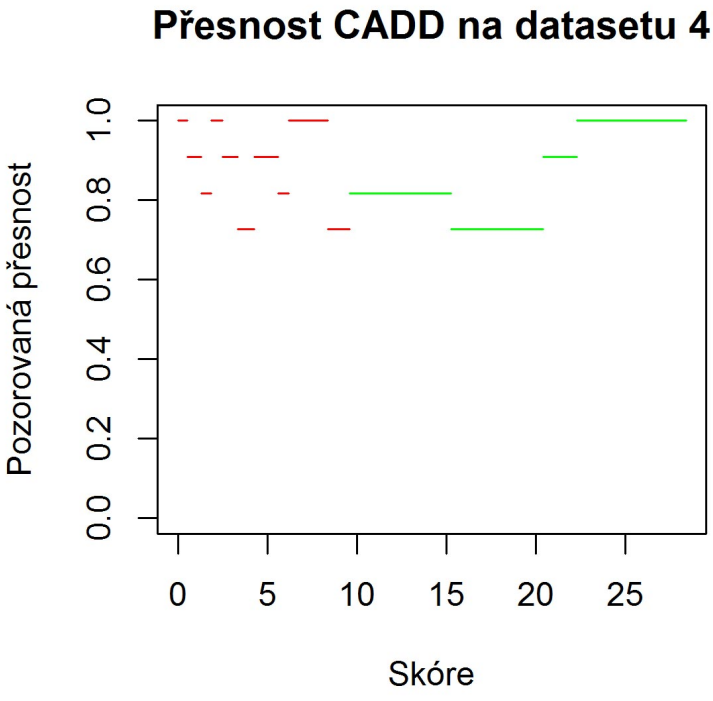
Obrázek B.29: Závislost přesnosti CADD na intervalu jeho skóre na datasetu 3.



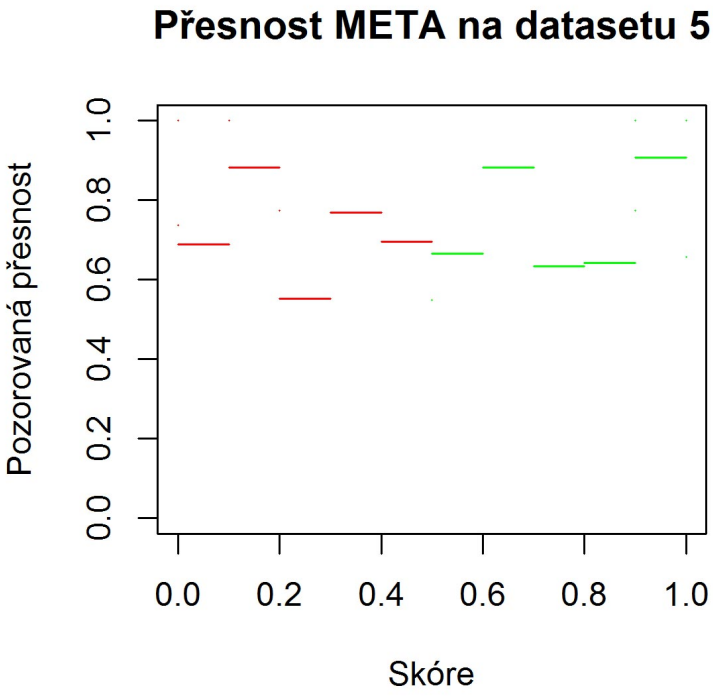
Obrázek B.30: Závislost přesnosti GWAVA na intervalu jeho skóre na datasetu 3.



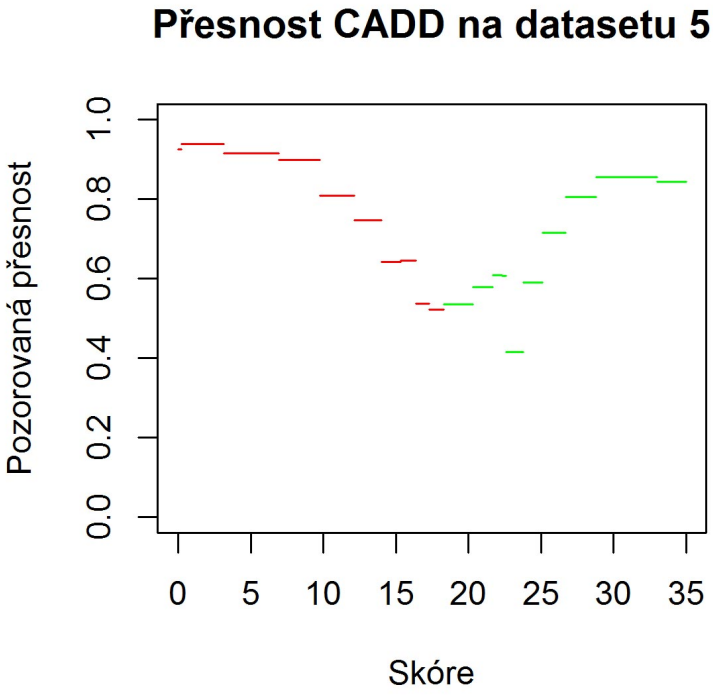
Obrázek B.31: Závislost přesnosti FATHMM na intervalu jeho skóre na datasetu 3.



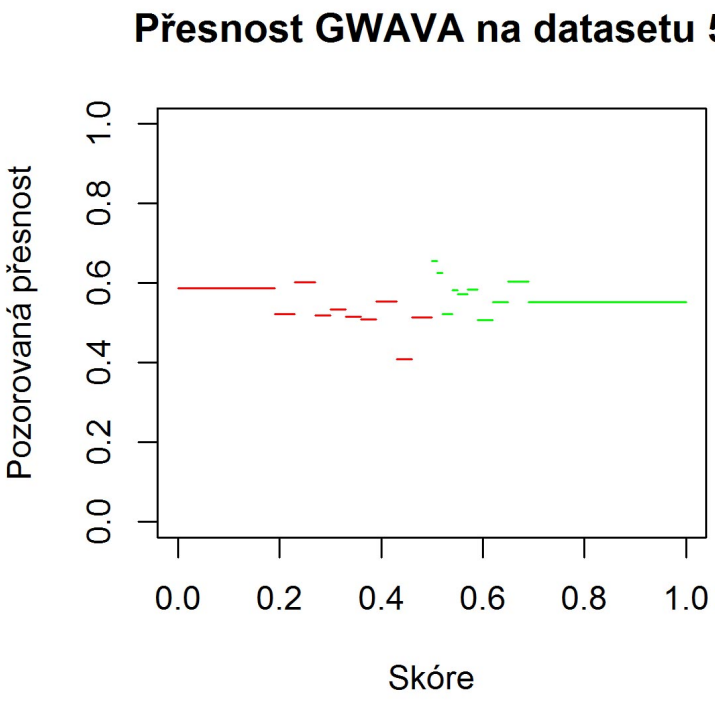
Obrázek B.32: Závislost přesnosti CADD na intervalu jeho skóre na datasetu 4.



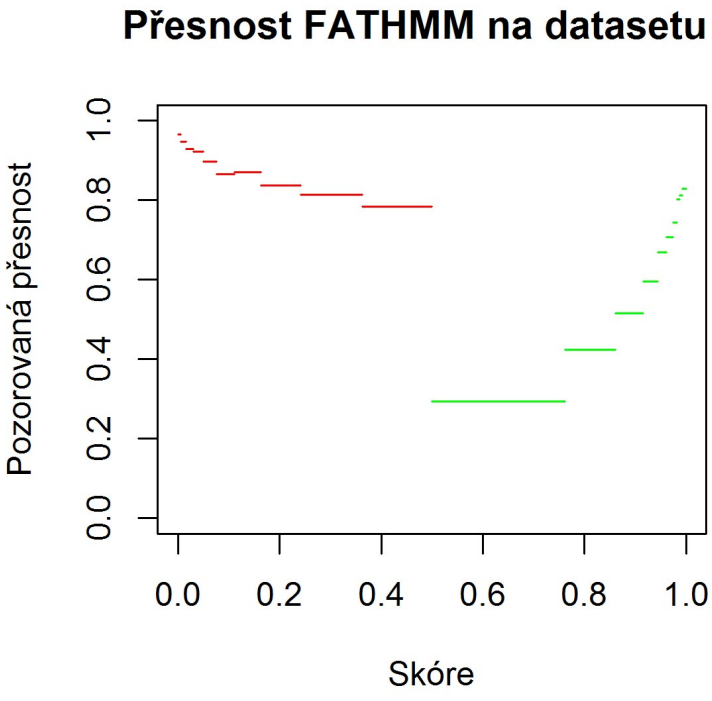
Obrázek B.33: Závislost přesnosti metanástroje na intervalu výstupního skóre na datasetu 5.



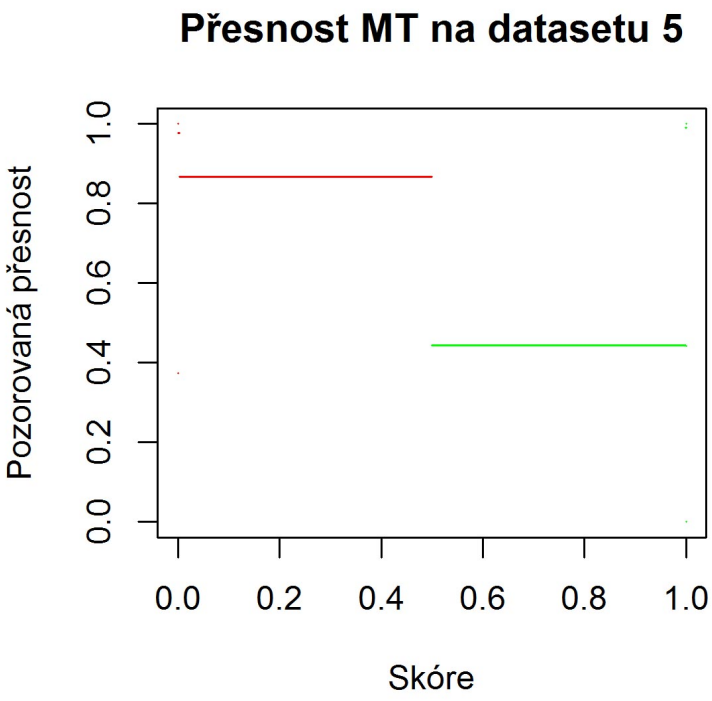
Obrázek B.34: Závislost přesnosti CADD na intervalu jeho skóre na datasetu 5.



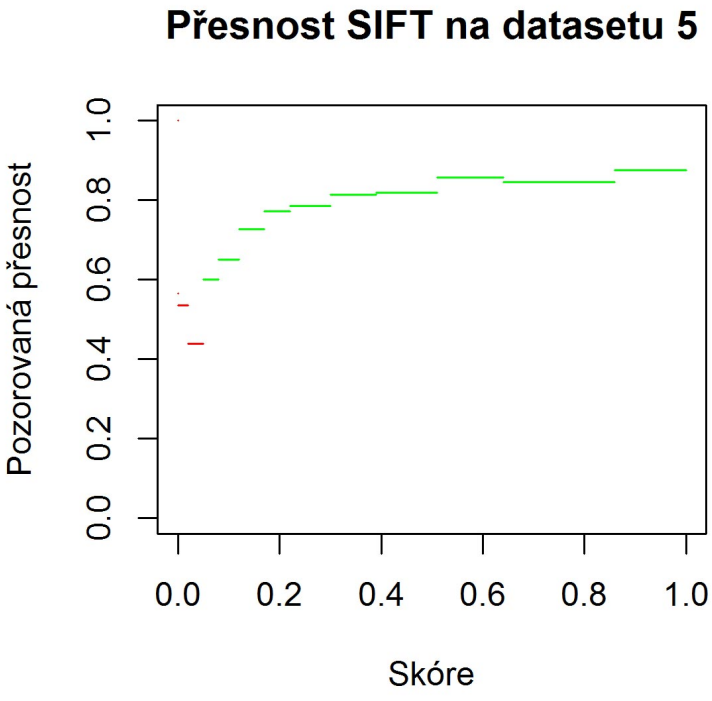
Obrázek B.35: Závislost přesnosti GWAVA na intervalu jeho skóre na datasetu 5.



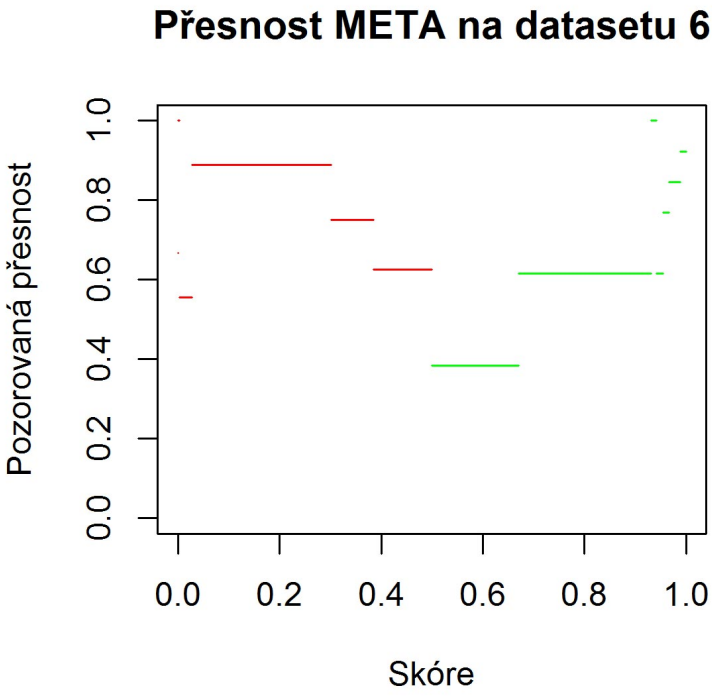
Obrázek B.36: Závislost přesnosti FATHMM na intervalu jeho skóre na datasetu 5.



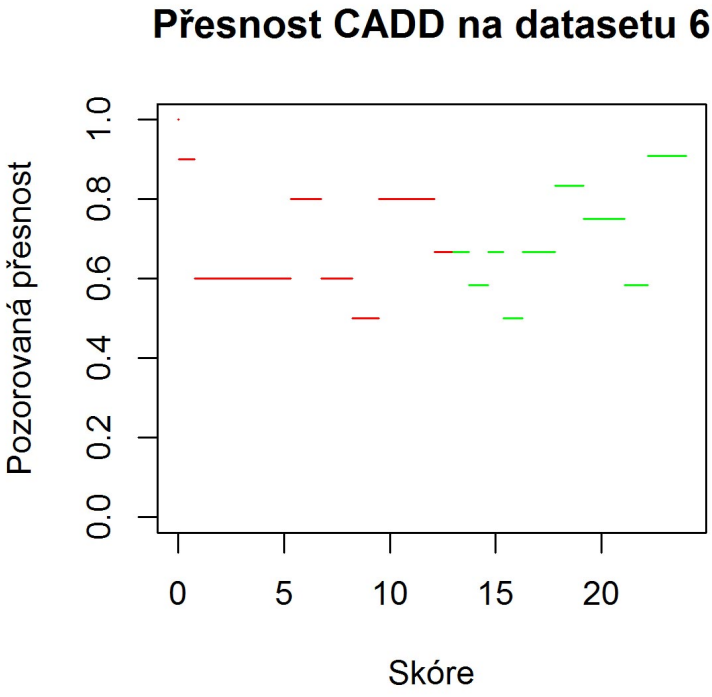
Obrázek B.37: Závislost přesnosti MutationTaster2 na intervalu jeho skóre na datasetu 5.



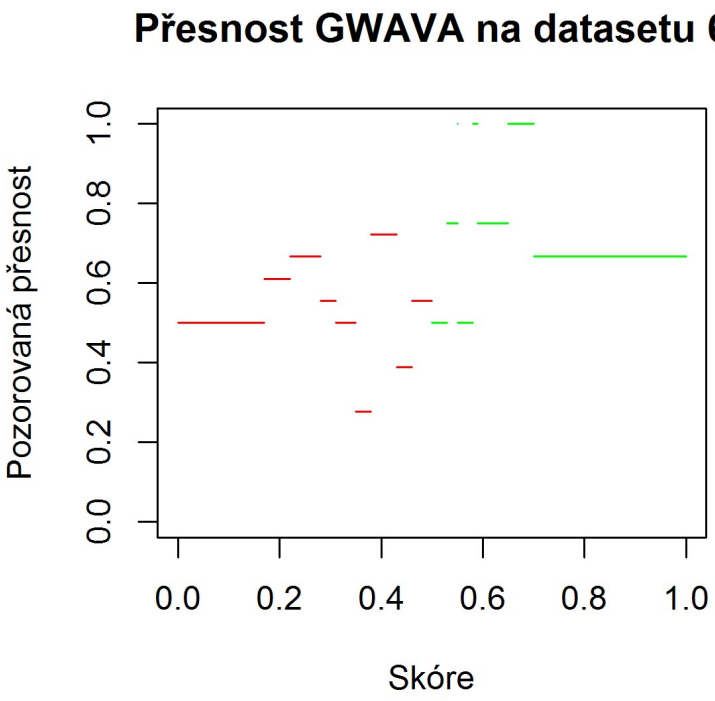
Obrázek B.38: Závislost přesnosti SIFT na intervalu jeho skóre na datasetu 5.



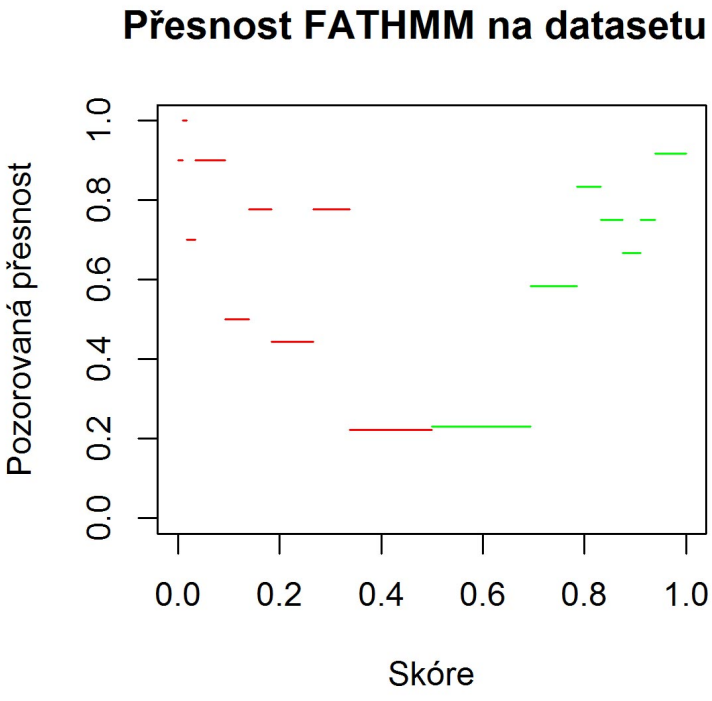
Obrázek B.39: Závislost přesnosti metanástroje na intervalu výstupního skóre na datasetu 6.



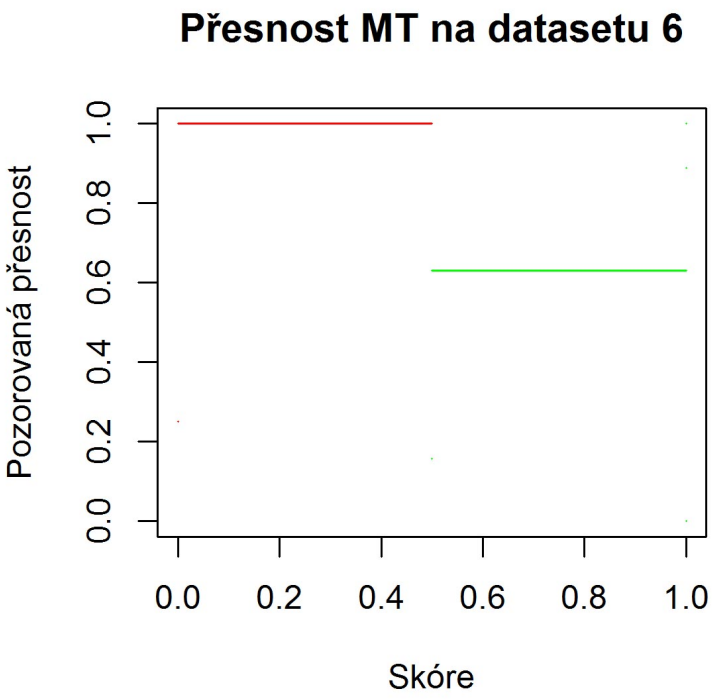
Obrázek B.40: Závislost přesnosti CADD na intervalu jeho skóre na datasetu 6.



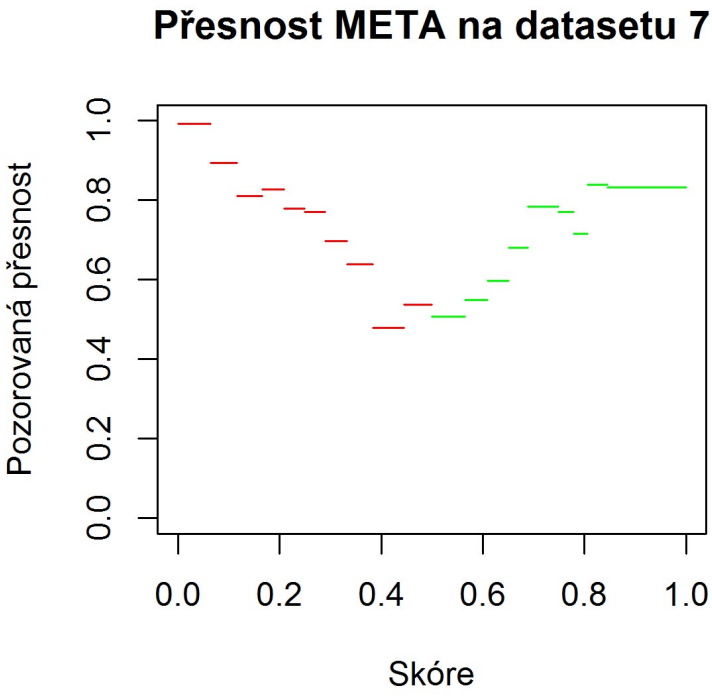
Obrázek B.41: Závislost přesnosti GWAVA na intervalu jeho skóre na datasetu 6.



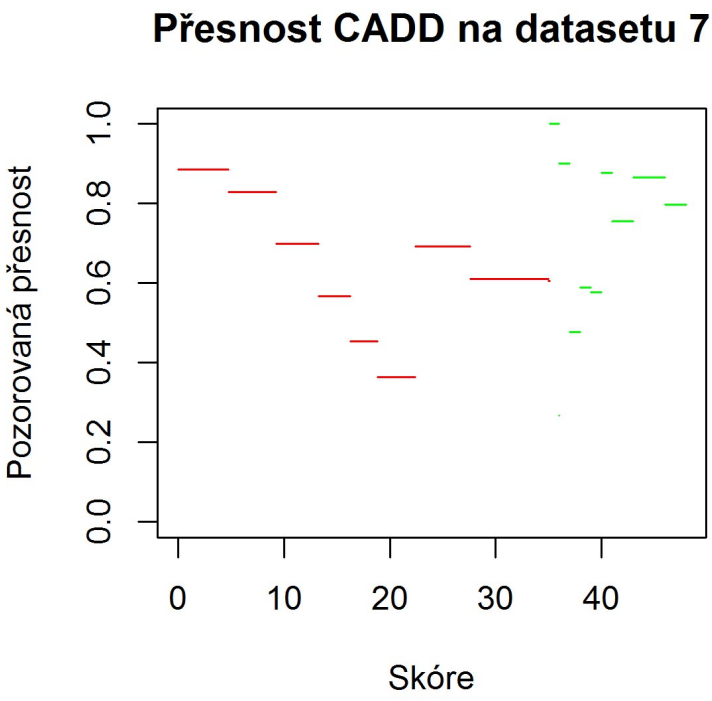
Obrázek B.42: Závislost přesnosti FATHMM na intervalu jeho skóre na datasetu 6.



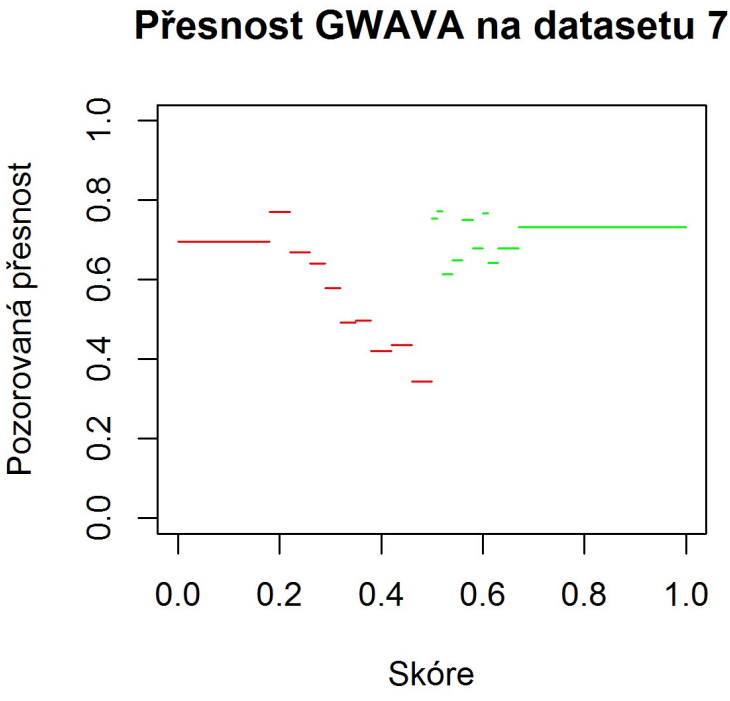
Obrázek B.43: Závislost přesnosti MutationTaster2 na intervalu jeho skóre na datasetu 6.



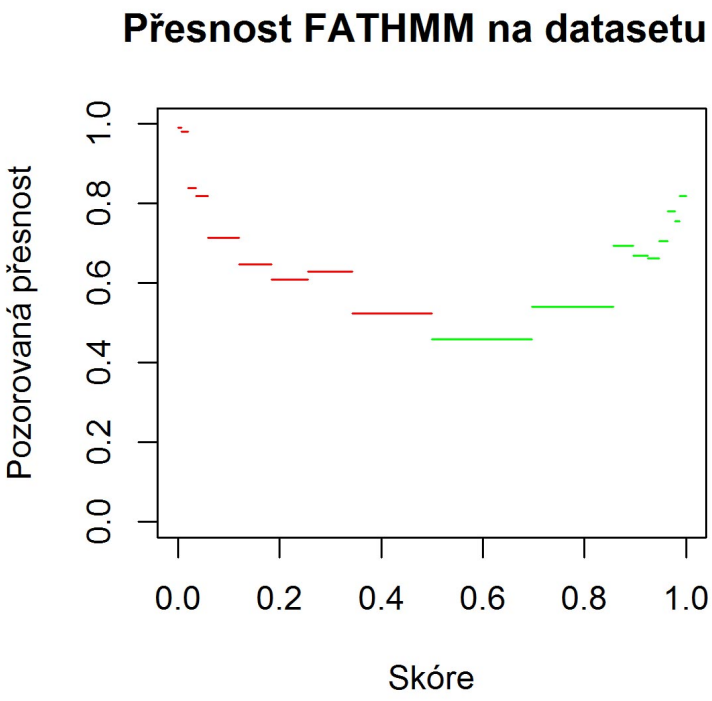
Obrázek B.44: Závislost přesnosti metanástroje na intervalu výstupního skóre na datasetu 7.



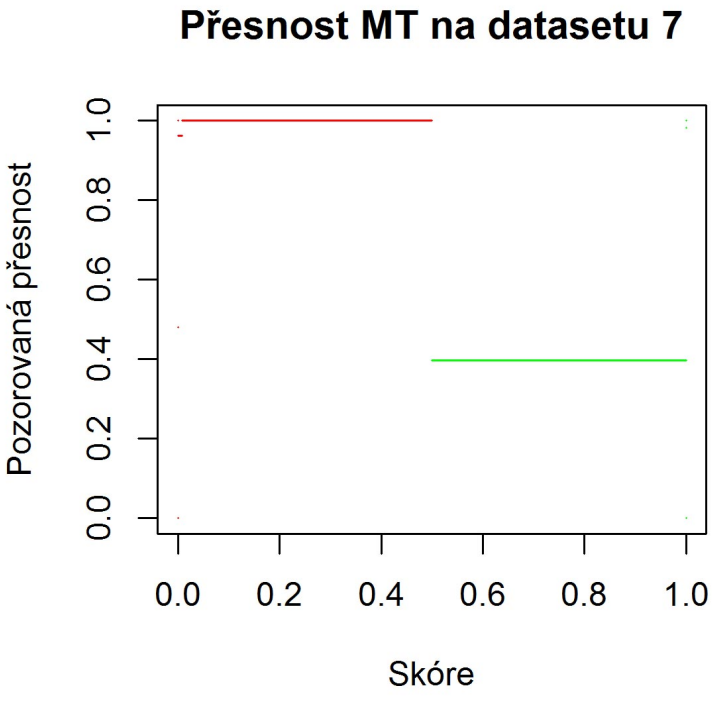
Obrázek B.45: Závislost přesnosti CADD na intervalu jeho skóre na datasetu 7.



Obrázek B.46: Závislost přesnosti GWAVA na intervalu jeho skóre na datasetu 7.



Obrázek B.47: Závislost přesnosti FATHMM na intervalu jeho skóre na datasetu 7.



Obrázek B.48: Závislost přesnosti MutationTaster2 na intervalu jeho skóre na datasetu 7.

Dodatek C

Webové rozhraní

Obrázek C.1: Snímek úvodní stránky webového rozhraní implementovaného klasifikátoru.

Dodatek D

Obsah CD

/analysis	výsledky, tabulky, grafy, experimenty
/grafy	skripty pro generování grafů v jazyce R
/summary	souhrnné výsledky trénování a validace
/ROC	soubory s ROC křivkami
/dataset	složka pro podklady k vytvoření a ohodnocení datasetu
/annovar	skripty pro anotaci mutací programem ANNOVAR
/create	skripty pro extrakci škodlivých a neutrálních mutací z databází
/db	skripty pro komunikaci s MySQL databází a SQL skript pro její vytvoření
/evaluate	v podsložkách podklady k ohodnocení datasetu nástroji
/rozdeleni	podklady k analýze pokrytí datasetu nástroji a jeho rozdělení na trénovací podmnožiny
/doc	zdrojové soubory pro LaTeX a pdf verze práce
/train	složka pro skripty související s trénováním
/datasets	konstrukce trénovacích podmnožin a převod do ARFF
/weka	Java projekt pro trénování a validaci modelů a vytváření vstupů pro ROC křivky, další grafy a analýzy
/web	projekt webového rozhraní v prostředí Eclipse