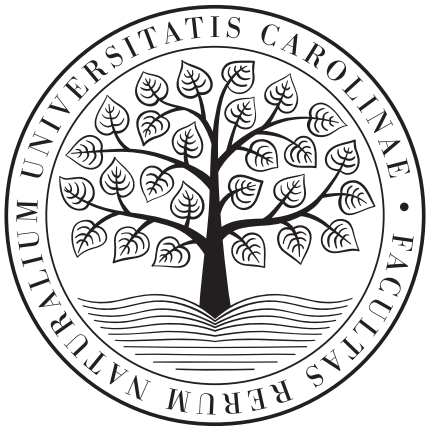
Univerzita Karlova
Přírodovědecká fakulta

BAKALÁŘSKÁ PRÁCE



Magdaléna Turinská

Vizualizace funkcí podobných proteinů

Katedra fyzikální a makromolekulární chemie (2600)

Vedoucí bakalářské práce:	prof. RNDr. Jiří Vondrášek, CSc.
Studijní program:	Bioinformatika (B0688A140003)
Studijní obor:	B-BINF (0688RA140003)

Praha 2024

Prohlašuji, že jsem tuto bakalářskou práci vypracovala samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Dále dle opatření děkana č. 13/2023 vydaném Přírodovědeckou fakultou Univerzity Karlovi, které nabylo účinnosti 1.10.2023, přiznávám využitím generativních jazykových modelů typu GPT při psaní této práce včetně její programové přílohy. Každé využití AI modelů je citováno včetně záznamu v seznamu použitých zdrojů včetně literatury a forma citace vychází z Brown University Library/AI Guides/Generative Artificial Intelligence/Citation and Attribution (verze stránky 29 leden 2024 8:56). Tato práce ani její podstatná část nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ....

Podpis autora

Děkuji vedoucímu prof. RNDr. Jiřímu Vondráškovi, CSc. za zajímavé téma na bakalářskou práci včetně všech cenných rad, které mi dal. Zároveň moc děkuji Mgr. Kateřině Balážové Faltejskové za velkou pomoc se psaním práce. Vážím si všech nově nabitých znalostí, které mi psaní bakalářské práce přineslo a děkuji moc celému okolí, které dodávalo nejen psychickou podporu.

Zároveň děkuji vývojářům všech software, které byli na bakalářkou práci využity, jedná se o programy:

- Python3 (včetně knihoven: obonet, networkx, sys, re, pygraphviz, pandas, requests, time, sys, gzip, tarfile, io, tempfile, os, BioPython, xml, CherryPy)
- LaTeX & LuaTeX & BibTeX & šablona better-mff-thesis
- Linux Ubuntu
- LibreOffice Calc
- ChatGPT
- Google
- Google scholar
- SCISPACE (typeset)
- sci-hub
- Microsoft Copilot
- ID mapping Uniprot
- BLAST
- FoldSeek
- ESMFold
- Graphviz

Název práce: Vizualizace funkcí podobných proteinů

Autor: Magdaléna Turinská

Katedra: Katedra fyzikální a makromolekulární chemie (2600)

Vedoucí bakalářské práce: prof. RNDr. Jiří Vondrášek, CSc., Ústav organické chemie a biochemie AV ČR

Abstrakt: Důležitou úlohou v molekulární biologii je hledání nových proteinů a stanovení jejich funkce. Přestože experimentální výzkum proteinů je stále nenahraditelný, výpočetní techniky umožnují získávat nové poznatky mnohem rychleji. Proto se čím dál častěji k predikcím 3D struktur a biologických funkcí využívají počítačové metody.

V této práci jsou diskutovány různé metody predikce funkčních anotací, včetně moderních metod využívajících hluboké strojové učení. Zároveň je představen nový program GOLizard, určený k vizualizaci funkcí podobných proteinů, které jsou získány pomocí programů BLAST a FoldSeek. K vizualizaci využívá hierarchické uspořádání Gene Ontology termínů pomocí orientovaného grafu, který zobrazuje vztahy mezi jednotlivými termíny.

Klíčová slova: funkční anotace proteinů, Gene Ontology, podobnost proteinů, vizualizace

Title: Visualization of the functions of similar proteins

Author: Magdaléna Turinská

Department: Department of Physical and Macromolecular Chemistry (2600)

Supervisor: prof. RNDr. Jiří Vondrášek, CSc., Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences

Abstract: An important task in molecular biology is the search for new proteins and the determination of their functions. Although experimental research on proteins remains irreplaceable, computational techniques allow us to obtain new insights much faster. Therefore, computer calculations are increasingly being used for predicting 3D structures and biological functions.

In this work, various methods for predicting functional annotations are discussed, including modern methods based on deep machine learning. Additionally, a new program called GOLizard is introduced, which can be used to visualize functions of similar proteins obtained through the BLAST and FoldSeek programs. It uses a hierarchical arrangement of Gene Ontology terms via a directed graph, showing the relationships between individual terms.

Keywords: functional annotation of proteins, Gene Ontology, protein similarity, visualization

Obsah

Seznam zkratek

A adenin - báze DNA. 4, 47
ATP Adenosintrifosfát. 8

BLAST Basic Local Alignment Search Tool. ii–iv, 5, 17, 18, 26, 29–32, 49–51, 55–57
BLOSUM Blocks Substitution Matrix. 10

C cytosin - báze DNA. 4, 47
CASP Critical Assessment of Structure Prediction. 22

DAG acyklický orientovaný graf. 14, 27
DNA Deoxyribonukleová kyselina. 3, 4, 8, 9, 32

EC Enzyme Commission Classification. 14

G guanin - báze DNA. 4, 47
Gag group specific antigen. 29, 30, 32
GO Gene Ontology. iii, iv, 5, 11–14, 25, 27–32, 34, 44, 49–60

HIV virus lidské imunitní nedostatečnosti. 28–32, 34

mRNA mediátorová ribonukleová kyselina. 4, 47

PAM Point Accepted Mutation. 10
PDB Protein Data Bank. 3, 19, 24
pol polymeráza. 29, 30, 32

RMSD střední kvadratická odchylka. 20, 21
RNA Ribonukleová kyselina. 3, 31

T thymin - báze DNA. 4
TC Transporter Classification. 14

U uracil - báze RNA. 47

wwPDB Worldwide PDB. 19, 24
k. *i* - *i*-tá kapitola v knize

Úvod

Klíčovým odvětvím molekulárně - biologického výzkumu je studium sekvence, struktury a funkce proteinů. Takto získané poznatky jsou cenné v aplikovaném farmaceutickém, medicínském i veterinárním výzkumu pro studium a léčbu nemocí. V rámci studia proteinů se využívají metody experimentální, teoretické a výpočetní.

Stěžejním vztahem molekulární biologie je takzvané centrální dogma molekulární biologie. To dává do souvislosti genetickou informaci uloženou v deoxyribonukleové kyselině (DNA) se sekvencí aminokyselin v proteinu a tok informací mezi molekulami.

DNA je složena ze čtyř základních bází adeninu (A), thyminu (T), guaninu (G) a cytosinu (C), které mohou být v téměř libovolném pořadí. Udávají sekvenci DNA, včetně její informační hodnoty. Základní jednotkou v DNA jsou geny, které typicky reprezentují právě jeden protein. Tato jednotka je celá přepsána do mediátorové ribonukleové kyseliny (mRNA), která ještě může být upravena (příkladem takové úpravy je sestřih mRNA)[1, k. 6].

Z takto připravené mRNA může být pomocí ribozomu syntetizován protein. Tato syntéza probíhá podle genetického kódu, který obsahuje 61 tripletových kodónů v mRNA pro 20 aminokyselin v budoucím proteinu a 3 tripletové kodóny značící konec syntézy aminokyselinového řetězce. Kodóny a jejich odpovídající aminokyseliny jsou uvedeny v příloze A [2]. V případě mutace, které budou rozebrány v kapitole 1.5, může v DNA vzniknou kodón pro jinou aminokyselinu a tím je způsobena změna v proteinu. Tato změna může razantně změnit jeho strukturu [3].

„Analýza struktury je zásadní pro pochopení funkce proteinů, protože se předpokládá, že jejich funkčnost je úzce spjata se strukturou“ [4]. Proto existují výzkumy zaměřující se na vztah mezi funkcí a strukturou [5]. Na poznatcích z těchto studií staví výpočetní bioinformatický obor zaměřující se na odhad funkce proteinu jak ze struktury, tak ze sekvence [6]. Různé metody předpovědi jsou diskutovány v této práci v kapitole 1.6.

Výzkum v oblasti přírodních věd produkuje data, která jsou archivována do různých databází. Databází existuje více z důvodu různých zaměření (například databáze sekvencí, nebo třeba proteinových struktur). Jednotlivé záznamy v nich jsou cenné, ale jejich přidaná hodnota spočívá v možnostech vzájemných porovnání.

Ty jsou prováděny na základě různých parametrů, přičemž stěžejním (biologicky relevantním) typem porovnání proteinů je na základě jejich sekvenční a strukturní podobnosti. Proto je v sekci 1.5 probíráno téma podobnosti proteinů a v kapitole 2 téma biologických databází, včetně možných metod vyhledávání v nich.

V poslední části (kapitole 3) této práce je představen program, který k dotazované sekvenci vyhledá podobné proteiny a následně ukáže jejich funkční anotace. Tento přístup dává uživateli možnost posouzení, která ze zobrazených funkční anotací je relevantní pro vkládanou sekvenci s ohledem na další jím známé skutečnosti (například podle původu vzorku proteinu). Program v rámci vyhledání uplatňuje metody BLAST (viz. kapitola 2.1) nebo FoldSeek dle výběru uživatele. Následně využívá projekci funkcí podobných proteinů do hierarchicky uspořádaných termínů v Gene Ontology (GO), a zobrazuje její relevantní podgraf.

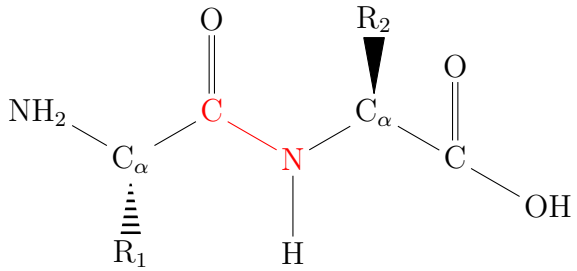
Kapitola 1

Proteiny: sekvence, struktura, funkce a homologie

1.1 Sekvence proteinů

Z biochemického hlediska jsou proteiny lineární polymery složené ze základních 20 aminokyselin [1, k. 2]. Jednotlivé aminokyseliny mají C_α atom, na kterém je navázána aminoskupina a karboxylová skupina, díky nimž jsou aminokyseliny spojeny peptidovou vazbou. Ta je zobrazena na obrázku 1.1 [1, k. 3]. Každá z těchto 20 aminokyselin má jiný postranní řetězec, který zajišťuje její specifické chemicko-fyzikální vlastnosti [7, k. 1]. Každé aminokyselině je přiděleno unikátní písmeno z abecedy. Takto je možné zakódovat sekvence proteinu jako posloupnost písmen. Tato písmena a k nim náležící postranní řetězce jsou uvedeny v příloze A. Typická délka proteinu je okolo 300 reziduí [1, k. 3], avšak existují proteiny složené z pouze desítek a nebo naopak z tisíců reziduí [8]. Největším známým proteinem v roce 2003 byl titin [9], který je v databázi Uniprot (v červnu 2024) čtvrtým nejdelším proteinem[1] [10]. Konkrétní délka tohoto proteinu (v dubnu 2024)

[1] Aktuálním nejdelším proteinem je peptidylprolyl isomeráza [10]



Obrázek 1.1 Peptidová vazba mezi dvěma aminokyselinami

u potkana je 35 375 aminokyselin [10].

Z teoretického hlediska může být libovolné ze zmíněných 20 písmen na každé pozici řetězce, což zajišťuje značnou variabilitu [1, k. 3]. Pro 20 různých reziduí na 300 pozicích proteinu vychází $\sum_{i=1}^{300} 20^i \approx 10^{390}$ možných kombinací. Z provedených statistických analýz plyne, že přirozené proteiny jsou podmnožinou celého kombinatorického prostoru všech možných proteinů [11]. Toto také naznačuje počet proteinů v proteinové databázi Uniprot. V této databázi se v dubnu 2024 nachází 248 805 733 různých záznamů [10], což je řádově méně než celková velikost kombinatorického prostoru. Předpokládá se, že tento velký rozdíl mezi teoretickou a reálnou velikostí vznikl i z důvodu evolučních fixací [12].

1.2 Struktura proteinů

Proteiny vykonávají velké množství funkcí, které jsou závislé na dosažení prostorového uspořádání aminokyselinového řetězce. Velká množina proteinů má svojí 3D strukturu v čase stabilní, těm říkáme proteiny strukturované [13]. Existují i další proteiny, které 3D-strukturu v čase stabilní nemají. Ty se nazývají vnitřně nestrukturované (intrinsically disordered proteins) [14, 15].

Strukturované proteiny zaujímají distinktní 3D strukturu, která je podle Anfinsenova dogmatu dána pořadím aminokyselin v sekvenci [16]. Složený protein vzniká díky interakcím mezi jednotlivými aminokyselinami a také díky interakcím s okolním prostředím. Tyto interakce mohou být vodíkové můstky, disulfidové můstky, van der Waalsovy síly, elektrostatické síly nebo také hydrofobní interakce [1, k. 3].

Strukturu proteinů dělíme do čtyř úrovní na primární (sekvence), sekundární (základní 3D strukturní elementy, kterými jsou hlavně α-helixy a β-listy), terciární a kvartérní. Sekundární elementy jsou udržovány vodíkovými můstky spojujícími C_α kostru proteinu. V α-helixu je C_α kostra stočena do pravotočivé šroubovice, v β-listu je C_α kostra uspořádaná ve více paralelních i antiparalelních řadách vedle sebe. Vodíkovými můstky jsou propojeny jednotlivá patra α-helixu, nebo řady β-listu [1, k. 3].

Terciární struktura popisuje prostorové uspořádání více sekundárních jednotek [7, k. 1]. Kvartérní popisuje uspořádání více aminokyselinových řetězců v jednom komplexu [1, k. 3].

Studium struktury proteinu dává informaci o biochemické funkci proteinu, jelikož funkce je závislá na interakčním místě proteinu, včetně konkretních interagujících aminokyselin [17].

1.3 Funkce proteinů

Proteiny mají různé funkce. Jednou z jejich důležitých funkcí je udržení struktury a tvaru buňky [1, k. 16]. Také zajišťují katalytické aktivity nebo například umožňují přenos malých molekul přes cytoplazmatickou membránu [1, k. 3].

Existuje velká skupina proteinů, které se nazývají strukturní a primárně slouží k udržení struktury a tvaru biologické hmoty (uvnitř buňky i mezibuněčné hmoty). Mezi tyto proteiny patří například keratin, který tvoří vlasy a nehty. Dalším zástupcem této skupiny je kolagen, nacházející se v extracelulárním matrixu [18]. Důležitým strukturním proteinem je elastin [19], který má elastické vlastnosti a proto se nachází v pojivových tkáních [18].

Katalytické aktivity se účastní enzymy, které dále kategorizujeme do tříd. Příkladem takové třídy jsou hydrolázy katalyzující hydrolytické štěpení [1, k. 3], mezi které patří protein NAD 5'- nukleotidáza [20]. Další třídou jsou oxido-reduktázy katalyzující přenos elektronu mezi molekulami [1, k. 3]. Příkladem oxido-reduktázy je biliverdin reduktáza [21].

Přenos molekul přes cytoplasmatickou membránu pomocí proteinů má více možností. Jednou z možností jsou kanálové a transportérové proteiny, které zajišťují přenos pomocí usnadněné difuze (to je přenos poháněný samovolně z míst s vysokou koncentrací do míst s nízkou koncentrací, tedy podle koncentračního gradientu) [22, k. 2]. Mezi tyto proteiny patří aquaporiny zajištující přenos molekul vody, příkladem je aquaporin-8 [23]. Další možností je přenos molekul aktivním transportem pomocí hydrolýzy adenosintrifosfátu (ATP). Takový způsob využívá Ca^{2+} pumpa [22, k. 2]. Mezi takové proteiny patří membránová ATPáza 1 transportující vápník [24].

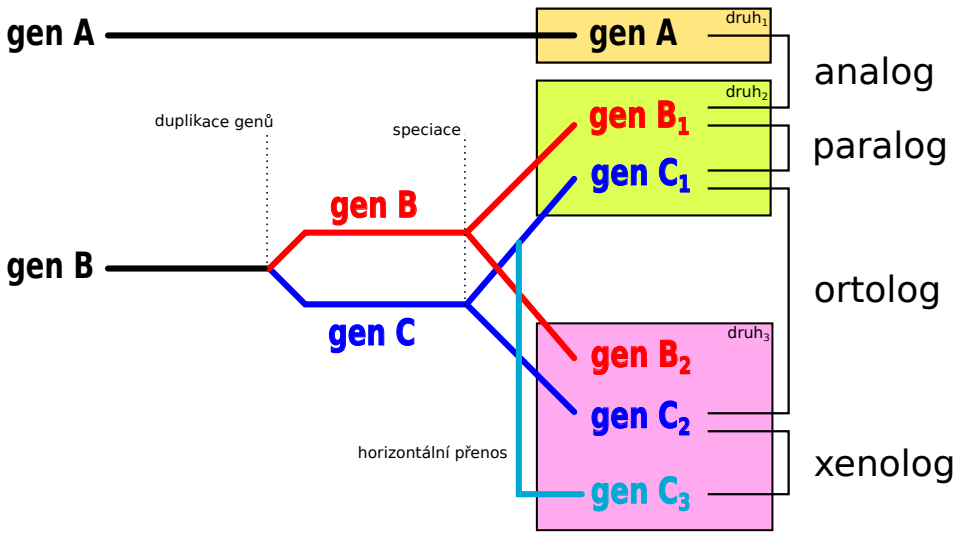
1.4 Homologie

Ačkoliv existuje mnoho různých definic homologie, tak obvyklou definicí je podobnost způsobená dědičností [25, 26]. Zásadní otázkou výzkumu homologie je způsob odlišení homologie od analogie, což je podobnost nezpůsobená dědičností [25].

Dědičnost je standardně spojena s geny a DNA, ale také je popsán vztah mezi DNA a proteiny (viz úvod této práce). Z toho je zjevné, že geny mají podobnou informační hodnotu jako proteiny v kontextu evoluční podobnosti, a proto jsou v následujícím textu tyto pojmy zaměňovány.

Homologii dělíme například podle úrovně srovnání na iterativní homologii zabývající se jedním jedincem v určitém čase, ontogenetickou homologii zabývající se jedním jedincem v různých časech, polymorfní homologií zabývající se více jedinci stejného druhu a supraspecifickou homologii zabývající se skupinou jedinců s podobnými znaky [26].

Dalším možným dělením homologie je podle typu evolučního vztahu mezi dvěma



Obrázek 1.2 Dělení homologie podle typu evolučního vztahu mezi dvěma geny.

geny. Jak je ukázáno na obrázku 1.2, tyto evoluční vztahy se dělí na paralogii definující vztah dvou genů vzniklých genovou duplikací, ortologii definující vztah dvou genů vzniklých speciací druhu, xenologii definující vztah dvou genů vzniklých horizontálním přenosem a případně analogii popisující neexistenci evolučního vztahu mezi dvěma geny, jejich produkty vykonávají podobnou funkci [27].

1.5 Podobnost proteinů v kontextu homologie

Jedním z biologicky relevantních srovnání proteinů je zarovnání (alignment) dvou sekvencí. Tento způsob je založen na představě, že proteiny podléhají mutacím v průběhu evoluce. Mutace v proteinech rozdělujeme na záměny (substituce), vložení (inzerce) a odstranění (delece) aminokyselin z řetězce [28]. Posloupností více mutací v DNA může vznikat odlišný protein.

Vhodným biologickým parametrem při srovnání dvou proteinových sekvencí je evoluční vzdálenost, tedy počet evolučních změn, které se odehrály mezi danými homologními proteiny (případně nekonečno počtu evolučních změn (∞) pro proteiny analogní). Při určování konkrétní hodnoty evoluční vzdálenosti nastává problém zjištění a následného započítání více evolučních změn na jedné pozici. Jako aproximace evoluční vzdálenosti se používá editační vzdálenost [29], která byla původně zavedena ke zpracování přirozeného jazyka [30].

Definice 1 (Editační vzdálenost [30]). *Máme-li dva řetězce a a b nad abecedou Σ, pak editační vzdálenost $d(a, b)$ je definována jako minimální počet operací potřebných ke změně řetězce a na b. Povolené operace jsou vložení znaku do řetězce, odebrání znaku ze řetězce a záměna znaku za jiný znak.*

Protože spočítaná editační vzdálenost nezohledňuje více změn na jedné pozici, tak je typicky nižší než skutečná evoluční vzdálenost, proto existují různé zpřesňující korekce a upravené definice [29]. Běžně užívanou zpřesňující úpravou je zavedení takzvaných skórovacích matic. Skórovací matice je tabulka 20×20, která přiřazuje každé substituci dvou aminokyselin jinou cenu, závislou na pozorované četnosti záměn těchto aminokyselin ve všech proteinech [31]. Tradičně využívané jsou dva typy těchto matic, jsou to Blocks Substitution Matrix (BLOSUM) a Point Accepted Mutation (PAM) [32]. Další možností úpravy editační vzdálenosti je zavedení lineární ceny za vložení nebo odebrání znaku ze řetězce. Tento model zavádí jednu cenu za nově přidanou mezeru a jinou cenu za více mezer vedle sebe [33].

„Za evolučně příbuzné, tedy homologní proteiny, považujeme takové sekvence, které vykazují významně vyšší úroveň podobnosti, než by se očekávalo u náhodných sekvencí [34]. Tato podobnost naznačuje, že proteiny mají společného předka a během evoluce prošly diverzifikací, která však nevymazala jejich základní společné rysy.“ [35].

Sekvenční identita $> 30\%$ mezi dvěma sekvencemi je statisticky významná a tedy můžeme dané proteiny považovat za homologní [34]. Existují ovšem i proteiny, které mají sekvenční identitu mezi 20% a 30% [36] a vykonávají podobnou funkci, což naznačuje jejich evoluční příbuznost [37].

1.6 Anotace funkce proteinů

Ideální anotace proteinu obsahuje informace o jeho funkci (případně více funkcích). Ta je primárně zajištěna experimentálním výzkumem. Tento postup je výrazně pomalejší než rychlost vkládání sekvenčních a strukturních dat do databází [38]. Kvůli tomu by vznikalo mnoho záznamů v databázích bez uvedené funkční anotace. Proto bylo vymyšleno mnoho metod pro počítačovou predikci funkčních anotací. Příklady těchto metod k roku 2021 jsou uvedeny v tabulkách 1.1 a 1.2. V roce 2022 byla vyvinuta metoda PANDA2 [39]. Metody predikce funkční anotace často zahrnují přenesení funkční anotace z podobného proteinu (případně agregace informací z více podobných proteinů) [40]. Další metodou předpovězení funkce proteinu je například na základě interakčních dat [41]. Příkladem takového přístupu je „metoda přiřazování funkcí založená na pravděpodobnostní analýze sousedství v interakční síti[2] protein-protein“ [42]. Neposlední možností predikce proteinových funkcí je využití metod strojového učení [43]. Moderní metody zahrnují kombinaci více metod, příkladem je DeepGOPlus, který kombinuje podobnost sekvencí s vyhledávání sekvenčních motivů pomocí neuronových sítí [44].

Postup přenesení funkce předpokládá, že podobnost proteinu nevznikla čistou náhodou. Jednou možností vzniku podobnosti dvou proteinů je na základě podobného evolučního původu. Takové proteiny se snažíme odhalit pomocí sekvenčního

[2]interakční síť je reprezentována informatickým grafem

Metoda	Název	Publikováno	Autoři	Dostupné
Na základě podobnosti	OntoBlast	2003	Zehetner [45]	-
	GOFigure	2003	Khan et al. [46]	-
	GOblet	2003	Hennig, Groth a Lehrach [47]	-
	Gotcha	2004	Martin, Berriman a Barton [48]	-
	PFP	2008	Hawkins et al. [49]	-
	INGA	2015	Piovesan et al. [50]	-
	GoFDR	2016	Gong, Ning a Tian [51]	-
Pravděpodobnostní	–	2003	Letovsky a Kasif [42]	-
	–	2007	Nariai, Kolaczyk a Kasif [52]	-
	BMRF		Kourmpetis et al. [53]	-
Strojové učení	GOPET	2006	Vinayagam et al. [54]	-
	PoGO	2010	Jung et al. [55]	-
	FFPred3[a]	2016	Cozzetto et al. [56]	+
	PANNZER2	2018	Törönen, Medlar a Holm [57]	+
	DeepText2GO	2018	You, Huang a Zhu [58]	-
	NetGo	2018	You et al. [59]	+

Tabulka 1.1 Metody založené na tradičních přístupech pro přiřazování termínů GO k proteinům. Webové odkazy, na kterých jsou tyto programy k dispozici (dostupným v červenci 2024) jsou v tabulce 1.3.
Převzato z Vu a Jung [60], upraveno podle aktuálně (k 17.7.2024) dostupných informací a přeloženo.

[a]Feature-based function prediction

Využitá vlastnost	Název	Pub-liko-váno	Autoři	Dostupné
Na základě podobnosti	–	2014	Chicco, Sadowski a Baldi [61]	-
	ProLanGO	2017	Cao et al. [62]	+
	SECLAF[a]	2018	Szalkai a Grolmusz [63]	+
	DEEPred	2019	Sureyya Rifaioglu et al. [64]	+
	DeepSeq	2019	Nauman et al. [65]	+
	DeepGO-Plus	2019	Kulmanov a Hoehndorf [44]	+
Integrovaná data/založená na struktuře	–	2016	Tavanaei et al. [66]	-
	DeepGO	2018	Kulmanov, Khan a Hoehndorf [67]	+
	deepNF[b]	2018	Gligorijević, Barot a Bonneau [68]	+
	–	2018	Fa et al. [69]	+
	DeepFunc	2019	Zhang et al. [70]	+
	DeepGOA	2020	Zhang et al. [71]	+
	SDN2GO	2020	Cai, Wang a Deng [72]	+
	DeepAdd	2020	Du et al. [73]	-
	FFPred-GAN[c]	2020	Wan a Jones [74]	+
	GONET	2020	Li et al. [75]	+

Tabulka 1.2 Metody založené na hlubokém učení predikující funkční anotaci přiřazením GO termínů. Webové odkazy, na kterých jsou tyto programy k dispozici (dostupným v červenci 2024) jsou v tabulce 1.3.
Převzato z Vu a Jung [60], upraveno podle aktuálně (k 17.7.2024) dostupných informací a přeloženo.

[a]Sequence Classification Framework
[b]deep Network Fusion
[c]FFPred - Generative Adversarial Networks

Název (identifikace)	Dostupné na
FFPred3	bioinf.cs.ucl.ac.uk/psipred/
PANNZER2	ekhidna2.biocenter.helsinki.fi/sanspanz/
NetGo	dmiip.sjtu.edu.cn/ng3.0
ProLanGO	github.com/caorenzhi/ProLanGO2
SECLAF	pitgroup.org/seclaf/
DEEPred	github.com/cansyl/DEEPred
DeepSeq	github.com/recluze/deepseq
DeepGO-Plus	deepgo.cbrc.kaust.edu.sa/deepgo/
DeepGO	github.com/bio-ontology-research-group/deepgo
deepNF	github.com/VGligorijevic/deepNF
Fa et al. [69]	bioinf.cs.ucl.ac.uk/downloads/mtdnn/
DeepFunc	csuligroup.com/DeepPFP/others
DeepGOA	csuligroup.com/DeepPFP/others
SDN2GO	github.com/Charrick/SDN2GO
FFPred-GAN	github.com/psipred/FFPredGAN
GONET	github.com/wanglx1874/GONET
PANDA2	dna.cs.miami.edu/PANDA2/

Tabulka 1.3 Seznam online dostupných programů pro predikci funkční anotace proteinů pomocí přiřazení GO termínů. Tento seznam zahrnuje webové servery a stažitelné verze programů, které jsou dostupné k 15. červenci 2024.

zarovnání.

Dalším možnou příčinou podobnosti je výskyt ve stejném okolním prostředí. Díky podobnosti okolních molekul, například substrátů, se daným proteinům mohlo vyvinout podobné interakční místo s velmi podobnou strukturou a fyzikálními vlastnostmi. Proto se pro nalezení podobného proteinu může využít i strukturní zarovnání [76].

Funkční anotace proteinu může být detailní popis definující funkci v konkrétních podmínkách (například daných konkrétním experimentem). Bohužel takto podrobný popis není možné automaticky třídit do podobných funkčních kategorií a tedy není možné říci zda dané dva proteiny sdílejí nějakou vlastnost. Proto vznikly přesně definované soubory popisných termínů, které jsou neredundantní a strojově čitelné [76]. U enzymů vznikla enzymová nomenklatura, ve které je definovaná klasifikace pomocí EC čísel (Enzyme Commission Classification), která jsou složena ze čtyř čísel reprezentujících hierarchické uspořádání [77]. Pro membránové transportní proteiny byla zavedena databáze klasifikace transportérů, která definuje TC čísla (Transporter Classification). Tato klasifikace je organizovaná jako pěti úrovňový hierarchický systém [78].

Výše uvedené klasifikační systémy jsou hodně omezené, tím, že klasifikují pouze podskupinu proteinů. Nezávisle na těchto systémech vznikla množina biologicky relevantních termínů s jasně definovanou hierarchickou strukturou, jedná se o GO [79, k. 9]. V této hierarchické struktuře je v dubnu 2024 uloženo 42 271 termínů [80]. GO je reprezentována jako acyklický orientovaný graf (DAG)[3]. Obsahuje tři hlavní vrcholy, ze kterých nevede žádná hrana. Tyto vrcholy reprezentují tři oblasti, ve kterých jsou uspořádány všechny ostatní termíny. Jedná se o oblasti **molekulární funkce** (GO:0003674), **buněčná komponenta** (GO:0005575) a **biologický proces** (GO:0008150). Hrany DAGu uvnitř oblastí reprezentují jednotlivé vztahy mezi dvěma termíny. Existuje několik druhů těchto vztahů, které jsou v GO podporovány. Příklady vztahů jsou *A je podtypem B, A je vždy část B, A vždy obsahuje B, A reguluje B* [80, 81, 79].

Existuje mnoho predikčních metod, které predikují funkci proteinu a mapují ji na GO. Například na základě podobnostních dat, nebo také pravděpodobnostních dat [60]. Nelze však snadno najít metodu přímého odvození funkce proteinu ze sekvence, která nejprve predikuje strukturu proteinu, následně využívá strukturní podobnost a zobrazuje odpovídající podgraf GO [60]. Vizualizovaný podgraf GO je uživatelsky lépe interpretovatelný než seznam funkcí. Navíc může uživateli zobrazit více vrcholů ve shluku, což uživateli umožňuje vizuální posouzení relevantnosti navržené funkce. Proto je v rámci této práci vyvinuta metoda, která tento postup implementuje a tím nabízí pro uživatele alternativní pohled. Tato metoda bude představena v kapitole 3.

[3] directed acyclic graph

Kapitola 2

Vyhledávání podobných proteinů

2.1 Sekvenční zarovnání

Standardním přístupem zarovnání sekvence délky m se sekvencí délky n je algoritmus navržený Needlemanem a Wunschem v roce 1970. Tento algoritmus byl navržen v kubickém výpočetním čase, protože byla umožněna penalizace mezer jako funkce ve velikosti mezer [82]. Moderní implementace této metody běží v čase $O(m \cdot n)$ [83, 84]. To znamená, že počet operací provedených za dobu výpočtu zarovnání dvou sekvencí je maximálně $m \cdot n \cdot k$, kde k je libovolná konstanta.

Představme si, že máme sekvenci a délky m, která činí 300 aminokyselin. K této sekvenci hledáme podobné sekvence pomocí výše uvedeného algoritmu. Prohledáváme databázi D neredundantních sekvencí, která v dubnu 2024 obsahovala 722 278 205 proteinových sekvencí[1] [85]. Pro jednoduchost výpočtu (2.1) definujme, že délka každého proteinu v databázi je 300 aminokyselin, což je obvyklá délka proteinu uvedená v kapitole 1.1.

$$|D| \cdot (m \cdot n \cdot k) = k \cdot 722\ 278\ 205 \cdot 300 \cdot 300 = k \cdot 65005038450000 \approx k \cdot 10^{13} \quad (2.1)$$

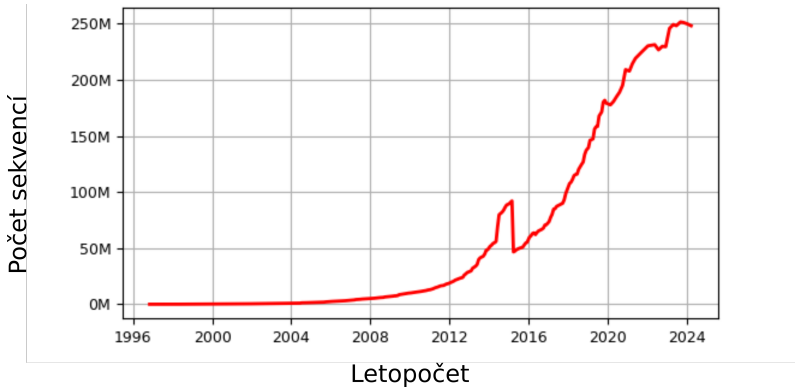
Běžný počítač (v roce 2022) vykoná 10^9 operací za sekundu [86, k. 2]. Za těchto podmínek algoritmus v nejhorším případě nalezne podobné sekvence za

$$k \cdot \frac{10^{13}}{10^9} = k \cdot 1000\text{sekund} \approx k \cdot 0.3\text{hodiny}. \quad (2.2)$$

Při představě, že „zanedbaná“ konstanta k je pouhých 5, je doba běhu výše uvedeného algoritmu 1.5 hodiny.

Výpočet (2.1) je závislý na parametrů $|D|$, což je velikost dotazované databáze. Ty budou detailně probírány v kapitole 2.3. Je uvedena hodnota $\approx 10^9$ k dubnu 2024,

[1] Využívám databázi *nr* (All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects)



Obrázek 2.1 Graf znázorňující zvětšování databáze UniProt. Pík zobrazený v roce 2015 je vzniklý rozsáhlým mazáním redundantních proteinů [87].
Obrázek převzat z aktuálních statistik UniProtu (v dubnu 2024) [10] a upraven přidáním popisků os.

ale proteiny jsou do databází stále přidávány, proto by bylo potřeba v brzké době započítat číslo vyšší, čímž by se zvýšili výsledky výše uvedených výpočtů(2.1)(2.2). Rychlost přírůstku databáze zavisí na mnoha faktorech, jako příklad rychlosti růstu databáze je uveden graf přírůstku UniProtu 2.1. Pro konkrétní představu je zde uvedeno, že v lednu 2023 bylo v databázi UniProtKB/TrEMBL 245 871 724 [10] sekvencí a v lednu 2024 zde bylo 249 751 891 [10] sekvencí, tedy o 3 880 167 více.

Pokud by se přesné algoritmy používaly na porovnání sekvence s každou sekvencí v databázi, tak by pro jednu sekvenci byli nalezeny podobné sekvence v řádu desítek minut. Navíc není časté, že by předmětem výzkumu bylo vyhledání podobných proteinů pro jednu sekvenci. Typicky předmětem výzkumu bývají komparativní analýzy, ve kterých jsou vyhledávány podobné proteiny pro více vstupních proteinů.

V případě potřeby vyhledání podobných sekvencí pro pouhých 100 vstupních sekvencí (při zanedbání k) bude čas výpočtu (2.2) 30hodin, což znamená více jak jeden den. Ovšem už při 1000 vstupních sekvencí je doba běhu algoritmu 300hodin, tedy více jak týden čistého času.

2.1.1 Heuristické metody

Na vyhledávání podobných sekvencí v databázích se využívají metody založené na heuristikách. Heuristiky jsou typicky rychlejší a prostorově úspornější algoritmy, což kompenzují přesností výsledku. Negarantují optimální řešení [88]. To znamená, že může existovat proteinová sekvence, která je dotazované sekvenci podobnější než všechny nalezené heuristickým algoritmem. Příklady programů určených k vyhledání podobných sekvencí, které využívají heuristické metody jsou uvedeny v tabulce 2.1.

2.1.2 Kvantifikace shody

Jak bylo zmíněno v kapitole 1, tak se předpokládá, že nenáhodná podobnost sekvencí souvisí s homologii daných proteinů [34]. Proto je k objevení homologií využíváno mnoho různých programů, které v databázích vyhledávají podobné proteiny. Vybrané (běžně užívané) programy jsou uvedeny v tabulce 2.1. V případě vyhledání podobných proteinů v databázi je důležitá informace, jak moc je daný nález relevantní.

Existuje hrubé skóre s, které je závislé na množství substitucí, inzercí a delecí mezi dotazovaným a vyhledaným proteinem. Tato hodnota reflektuje vzdálenost těchto proteinů. Statistickou normalizací této hodnoty získáme bitové skóre s', které se typicky využívá k vyjádření míry relevance [96].

Bohužel databázové vyhledání může nalézt některé proteiny náhodou. Proto se používají statistické modely k odhadu, zda lze očekávat, že nalezené skóre bylo získáno náhodou. Tuto náhodu je vhodné zohlednit ve všech nalezených shodách jako parametr nejistoty homologie.

Definice 2 (E-hodnota [34])**.** *Nechť s' je skóre a D velikost databáze, pak rovnicí*

$$E(s') \leq p(s' \geq x) \leq 1 - \exp(-\exp(-x)) \times D$$

je definována E-hodnota závislá na s' skóre.

Název	Publikováno	Autoři	Dostupné
SSEARCH	1981	Smith, Waterman et al. [89]	+
FASTA[a]	1988	Pearson a Lipman [90]	+
BLAST[b]	1990	Altschul et al. [85]	+
PSI-BLAST[c]	1997	Altschul et al. [91]	+
HMMER3[d]	1998	Durbin et al. [92]	+
USEARCH[e]	2010	Edgar [93]	+
OSWALD[f]	2016	Rucci et al. [94]	+
MMseqs2[g]	2017	Steinegger a Söding [95]	+

Tabulka 2.1 Příklad programů vyhledávající podobné sekvence v databázi.
Webové odkazy, na kterých jsou tyto programy k dispozici v červenci 2024 jsou v tabulce 2.2.

[a] Fast alignment
[b] Basic Local Alignment Search Tool
[c] Position-Specific Iterated BLAST
[d] Biosequence analysis using profile hidden Markov models
[e] Ultra-fast sequence analysis tool
[f] OpenCL Smith-Waterman Algorithm on Altera FPGA for Large Protein Databases
[g] Many-against-Many sequence searching

Název (identifikace)	Dostupné na
SSEARCH	npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_ssearch.html
FASTA	www.genome.jp/tools/fasta/
BLAST	blast.ncbi.nlm.nih.gov/Blast.cgi
PSI-BLAST	blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&PROGRAM=blastp&BLAST_PROGRAMS=psiBlast
HMMER3	hmmer.org/
USEARCH	www.drive5.com/usearch/
OSWALD	github.com/enzorucci/OSWALD
MMseqs2	seqsearch.east.k8s.rcsb.org/search

Tabulka 2.2 Seznam online dostupných programů pro vyhledávání podobných sekvencí v databázích. Tento seznam zahrnuje webové servery a stažitelné verze programů, které jsou dostupné k 15. červenci 2024.

Přesně E-hodnota pro konkrétní shodu udává počet očekávaných náhodných zarovnání se stejným nebo lepším skóre[2]. Hodnotu s podobnou informací uvádí i níže diskutovaný FoldSeek [97].

2.2 Strukturní zarovnání

Druhou standardní variantou zarovnání proteinů je na základě proteinové struktury. Tento přístup zobrazuje souvislosti mezi proteiny, u kterých podobnost vznikla reakcemi na stejné okolní prostředí, jak bylo diskutováno v kapitole 1.6.

„Proteinovou strukturu nelze popsat pouhou posloupností písmen, jako je tomu u sekvence aminokyselin. Místo toho je nutné specifikovat prostorové uspořádání proteinu, které je klíčové pro jeho funkci. Toto prostorové uspořádání je standardně zapisováno pomocí koordinát jednotlivých atomů, které udávají jejich přesnou polohu v trojrozměrném prostoru. Tímto způsobem lze detailně popsat, jak jsou jednotlivé části proteinu vůči sobě navzájem orientovány a jak vytvářejí složitou trojrozměrnou strukturu“[98].

Pro popis 3D struktur proteinů byly zavedeny různé automaticky zpracovatelné formáty. Základním typem souboru je soubor s příponou *.pdb*. Jedná se o typ

[2]Jedná se o vysvětlení převzaté z www.nlm.nih.gov/ncbi/workshops/2023-08_BLAST_evol/e_value.html (7.8.2024)

souboru zavedený před rokem 1977 organizací Worldwide PDB (wwPDB) [99]. Jsou v něm obsaženy informace o souřadnicích jednotlivých atomů a sekundárních strukturách. Tento soubor má přesně definovanou strukturu, a proto bohužel není možné do něj uložit molekulu delší než 99 999 atomů. Z toho důvodu je v dnešní době již zastaralý a tudíž by se neměl používat [100]. I přes to je stále velmi využívaným typem formátu. Příklad proteinu uloženém v tomto formátu je uveden na následujícím odkazu: `www.ebi.ac.uk/pdbe/entry-files/pdb1aoi.ent`[3].

Průměrný počet atomů v aminokyselině je 19.2 [4]. Pokud může v souboru typu *.pdb* být uloženo maximálně 99 999 atomů, tak jde uložit maximálně 5 200[5] průměrných aminokyselin v řetězci. To představuje omezení, které může být problematické při práci s většími proteinovými komplexy nebo strukturami obsahujícími více domén. Z tohoto důvodu vznikl novější formát mmCIF/PDBx [100], který nemá pevně daný rozsah a tím lze do něj uložit libovolně velké proteinové struktury [101].

2.2.1 Metody strukturního vyhledávání

V této části jsou diskutovány výhody jednotlivých metod využívaných k vyhledání podobných proteinových struktur. Příklady metod používaných ve vědecké komunitě[6] určených k vyhledávání ve strukturních databázích jsou uvedeny v tabulce 2.3.

Programy na vyhledávání strukturní podobnosti v databázích byly dlouhodobě pomalé. Toto se změnilo v roce 2021, když byl uveden do provozu FoldSeek [97]. Jedním z prvních algoritmů byl v roce 1995 Dali [102]. Tento algoritmus využíval informace o atomových souřadnicích. Metody vyvinuté později snižovali výpočetní čas heuristickými přístupy. Příkladem takových metod je CE a mTM-align [104, 106]. Mezi moderní využívané metody se řadí programy TM-search a FoldSeek. TM-search je specifický tím, že vyhledává v předpřipravených databázích, které uspořádal hierarchicky podle matice podobnosti. Poté porovnává jen se zástupci proteinů v dané hierarchické kategorii, a tím značně omezuje jím prohledávaný prostor proteinových struktur [107]. FoldSeek je moderní metoda vyvinutá za účelem velmi rychlého prohledávání databází. Rychlost výpočtu je zajištěna mnoha algoritmickými prvky. Prvním stěžejním bodem je reprezentace aminokyselinové kostry proteinu pomocí speciálně vyvinuté strukturální abecedy. Ta popisuje prostorovou orientaci (včetně terciálních interakcí) mezi dvěma nejbližšími strukturními motivy. Dalším urychlením výpočtu je využití filtrování

[3]Odkaz funkční 28. dubna 2024

[4]Průměr vypočten na odkaze: `www.imgt.org/IMGTeducation/Aide-memoire/_UK/aminoacids/abbreviation.html` (7.8.2024)

[5]$99\ 999/19.2 \approx 5200$

[6]počty citací uvedené na Google Scholar: Dali: 1619, VAST: 1392, CE: 2611, FATCAT: 730, mTM-align: 88, TM-Search: 3, FoldSeek: 435 (stav ke dni 19.6.2024)

dat před samotným algoritmem, které je navrženo s ohledem na snížení počtu falešně pozitivních výsledků. V neposlední řadě využívá FoldSeek paralelního běhu programu. Stěžejním prvkem výpočtu zajištujícím časově efektivní a velmi přesné vyhledávání je speciálně navržená a natrénovaná neuronová sít. Rychlost FoldSeeku při prohledávání databáze AlphaFoldDB (verze 1) je 184 600× rychlejší než původní program Dali [97].

Zajímavé je, že již program FATCAT z roku 2003 přiřazuje různým podčástem proteinu různé váhy vyjadřující konzervovanost části proteinu. Tato informace je cenná, jelikož zarovnání rozvolněných míst má minimální informační hodnotu z důvodu častých konformačních změn [105]. Konzervovanost struktur také zohlednili vývojáři FoldSeek při trénovaní neuronové sítě [97].

2.2.2 Kvantifikace shody

Při porovnávání struktur je stěžejní získat informaci o míře jejich podobnosti. Obvyklým měřítkem podobnosti je střední kvadratická odchylka (RMSD[7]). Ta vyjadřuje vzdálenost vztaženou na jednu aminokyselinu mezi dvěma optimálně zarovnanými sekvencemi.

Definice 3 (RMSD [108])**.** *„Hodnota RMSD mezi dvěma molekulami A a B složenými z n atomů je definována s ohledem na ortogonální matici R (rotace)*

[7]Root mean square deviation

Název	Publikováno	Autoři	Dostupné
Dali[a]	1995	Holm a Sander [102]	+
VAST[b]	1996	Gibrat, Madej a Bryant [103]	+
CE[c]	1998	Shindyalov a Bourne [104]	-
FATCAT[d]	2003	Ye a Godzik [105]	+
mTM-align	2018	Dong et al. [106]	-
TM-search	2022	Liu et al. [107]	+
FoldSeek[e]	2023	Van Kempen et al. [97]	+

Tabulka 2.3 Příklad programů vyhledávající podobné struktury v databázi. Webové odkazy, na kterých jsou tyto programy k dispozici v červenci 2024 jsou v tabulce 2.4.

[a]Protein structure comparison server
[b]Vector Alignment Search Tool
[c]Combinatorial extension
[d]Flexible structure AlignmenT by Chaining Aligned fragment pairs allowing Twists
[e]Fast and accurate protein structure alignment and visualisation

Název (identifikace)	Dostupné na
Dali	`ekhidna2.biocenter.helsinki.fi/dali/`
VAST	`structure.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html`
FATCAT	`fatcat.godziklab.org/fatcat/fatcat_search.html`
TM-search	`zhanggroup.org/TM-search/`
FoldSeek	`search.foldseek.com/search`

Tabulka 2.4 Seznam online dostupných programů pro vyhledávání podobných struktur v databázích. Tento seznam zahrnuje webové servery a stažitelné verze programů, které jsou dostupné k 17. červenci 2024.

a translační vektor v:“[8]

$$RMSD := \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|a_i - Rb_i - v\|^2}. \tag{2.3}$$

Existuje mnoho dalších odvozených metrik, příkladem je relativní RMSD zohledňující rozdílnou délku mezi proteiny. Tato metrika nabývá hodnot mezi 0 a 1, přičemž 1 vyjadřuje shodné proteiny a 0 velmi rozdílné proteiny [109].

Další metodou vhodnou na kvantifikaci podobnosti dvou struktur je TM-skóre[9]. Jedná se o jediné zde představené skóre, které bylo vyvinuto přímo na porovnání proteinových struktur. Nabývá hodnot mezi $(0, 1]$ (kde vyšší skóre odpovídá vyšší shodě) a dobře koreluje s lidským vizuálním vnímáním podobnosti [110].

Definice 4 (TM-score [110])**.** *Nechť L_n je délka nativní struktury, L_T je počet zarovnaných zbytků, d_i je vzdálenost mezi i-tým párem zarovnaných zbytků, d_0 normalizační faktor pak*

$$TM\text{-}score := \max \frac{1}{L_n}\sum_{i=1}^{L_T} \frac{1}{1 + (\frac{d_i}{d_0})^2}$$

Toto skóre zohledňuje jak délku proteinu, tak kvalitu zarovnání v celé délce proteinu a tedy je určeno k posouzeni globálního zarovnání [110].

2.2.3 Strukturní zarovnání na základě sekvence

Jak bylo uvedeno výše v dnešní době již existují rychlé a spolehlivé programy vhodné k vyhledávání ve strukturních databázích. V kapitole 1.6 byl diskutován

[8] Jedná se o přesně převzatou a přeloženou definici (i s rovnicí) ze článku (Fukutani et al. [108]).

[9] název vychází z anglického *template modeling score*

rozdíl mezi sekvenční a strukturní podobností. Jelikož sekvenční podobnost je založená na homologii a strukturní na funkci, tak se nabízí otázka zda lze ze sekvence odvodit i strukturní podobnost[10]. Přímé odvození strukturní podobnosti ze sekvence možné není, ale existují moderní metody predikce proteinové struktury ze sekvence. Tímto způsobem lze získat přibližnou strukturu proteinu, kterou lze následně využít na vyhledání strukturně podobných proteinů v databázi.

Existuje více metod predikce struktury proteinu, ale moderní metody využívají převážně hluboké strojové učení. Příklad takových metod je uveden v tabulce 2.5.

Program AlphaFold vyvinutý společností DeepMind byl první relativně přesnou metodou založenou na hlubokém strojovém učení. Ve 14. kole soutěže Critical Assessment of Structure Prediction (CASP) dokázal určit strukturu na atomární přesnost i v případech kdy v databázi nebyl nalezen homologní protein [111]. Od té doby vznikají další metody dosahujících podobných výsledků, ty jsou uvedeny v tabulce 2.5. AlfaFold byl primárně testován na vstupech obsahujících více zarovnaných sekvencí. Oproti tomu ESMFold se zaměřuje na predikci struktury z jediné sekvence. ESMFold předpoví strukturu proteinu v kratším čase než AlphaFold za cenu nižší přesnosti (postranních řetězců) [112].

Využitá vlastnost	Název	Publikováno	Autoři	Dostupné
Více-násobné zarovnání sekvencí	AlphaFold	2021	Jumper et al. [111]	+
	RoseTTAFold	2021	Baek et al. [113]	+
Předem připravené modely proteinových jazyků	ESMFold[a]	2021	Rives et al. [112]	+
	OmegaFold	2022	Wu et al. [114]	+
více pohledového kontrastní učení	S-PLM[b]	2024	Wang et al. [115]	+

Tabulka 2.5 Příklad programů predikující 3D strukturu proteinu pomocí hlubokého strojového učení. Webové odkazy, na kterých jsou tyto programy k dispozici v červenci 2024 jsou v tabulce 2.6.
Seznam metod inspirován článkem od Jeliazkov, Alamo a Karpiak [116].

[a] Evolutionary Scale Modeling
[b] Structure-aware Protein Language Model

[10] Nabízí se i opačná otázka zda lze využít strukturu k sekvenčnímu vyhledávání, ale zde je důležité si uvědomit, že strukturní informace je robustnější a obsahuje uvnitř sebe informaci o sekvenci. Tedy tato metoda lze využít, ale za cenu ztráty informace.

Název (identifikace)	Dostupné na
AlphaFold	`colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb`
RoseTTAFold	`github.com/RosettaCommons/RoseTTAFold`
ESMFold	`esmatlas.com/resources?action=fold`
OmegaFold	`github.com/HeliXonProtein/OmegaFold`
S-PLM	`github.com/duolinwang/S-PLM`

Tabulka 2.6 Seznam online dostupných programů predikující 3D strukturu proteinu. Tento seznam zahrnuje webové servery a stažitelné verze programů, které jsou dostupné k 17. červenci 2024.

2.3 Prohledávané databáze

Při vyhledávání podobných proteinů k dotazovanému je stěžejním bodem výběr vhodné prohledávané databáze. Mohlo by se zdát, že čím více záznamů v databázi je, tím více podobných proteinů lze nalézt. Toto není pravda, protože u každého záznamu musíme filtrovat náhodné zarovnání (bylo probráno v sekci 2.1.2), pravděpodobnější je u velkých databází s hodně záznamy [34]. Proto je pro jisté výzkumy vhodné zvolit vyhledávání pouze v nějaké filtrované podmnožině zvolené databáze.

Dalším důležitým parametrem při výběru vhodné databáze může být informace o relevanci v ní uložených dat. Existují databáze, které obsahují záznamy, které nejsou podloženy experimentálním výzkumem, ale jsou jen predikovány. Příkladem takové proteinové strukturní databáze je AlfafoldDB[117].

Informaci o relevanci nám může dát i množství dat zkontrolovaných expertem. Zkontrolované databáze (curated databases) jsou výrazně kvalitnější a organizovanější než pouze automaticky udržované databáze, ale za cenu velmi pomalého růstu [118]. Příkladem plně zkontrolované sekvenční databáze je Swiss-Prot. K tomu v roce 1996 vznikla druhá část TREMBL, obsahující ručně nekontrolovaná data [119]. V dnešní době jsou obě tyto množiny dat k dispozici v databázi Uniprot [10].

Příklady metod vyvinutých k prohledávání databází jsou uvedeny v tabulkách 2.1 (metody sekvenčního vyhledávání) a 2.3 (strukturního).

Většina metod uvedených v tabulkách 2.1 a 2.3 je také k dispozici ke stažení. V případě využívání stažené varianty programu je standardně možné zvolit si pro uživatele vhodnou sadu prohledávaných sekvencí s ohledem na téma výzkumu. Druhou možností je využití webových serverů poskytnutých typicky vývojářem daného software. Tato možnost je komfortnější pro rychlé a snadné užití programu,

ale velmi omezující ve výběru možných prohledávaných databázích, protože si musíme vystačit se sadami dat poskytnutými daným serverem. Pro sekvenční zarovnání na webových serverech je u SSEARCH, FASTA, BLAST a PSI-BLAST k dispozici velká neredundatní databáze a Swiss-Prot. Webové servery SSEARCH, FASTA, BLAST a PSI-BLAST nabízejí i další databáze, ale ty jsou pro každý uvedený program různé.

Webové servery strukturních vyhledávacích programů Dali, VAST, FATCAT, TM-search a FoldSeek vždy nabízí vyhledávání v Protein Data Bank (PDB), která je spravovaná organizací wwPDB [100]. Některé z uvedených programů nabízí i další databáze pro vyhledávání podobných struktur. Program FoldSeek nabízí sedm různých databází, a navíc i všechny jejich kombinace.

Kapitola 3

Vizualizace funkce podobných proteinů

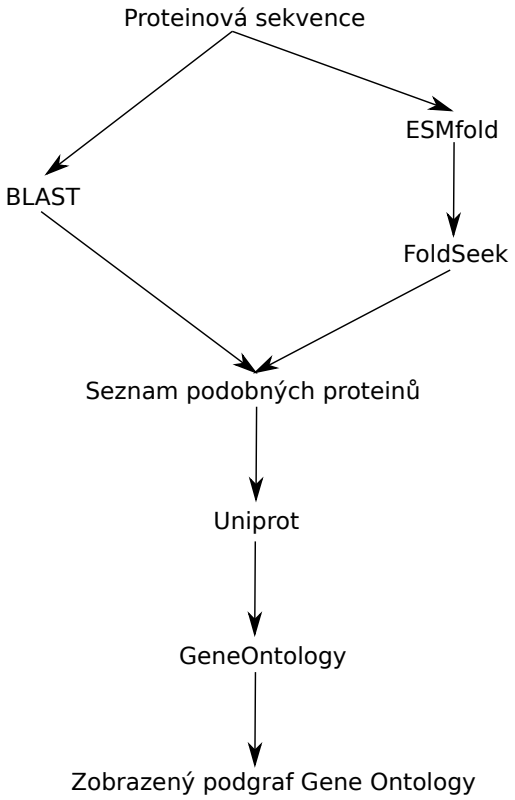
V této práci je provedena analýza různých metod přenesení funkční anotace (viz kapitola 1.6.) a ní byl nenalezen nástroj zaměřující se na vizualizaci funkcí podobných proteinů pomocí grafu GO. Proto je tato metoda v této práci představena.

Představovaný program se jmenuje GOlizard. Originalitou navrženého softwaru je zobrazení všech funkcí podobných proteinů. Dle kapitoly 1 lze předpokládat, že sekvenčně podobné proteiny reflektují evoluci a díky ní sdílejí stejnou funkci. Strukturně podobné proteiny sdílejí stejné funkce z důvodu reakce na podobné okolní podmínky. Proto je předložen výčet funkčních anotací podobných proteinů s předpokladem relevance přenesení těchto anotací na vstupní protein. Tento přístup přináší uživateli více informací a zároveň mu přenechává rozhodnutí výběru věrohodné charakteristiky proteinu, jak bylo zmíněno v úvodu. Výstupem programu je orientovaný graf s projekcí všech navržených anotací GO (viz kapitola 1.6). Podstatným záměrem je, že vizualizovaný podgraf není zobrazen ve vrstvách reprezentujících vzdálenost od hlavních vrcholů[1]. „Tato varianta je vybrána, protože vzdálenost není relevantním měřítkem kvality odhadu funkce. Přiřazení GO anotace k proteinu totiž odráží hloubku vědeckého poznání v konkrétní tématické oblasti. Zohledňuje také další informace, které jsou o daném proteinu známy v době anotace“ [120, 121].

Na obrázku 3.1 je zobrazeno schéma jednotlivých kroků výpočtu. Program na vstupu bere proteinovou sekvenci ve FASTA formátu. To je standardizovaný formát, který na prvním řádku obsahuje informace o proteinu a na dalších řádcích samotnou sekvenci proteinu zapsanou v přiřazených písmenech (viz kapitola 1.)[2] [122]. Příklad proteinu, který je ve FASTA formátu je uveden na obrázku 3.2. Jedná se

[1] viz kapitola 1.6: molekulární funkce, buněčná komponenta a biologický proces

[2] Neoficiální specifikace na `zhanggroup.org/FASTA/` dostupná 7.8.2024



Obrázek 3.1 Schéma software pro vizualizaci predikce funkce proteinu

```
>sp|P01308|INS_HUMAN Insulin OS=Homo sapiens OX=9606 GN=INS PE=1 SV=1
MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRR
EAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
```

Obrázek 3.2 Příklad proteinu uloženého ve FASTA formátu

o lidský inzulín (UniprotID: P01308) stažený v červnu 2024 z databáze Uniprot.

Následně program vyhledá proteiny podobné vstupní sekvenci. Tuto část program provádí dle nastavení uživatelem jednou ze dvou různých metod. Buď využije sekvenční vyhledávání pomocí programu BLAST, nebo z vložené sekvence predikuje strukturu pomocí programu ESMFold, a poté využije strukturní vyhledání pomocí programu FoldSeek. Programy FoldSeek a ESMFold jsou vybrány z důvodu krátkého výpočetního času a dostupnosti webových serverů. Bohužel tato varianta je z důvodu omezení predikce struktury v ESMFold omezena maximální délkou vstupní sekvence na 400 aminokyselin.

Jelikož vyhledání nalezne i proteiny, které jsou si navzájem velmi podobné (a tedy mohou nést podobnou informační hodnotu o funkci). Je také naimplementována

Barva	Název vztahu
Černá	A je podtypem B
Tmavě modrá	A je vždy část B
Žlutá	A reguluje B
Zelená	A pozitivně reguluje B
Červená	A negativně reguluje B
Růžová	A vždy obsahuje B
Tyrkysová	A se vyskytuje v B

Tabulka 3.1 Barvy vztahů GO ve výstupních grafech programu GoLizard. Zvolené barvy odpovídají barvám v aplikaci QuickGO.

možnost využití hierarchického klastrování. Tato možnost znamená, že proteiny se sekvenční identitou vyšší než nastavená prahová hodnota jsou nahrazeny jedním reprezentantem, tím se stává nejdelší ze seskupených proteinů. Ve výsledných proteinech, které jsou využity pro další zpracovaní, se tedy nebudou nacházet proteiny podobnější než nastavená prahová hodnota. Důsledkem této konfigurace je zahrnutí širšího spektra GO anotací, které se nacházejí u méně příbuzných proteinů.

Program následně projde nalezené podobné proteiny a ke každému z nich extrahuje z databáze Uniprot všechny uvedené GO termíny. Takto získá množinu možných funkčních anotací reprezentovanou vrcholy DAGu. Dalším krokem programu je detailní vyčištění přebytečných vrcholů v grafu GO. Přebytečnými vrcholy jsou označeny takové, které neleží na žádné cestě mezi nějakou navrženou funkční anotací vstupní sekvence a vrcholy reprezentujícími některý z oblastí molekulární funkce (GO:0003674), buněčná komponenta (GO:0005575) a biologický proces (GO:0008150) (viz kapitola 1.6).

Takto upravený graf GO následně uživateli zobrazí. V zobrazeném grafu jsou hrany odlišeny různými barvami, dle typu vztahů uvnitř GO. Barvy a jim odpovídající vztahy jsou ukázány v tabulce 3.1. Dalším barevným odlišením je obarvení horního nadpisu každého termínu GO, u kterého barva reprezentuje oblast, ve které se tento termín primárně nachází. Většina barev je konsistentní s grafy dostupnými v aplikaci QuickGO [123].

Navíc je možno vygenerovat textový, strojově zpracovatelný, výstup ve formě tabulky obsahující všechny navržené funkční anotace GO a k nim odpovídající proteiny z Uniprotu. Program také umožňuje vytvoření exportu grafu do formátu OBO. Jedná se o textový soubor s předem definovanou strukturou určený k ukládání ontologií [79, 124]. Konkrétně obsahuje hlavičku s podrobnostmi o verzi a stáří daného souboru. Pak následuje část obsahující informace o jednotlivých GO termínech.

```
[Term]
id: GO:0008135
name: translation factor activity, RNA binding
namespace: molecular_function
def: "Functions during translation
by binding to RNA during polypeptide synthesis
at the ribosome." [GOC:ai, GOC:vw]
subset: goslim_chembl
subset: goslim_plant
subset: goslim_yeast
synonym: "translation factor activity,
nucleic acid binding" BROAD [GOC:mah]
is_a: GO:0003676 ! nucleic acid binding
is_a: GO:0045182 ! translation regulator activity
relationship: has_part GO:0003723 ! RNA binding
relationship: part_of GO:0006412 ! translation
```

Obrázek 3.3 Příklad jednoho GO termínu v OBO file formátu.
Jedná se o term GO:0008135, z Plant subset - staženo 24.6.2024.

V nich jsou specifikovány konkrétní vztahy s ostatními termíny. Také je zde rozepsán detailní textový popis co tato GO anotace vyjadřuje. Na obrázku 3.3 je vidět příklad uložení jednoho GO termínu ve formátu OBO. V poslední části souboru jsou definovány vztahy, které jsou mezi uloženými termíny [124].

3.1 Příklad vizualizace

Pro ilustraci výpočtu bylo provedeno několik spuštění. Vstupní sekvence byla vybrána tak, aby dosud nebyla publikována. Taková sekvence byla zvolena z důvodu nemožnosti nalezení stejného proteinu v žádné veřejné databázi[3]. Výsledky tohoto testu jsou uvedeny v příloze B.

3.1.1 Technické informace o vstupních datech

Jako vstup programu byla využita ze 38% mutovaná sekvence vycházející z HIV-1 proteázy (UniprotID: O90777)[4], bohužel konkrétní vstupní sekvence zde není uvedena z důvodu pokračujícího výzkumu na Ústavu organické chemie a biochemie AV ČR, v.v.i. ve skupině Jiřího Vondráška. Jako podkladový graf GO byl zvolen `go-basic.obo` stažený 13.3.2024 na webovém odkaze `geneontology.org/docs/download-ontology/`. Po každém vyhledání

[3] v červenci 2024

[4] virus lidské imunitní nedostatečnosti (HIV), anglicky: Human Immunodeficiency Virus

a zklastrování podobných proteinů bylo vybráno 10 nejlepších shod dle procenta identických pozic. K těmto 10 proteinů byly nalezeny odpovídající GO anotace, ke kterým byla udělaná grafová analýza a následné zobrazení.

Vyhledávání podobných sekvencí pomocí programu BLAST bylo provedeno nad databází Swiss-Prot. Podkladové databáze pro strukturní vyhledávání (FoldSeek) byly zvoleny AlphaFold/UniProt50 v4, AlphaFold/Swiss-Prot v4 a AlphaFold/Proteome v4[5].

U základní HIV-1 proteázy (UniprotID: O90777) jsou v červenci 2024 v aplikaci QuickGO uvedeny dvě GO anotace. Jedná se o aktivitu endopeptidázy asparagového typu (GO:0004190) z oblasti molekulární funkce (GO:0003674)a proteolýzu (GO:0006508) z oblasti biologického procesu (GO:0008150).

3.1.2 Popis výstupních dat

V příloze B jsou k prohlédnutí výstupy ve všech třech oblastech GO využívající BLAST bez klastrování, BLAST s klastrováním, které shlukuje proteiny se sekvenční identitou alespoň 90%, FoldSeek bez klastrování a FoldSeek s klastrováním, které shlukuje proteiny se sekvenční identitou alespoň 90%. Celkem je tedy zobrazeno 12 podgrafů GO.

Posouzení dle způsobu vyhledání podobných proteinů

Při vyhledávání podobných proteinů programem FoldSeek bez klastrování byl mezi 10 nejlepšími výsledky nalezen protein s UniprotID A0A0K0EQP9, který byl v databázi[6] Uniprot smazán. Kvůli tomu k němu nejsou přiřazeny žádné GO termíny. V návaznosti jednoho smazaného UniprotID se v zobrazených grafech B.4, B.5 a B.6 nalézá pouze devět UniprotID. Ve variantě s FoldSeekem s klastrováním byly mezi 10 nejlepšími shodami nalezeny dva smazené UniprotID (A0A2H2Y5B9 a A0A351EBB9) a navíc bylo nalezeno pět proteinů, u kterých zatím není uvedena funkční anotace. Jsou to UniprotID: A0A5D3BRE1, A0A0C9T4L2, B4W551, A0A818XM81 a A0A7C4F495. Z těchto důvodů se bohužel v grafech B.10, B.11 a B.12 nalézají pouze tři UniprotID. Pravděpodobným důsledkem této komplikace je výrazně menší velikost (a nižší informační hodnota) těchto grafů, oproti všem ostatním v příloze B.

Ve variantách se sekvenčním vyhledáváním pomocí programu BLAST měly všechny nalezené proteiny přiřazené GO anotace. Při spuštění výpočtu bez klastrování program nalezl 10 „Gag-Pol polyprotein“[7] u různých podtypů „viru lidské imunitní

[5]Všechny databáze ve verzi 7/2024.

[6]verze 7/2024

[7]vysvětlení zkratek: group specific antigen (Gag), polymeráza (pol) [125]

nedostatečnosti (HIV) typu 1 skupiny M“[8]. Jedná se o proteiny: O93215, P04588, Q9QBZ5, Q9QBZ1, Q75002, Q9QSR3, O89290, Q9QBZ9, Q9QBY3, Q9WC63 [10]. Jelikož se jedná o stejné proteiny u velmi podobných organizmů, tak měli i přiřazené stejné GO anotace a tedy z pohledu grafové analýzy GO i stejnou informační hodnotu. Ve výsledcích B.1, B.2 a B.3 je tato vlastnost zřetelně vidět. Na základě této skutečnosti byla doimplementována výše představená varianta s klastrováním, které shlukuje proteiny s vyšší (než prahovou hodnotou) sekvenční identitou. Tato úprava odhalila přínosné informace, jelikož se ve výsledných 10 proteinech, vzešlých z BLAST s klastrování, nalézají také pouze „Gag-Pol polyproteiny“, avšak tři z nich u „opičího viru imunitní nedostatečnosti (šimpanzího viru imunitní nedostatečnosti)“[9]. Jedná se o proteiny: Q9WC54, O93215, Q1A249, Q1A267, O89290, Q9QBY3, P17283, O12158, O41798, Q77373 [10]. Efekt se očividně projevuje při porovnání grafů z oblasti biologický proces, kde graf s klastrováním (ve kterém jsou proteiny s alespoň 90% sekvenční identitou zklastrovány B.9) je nadmnožinou grafu bez klastrování B.3. V oblastech molekulární funkce a buněčná komponenta jsou totožné GO anotace v grafech s klastrování a bez klastrování, ale s různou distibucí UniprotID. Ty jsou k nahlédnutí na grafech B.7 a B.8.

Posouzení dle podstaty GO anotace

Následující řádky popisují výsledné grafy z hlediska oblastí GO. Jsou představeny souvislosti mezi grafy s různým vyhledáním podobných proteinů. V textu u sdílených vlastností jsou uváděny pouze nejkonkrétnější společné GO anotace, jelikož všechny termíny hierarchicky obecnější se již také musí překrývat. To z důvodu způsobu čištění grafu, protože program při čištění nechává všechny cesty k obecným termínům. Naopak v případě navržení jiných částí grafu je, kvůli přehlednosti uváděn nejobecnější rozdílný termín.

[8] Původní anglická verze: Human immunodeficiency virus type 1 group M

[9] Původní anglická verze: Simian immunodeficiency virus (Chimpanzee immunodeficiency virus)

Způsob analýzy	BLAST	BLAST & FoldSeek	FoldSeek
Bez klastrování	vazba lipidů (GO:0008289)	aktivita endopeptidázy asparagového typu (GO:0004190)	podoblast vazba na proteiny (GO:0005515)
	vazba RNA pomocí vlásenky (GO:0035613)		
	podoblast vazba na proteiny (GO:0005515)		
S klastrováním	ostatní (modře neoznačené) termíny	aktivita endopeptidázy asparagového typu (GO:0004190)	

Tabulka 3.2 Přehled dále diskutovaných GO termínů, které jsou uvedeny v grafech v oblasti molekulární funkce (GO:0003874).
V druhém sloupci jsou GO termíny, které obsahuje pouze graf vzniklý BLASTovou variantou programu. Ve třetím sloupci jsou termíny, které sdílejí grafy s odlišným způsobem vyhledání podobných proteinů. V posledním sloupci se nacházejí termíny unikátní pro FoldSeekoví graf.

První probíranou oblastí výsledkových grafů je molekulární funkce (GO:0003874). Přehled rozdílných a společných termínů mezi jednotlivými grafy je sepsán v tabulce 3.2. Zajímavým poznatkem v této oblasti je, že grafy bez klastrování (BLAST B.1, FoldSeek B.4) jsou si hodně podobné, což je naprostým opakem k výsledkům z oblasti buněčné komponenty (GO:0005575). Konkretně 37 termínů GO z oblasti molekulární funkce (GO:0003874) je identických v obou grafech, 7 jich je pouze v BLASTové variantě a 5 jich je pouze ve FoldSeekové variantě. U vyhledávání pomocí BLAST se jedná o funkce vazba lipidů (GO:0008289), vazba RNA pomocí vlásenky (GO:0035613) a termíny z podoblasti exonukleázové aktivity (GO:0004527). U strukturního vyhledávání pomocí FoldSeeku byla zobrazena podoblast vazba na proteiny (GO:0005515). Navíc jediný termín na kterém se shodnou všechny grafy z oblasti molekulární funkce (GO:0003874) (B.1, B.4, B.7 a B.10) je aktivita endopeptidázy asparagového typu (GO:0004190). Pro přehlednost je tato anotace zvýrazněna modrou barvou, včetně celé cesty k termínu molekulární funkce (GO:0003874). Zároveň se jedná o anotaci uvedenou u původní HIV-1 proteázy (UniprotID: O90777) (viz sekce 3.1.1).

V oblasti biologického procesu (GO:0008150) se všechny grafy (B.3, B.6, B.9 a B.12) shodují pouze v proteolýze (GO:0006508). Jedná se o pozitivní zjištění, jelikož proteolýza (GO:0006508) je uvedena jako GO anotace původní HIV-1 proteázy (UniprotID: O90777) (viz sekce 3.1.1). Druhý velký překryv je na metabolickém procesu DNA (GO:0006259). Zde se shodují obě varianty BLAST i FoldSeek bez klastrování. Třetím společným termínem je virový proces (GO:0016032), který FoldSeek bez klastrování ukazuje jako „nejkonkrétnější“ anotaci, naproti tomu obě BLASTové verze tuto podoblast detailně zobrazují včetně 10ti prvkového grafu předchůdců (konkrétnějších termínů).

Dále se v grafech týkajících se biologického procesu projevují jednotlivá specifika různých vyhledávání. Například varianta Foldseeku bez klastrování (obrázek B.6) nalezla pozoruhodnou informaci. Jedná se o dva termíny, které lze zobecnit na proteolýzu (GO:0006508). Jsou to zrání proteinu (GO:0016485) a katabolický proces zprostředkovaný proteiny v proteazomu (GO:0010498). Dále tato varianta nabídla úplně nové části GO grafu. Jedná se o tématické okruhy reakce na podmět (GO:0050896) a regulace biologické kvality (GO:0065008). Překvapivým výsledkem je, že FoldSeek bez klastrování také přinesl nové informace a to podoblasti reprodukce (GO:0000003) a vývojového procesu (GO:0032502). Poslední zde rozebranou částí je podoblast zabíjení buněk (GO:0001906), kterou navrhuje jediný BLAST s klastrováním B.9. Může se jednat o náhodný (nerelevatní) nález. Přesto je tato informace zajímavá, jelikož v roce 2003 bylo u HIV-1 proteázy experimentálně zjištěno, že způsobuje smrt buňce, která byla virem HIV napadena [126].

V oblasti týkající se buněčné komponenty (GO:0005575) je na grafech B.2, B.5, B.8 a B.11 přítomná membrána (GO:0016020). Vysvětlení membrány jako anotace HIV-1 proteázy může vycházet z poznatku, že celý virion HIV získává svůj lipidový obal na plazmatické membráně hostitelské buňky [127]. Jedná se primárně o anotaci celého Gag-pol polyproteinu.

Při porovnání grafů bez klastrování z oblasti buněčné komponenty (GO:0005575), které vznikly BLASTem (B.2) a FoldSeekem (B.5) je vidět značný rozdíl v zobrazených GO anotacích. To naznačuje podstatný rozdíl mezi sekvenční a strukturní podobností, jak bylo diskutováno v kapitole 1.6. Graf využívající strukturní vyhledávání jediný nabízí oblast komplexu makromolekul s proteinovou složkou (GO:0032991). Oba diskutované grafy (B.2, B.5) obsahují ústřední termín buněčný anatomický celek (GO:0110165), ale struktura jejich mnoha předchůdců (konkrétnějších termínů) je rozdílná.

Program představený v této práci úspěšně identifikoval obě GO anotace, které jsou uvedeny u nemutované HIV-1 proteázy (UniprotID: O90777).

3.2 Budoucí vývoj

Programová část této práce je v červenci 2024 stále ve vývoji. Je předpokládáno zveřejnění zdrojových kódů finální verze na githubu, včetně podrobněji sepsaného souboru `README.md`, který bude přeložen do anglického jazyka. Zároveň bude dán do provozu webový server, na kterém bude program GOLizard nasazen. K dispozici bude jako webový formulář vhodný pro uživatelsky jednoduché vyhledání anotací podobných proteinů.

Závěr

V této práci byl popsán vztah mezi proteinovou sekvencí, strukturou a homologií. Následně byly ukázány různé metody hledání podobných proteinů na základě sekvenční a strukturní podobnosti. Byly představeny jednotlivé programy určené k vyhledávání těchto proteinů. Byla uvedena možnost, při které je z neznámé sekvence predikována struktura a následně jsou vyhledány strukturně podobné proteiny. Dále byly prezentovány různé programy pro predikci struktury proteinu. Práce obsahovala vysvětlení v dnešní době používaných systémů pro uložení funkční anotace proteinů. Dále byly představeny různé programy, které anotace predikují.

Stěžejním prvkem práce bylo představení programu GOLizard. Jedná se o program, který vizualizuje funkce proteinů, které jsou podobné neanotované vstupní sekvenci. K vizualizaci program využívá hierarchický anotační systém GO. Ten je reprezentován orientovaným grafem. Program nejprve vyhledá podobné proteiny, poté extrahuje jejich funkční anotace a nakonec zobrazí odpovídající podgraf GO. Použití programu bylo ilustrováno na nezveřejněné sekvenci, kterou se stala ze 38% mutovaná HIV-1 proteáza.

Navržený program může být užitečný pro odhad vlastností neznámých proteinů. Lze předpokládat i využití u metagenomických a jiných podobných sběrů dat. Z důvodu anotovaní nepopsaných proteinů je důraz kladen na navržení širokého rozsahu potenciálních anotací a přenechání interpretace na uživateli software.

Seznam použitých zdrojů včetně literatury

[1] Bruce Alberts. *Molecular biology of the cell*. eng. Sixth edition. Boca Raton, FL: CRC Press, an imprint of Garland Science, 2017. ISBN: 1-315-73536-9.

[2] Francis HC Crick. "The origin of the genetic code". In: *Journal of molecular biology* **38**.*3* (1968), s. 367–379.

[3] Yoshiaki Urakubo, Teikichi Ikura a Nobutoshi Ito. "Crystal structural analysis of protein–protein interactions drastically destabilized by a single mutation". In: *Protein Science* **17**.*6* (2008), s. 1055–1065.

[4] OpenAI. *Odpověď na dotaz: souvislost mezi strukturou a funkcí proteinu*. Poskytnuto službou ChatGPT od OpenAI. PROMPT: „Zkus tuhle větu úplně přepsat (předtím je DNA → protein): Studium struktury je podstatné při výzkumu proteinové funkce, jelikož se předpokládá, že funkce proteinů je závislá na jeho struktuře.", [příspěvek vytvořený umělou inteligencí]. 2024. URL: `https://www.openai.com/chatgpt` (cit. 18. 07. 2024).

[5] Hedi Hegyi a Mark Gerstein. "The relationship between protein structure and function: a comprehensive survey with application to the yeast genome". In: *Journal of molecular biology* **288**.*1* (1999), s. 147–164.

[6] James C. Whisstock a Arthur M. Lesk. "Prediction of protein function from protein sequence and structure". In: *Quarterly Reviews of Biophysics* **36**.*3* (2003), s. 307–340. DOI: `S0033583503003901`.

[7] Anna Tramontano. *Protein structure prediction: concepts and applications*. eng. John Wiley & Sons, 2006. ISBN: 978-3-527-31167-5.

[8] Jianzhi Zhang. "Protein-length distributions for the three domains of life". In: *Trends in Genetics* **16**.*3* (2000), s. 107–109.

[9] Christiane A Opitz et al. "Damped elastic recoil of the titin spring in myofibrils of human myocardium". In: *Proceedings of the National Academy of Sciences* **100**.*22* (2003), s. 12688–12693.

[10] Alex Bateman et al. “UniProt: the Universal Protein Knowledgebase in 2023”. In: *NUCLEIC ACIDS RESEARCH* **51**.*D1* (2023), s. D523–D531. ISSN: 0305-1048. DOI: gkac1052.

[11] Aditya Mittal, Anandkumar Madhavjibhai Changani a Sakshi Taparia. “What limits the primary sequence space of natural proteins?” In: *Journal of Biomolecular Structure and Dynamics* **38**.*15* (2020), s. 4579–4583.

[12] Vyacheslav Tretyachenko et al. “Random protein sequences can form defined secondary structures and are well-tolerated in vivo”. In: *Scientific Reports* **7**.*1* (2017), s. 15449.

[13] IN Serdyuk. “Structured proteins and proteins with intrinsic disorder”. In: *Molecular Biology* **41** (2007), s. 262–277.

[14] Antonio Deiana et al. “Intrinsically disordered proteins and structured proteins with intrinsically disordered regions have different functional roles in the cell”. In: *PloS one* **14**.*8* (2019). DOI: journal.pone.0217889.

[15] A Keith Dunker et al. “Function and structure of inherently disordered proteins”. In: *Current opinion in structural biology* **18**.*6* (2008), s. 756–764.

[16] Jorge A Vila. “Metamorphic proteins in light of Anfinsen’s dogma”. In: *The Journal of Physical Chemistry Letters* **11**.*13* (2020), s. 4998–4999.

[17] Oliver C Redfern, Benoit Dessailly a Christine A Orengo. “Exploring the structure and function paradigm”. In: *Current opinion in structural biology* **18**.*3* (2008), s. 394–402.

[18] Keiji Numata. “How to define and study structural proteins as biopolymer materials”. In: *Polymer Journal* **52**.*9* (2020), s. 1043–1056.

[19] Christian EH Schmelzer, Melkamu Getie a Reinhard HH Neubert. “Mass spectrometric characterization of human skin elastin peptides produced by proteolytic digestion with pepsin and thermitase”. In: *Journal of Chromatography A* **1083**.*1-2* (2005), s. 120–126.

[20] Silvia Garavaglia et al. “The high-resolution crystal structure of periplasmic Haemophilus influenzae NAD nucleotidase reveals a novel enzymatic function of human CD73 related to NAD metabolism”. In: *Biochemical Journal* **441**.*1* (2012), s. 131–141.

[21] Tokio Yamaguchi, Yasuo Komoda a Hiroshi Nakajima. “Biliverdin-IX alpha reductase and biliverdin-IX beta reductase from human liver. Purification and characterization.” In: *Journal of Biological Chemistry* **269**.*39* (1994), s. 24343–24348.

[22] Geoffrey Cooper a Kenneth Adams. *The cell: a molecular approach*. Oxford University Press, 2022. ISBN: 9780197583722.

[23] Kun Liu et al. “Purification and functional characterization of aquaporin-8”. In: *Biology of the Cell* **98**.*3* (2006), s. 153–161.

[24] Meer Jacob Rahimi et al. “De novo variants in ATP2B1 lead to neurodevelopmental delay”. In: *The American Journal of Human Genetics* **109**.*5* (2022), s. 944–952.

[25] Paula M Mabee et al. “A logical model of homology for comparative biology”. In: *Systematic biology* **69**.*2* (2020), s. 345–362.

[26] Gerhard Haszprunar. “The types of homology and their significance for evolutionary biology and phylogenetics”. In: *Journal of evolutionary Biology* **5**.*1* (1992), s. 13–24.

[27] Helga Ochoterena et al. “The search for common origin: homology revisited”. In: *Systematic Biology* **68**.*5* (2019), s. 767–780.

[28] Prathima Iengar. “An analysis of substitution, deletion and insertion mutations in cancer genes”. In: *Nucleic acids research* **40**.*14* (2012), s. 6401–6413.

[29] Krister M Swenson et al. “Approximating the true evolutionary distance between two genomes”. In: *Journal of Experimental Algorithmics (JEA)* **12** (2008), s. 1–17.

[30] Daniel Jurafsky a James H Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. PEARSON INDIA, 2014. ISBN: 9789332518414.

[31] JL Risler et al. “Amino acid substitutions in structurally related proteins a pattern recognition approach: Determination of a new and efficient scoring matrix”. In: *Journal of molecular biology* **204**.*4* (1988), s. 1019–1029.

[32] Rakesh Trivedi a Hampapathalu Adimurthy Nagarajaram. “Substitution scoring matrices for proteins-An overview”. In: *Protein Science* **29**.*11* (2020), s. 2150–2163.

[33] Kun-Mao Chao. “Calign: aligning sequences with restricted affine gap penalties.” In: *Bioinformatics (Oxford, England)* **15**.*4* (1999), s. 298–304.

[34] William R Pearson. “An introduction to sequence similarity (“homology”) searching”. In: *Current protocols in bioinformatics* **42**.*1* (2013), s. 3–1.

[35] OpenAI. *Odpověď na otázku o homologních proteinech*. Poskytnuto službou ChatGPT od OpenAI. PROMPT: „víc rozepsat: Za evolučně příbuzné, tedy homologní proteiny považujeme takové sekvence, které mají větší podobnost, než by odpovídalo náhodným sekvencím [34].“, [příspěvek vytvořený umělou inteligencí]. 2024. URL: `https://www.openai.com/` (cit. 18. 06. 2024).

[36] Burkhard Rost. "Twilight zone of protein sequence alignments". In: *Protein engineering* **12**.*2* (1999), s. 85–94.

[37] S Balaji a N Srinivasan. "Comparison of sequence-based and structure-based phylogenetic trees of homologous proteins: Inferences on protein evolution". In: *Journal of biosciences* **32** (2007), s. 83–96.

[38] Valérie de Crécy-Lagard et al. "A roadmap for the functional annotation of protein families: a community perspective". In: *Database* **2022** (2022).

[39] Chenguang Zhao, Tong Liu a Zheng Wang. "PANDA2: protein function prediction using graph neural networks". In: *NAR genomics and bioinformatics* **4**.*1* (2022). DOI: lqac004.

[40] Daisuke Kihara, ed. *Protein Function Prediction: Methods and Protocols*. Methods in Molecular Biology. Citace vygenerována pomocí chatGPT [128]. New York: Humana Press, 2017. ISBN: 978-1-4939-7013-1. DOI: 978-1-4939-7015-5.

[41] Minghua Deng et al. "Prediction of protein function using protein-protein interaction data". In: *IEEE* (2002), s. 197–206. DOI: CSB.2002.1039342.

[42] Stanley Letovsky a Simon Kasif. "Predicting protein function from protein/protein interaction data: a probabilistic approach". In: *Bioinformatics* **19**.*suppl_1* (2003), s. i197–i204.

[43] Rosalin Bonetta a Gianluca Valentino. "Machine learning techniques for protein function prediction". In: *Proteins: Structure, Function, and Bioinformatics* **88**.*3* (2020), s. 397–413. DOI: prot.25832.

[44] Maxat Kulmanov a Robert Hoehndorf. "DeepGOPlus: improved protein function prediction from sequence". In: *Bioinformatics* **36**.*2* (čvc. 2019), s. 422–429. ISSN: 1367-4803. DOI: btz595.

[45] Gunther Zehetner. "OntoBlast function: From sequence similarities directly to potential functional annotations by ontology terms". In: *Nucleic acids research* **31**.*13* (2003), s. 3799–3803.

[46] Salim Khan et al. "GoFigure: Automated gene ontology™ annotation". In: *Bioinformatics* **19**.*18* (2003), s. 2484–2485.

[47] Steffen Hennig, Detlef Groth a Hans Lehrach. "Automated Gene Ontology annotation for anonymous sequence data". In: *Nucleic Acids Research* **31**.*13* (2003), s. 3712–3715.

[48] David MA Martin, Matthew Berriman a Geoffrey J Barton. "GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes". In: *BMC bioinformatics* **5** (2004), s. 1–17.

[49] Troy Hawkins et al. “PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data”. In: *Proteins: Structure, Function, and Bioinformatics* **74**.*3* (2009), s. 566–582.

[50] Damiano Piovesan et al. “INGA: protein function prediction combining interaction networks, domain assignments and sequence similarity”. In: *Nucleic acids research* **43**.*W1* (2015), W134–W140.

[51] Qingtian Gong, Wei Ning a Weidong Tian. “GoFDR: a sequence alignment based method for predicting protein functions”. In: *Methods* **93** (2016), s. 3–14.

[52] Naoki Nariai, Eric D Kolaczyk a Simon Kasif. “Probabilistic protein function prediction from heterogeneous genome-wide data”. In: *Plos one* **2**.*3* (2007), e337.

[53] Yiannis AI Kourmpetis et al. “Bayesian Markov Random Field analysis for protein function prediction based on network data”. In: *PloS one* **5**.*2* (2010), e9293.

[54] Arunachalam Vinayagam et al. “GOPET: a tool for automated predictions of Gene Ontology terms”. In: *BMC bioinformatics* **7** (2006), s. 1–7.

[55] Jaehee Jung et al. “PoGO: Prediction of Gene Ontology terms for fungal proteins”. In: *BMC bioinformatics* **11** (2010), s. 1–9.

[56] Domenico Cozzetto et al. “FFPred 3: feature-based function prediction for all Gene Ontology domains”. In: *Scientific reports* **6**.*1* (2016), s. 31865.

[57] Petri Törönen, Alan Medlar a Liisa Holm. “PANNZER2: a rapid functional annotation web server”. In: *Nucleic acids research* **46**.*W1* (2018), W84–W88.

[58] Ronghui You, Xiaodi Huang a Shanfeng Zhu. “DeepText2GO: improving large-scale protein function prediction with deep semantic text representation”. In: *Methods* **145** (2018), s. 82–90.

[59] Ronghui You et al. “NetGO: improving large-scale protein function prediction with massive network information”. In: *Nucleic acids research* **47**.*W1* (2019), W379–W387.

[60] Thi Thuy Duong Vu a Jaehee Jung. “Protein function prediction with gene ontology: from traditional to deep learning models”. In: *PeerJ* **9** (2021), e12019.

[61] Davide Chicco, Peter Sadowski a Pierre Baldi. “Deep autoencoder neural networks for gene ontology annotation predictions”. In: *Proceedings of the 5th ACM conference on bioinformatics, computational biology, and health informatics*. 2014, s. 533–540.

[62] Renzhi Cao et al. “ProLanGO: protein function prediction using neural machine translation based on a recurrent neural network”. In: *Molecules* **22**.*10* (2017), s. 1732.

[63] Balázs Szalkai a Vince Grolmusz. “SECLAF: a webserver and deep neural network design tool for hierarchical biological sequence classification”. In: *Bioinformatics* **34**.*14* (2018), s. 2487–2489.

[64] Ahmet Sureyya Rifaioglu et al. “DEEPred: automated protein function prediction with multi-task feed-forward deep neural networks”. In: *Scientific reports* **9**.*1* (2019), s. 7344.

[65] Mohammad Nauman et al. “Beyond homology transfer: Deep learning for automated annotation of proteins”. In: *Journal of Grid Computing* **17** (2019), s. 225–237.

[66] Amirhossein Tavanaei et al. “Towards recognition of protein function based on its structure using deep convolutional networks”. In: *2016 IEEE international conference on bioinformatics and biomedicine (BIBM)*. IEEE. 2016, s. 145–149.

[67] Maxat Kulmanov, Mohammed Asif Khan a Robert Hoehndorf. “DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier”. In: *Bioinformatics* **34**.*4* (2018), s. 660–668.

[68] Vladimir Gligorijević, Meet Barot a Richard Bonneau. “deepNF: deep network fusion for protein function prediction”. In: *Bioinformatics* **34**.*22* (2018), s. 3873–3881.

[69] Rui Fa et al. “Predicting human protein function with multi-task deep neural networks”. In: *PloS one* **13**.*6* (2018). DOI: `journal.pone.0198216`.

[70] Fuhao Zhang et al. “DeepFunc: a deep learning framework for accurate prediction of protein functions from protein sequences and interactions”. In: *Proteomics* **19**.*12* (2019), s. 1900019.

[71] Fuhao Zhang et al. “A deep learning framework for gene ontology annotations with sequence-and network-based information”. In: *IEEE/ACM transactions on computational biology and bioinformatics* **18**.*6* (2020), s. 2208–2217.

[72] Yideng Cai, Jiacheng Wang a Lei Deng. “SDN2GO: an integrated deep learning model for protein function prediction”. In: *Frontiers in bioengineering and biotechnology* **8** (2020), s. 391.

[73] Zhihua Du et al. “Deepadd: protein function prediction from k-mer embedding and additional features”. In: *Computational Biology and Chemistry* **89** (2020), s. 107379.

[74] Cen Wan a David T Jones. “Protein function prediction is improved by creating synthetic feature samples with generative adversarial networks”. In: *Nature Machine Intelligence* **2**.*9* (2020), s. 540–550.

[75] Junyi Li et al. “Gonet: a deep network to annotate proteins via recurrent convolution networks”. In: *2020 IEEE international conference on bioinformatics and biomedicine (BIBM)*. IEEE. 2020, s. 29–34.

[76] Iddo Friedberg. “Automated protein function prediction-the genomic challenge”. In: *Briefings in bioinformatics* **7**.*3* (2006), s. 225–242.

[77] Andrew G McDonald, Sinead Boyce a Keith F Tipton. “Enzyme classification and nomenclature”. In: *eLS* (2015), s. 1–11.

[78] Milton H Saier Jr et al. “The transporter classification database”. In: *Nucleic acids research* **42**.*D1* (2014), s. D251–D258.

[79] Christophe Dessimoz a Nives Škunca. *The gene ontology handbook*. Springer Nature, 2017. ISBN: 1013267710.

[80] Michael Ashburner et al. “Gene ontology: tool for the unification of biology”. In: *Nature genetics* **25**.*1* (2000), s. 25–29.

[81] Suzi A Aleksander et al. “The gene ontology knowledgebase in 2023”. In: *Genetics* **224**.*1* (2023). DOI: iyad031.

[82] Saul B Needleman a Christian D Wunsch. “A general method applicable to the search for similarities in the amino acid sequence of two proteins”. In: *Journal of molecular biology* **48**.*3* (1970), s. 443–453.

[83] Haider Syed a Amar K Das. “Temporal needleman-wunsch”. In: *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE. 2015, s. 1–9.

[84] Michael Sipser. “Introduction to the Theory of Computation”. In: *ACM Sigact News* **27**.*1* (1996), s. 27–29.

[85] Stephen F Altschul et al. “Basic local alignment search tool”. In: *Journal of molecular biology* **215**.*3* (1990), s. 403–410.

[86] Martin Mareš a Tomáš Valla. *Průvodce labyrintem algoritmů*. CZ.NIC, z.s.p.o., 2017. ISBN: 978-80-88168-63-8.

[87] Qingyu Chen et al. “Evaluation of CD-HIT for constructing non-redundant databases”. In: *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2016, s. 703–706.

[88] Natallia Kokash. “An introduction to heuristic algorithms”. In: *Computer Science, Mathematics* (2005), s. 1–8.

[89] Temple F Smith, Michael S Waterman et al. "Identification of common molecular subsequences". In: *Journal of molecular biology* **147**.*1* (1981), s. 195–197.

[90] William R Pearson a David J Lipman. "Improved tools for biological sequence comparison." In: *Proceedings of the National Academy of Sciences* **85**.*8* (1988), s. 2444–2448.

[91] Stephen F Altschul et al. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". In: *Nucleic acids research* **25**.*17* (1997), s. 3389–3402.

[92] Richard Durbin et al. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998. ISBN: 9780511790492. DOI: CBO9780511790492.

[93] Robert C. Edgar. "Search and clustering orders of magnitude faster than BLAST". In: *BIOINFORMATICS* **26**.*19* (2010), s. 2460–2461. ISSN: 1367-4803. DOI: btq461.

[94] Enzo Rucci et al. "Oswald: O pencl s mith–w aterman on a ltera's fpga for l arge protein d atabases". In: *The International Journal of High Performance Computing Applications* **32**.*3* (2018), s. 337–350.

[95] Martin Steinegger a Johannes Söding. "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets". In: *Nature biotechnology* **35**.*11* (2017), s. 1026–1028.

[96] Jan Fassler a Peter Cooper. "BLAST glossary". In: *BLAST® Help* (2011).

[97] Michel Van Kempen et al. "Fast and accurate protein structure search with Foldseek". In: *Nature Biotechnology* **42**.*2* (2024), s. 243–246.

[98] OpenAI. *Odpověď na dotaz: Rozšíření popisu proteinové struktury*. Poskytnuto službou ChatGPT od OpenAI. PROMPT: "Prosím zkus mi tohle rozepsat na víc vět (a doplnit): Proteinová struktura nelze popsat posloupností písmen, ale je nutno uvést prostorové uspořádání. To je standardně zapisováno koordináty jednotlivých atomů.", [příspěvek vytvořený umělou inteligencí]. 2024. URL: https://www.openai.com/chatgpt (cit. 17. 07. 2024).

[99] Frances C Bernstein et al. "The Protein Data Bank: a computer-based archival file for macromolecular structures". In: *Journal of molecular biology* **112**.*3* (1977), s. 535–542.

[100] Helen Berman, Kim Henrick a Haruki Nakamura. "Announcing the worldwide protein data bank". In: *Nature structural & molecular biology* **10**.*12* (2003), s. 980–980.

[101] OpenAI. *Následně upravená odpověď na dotaz: rozvedení textu o mmCIF/PDBx.* Poskytnuto službou ChatGPT od OpenAI. PROMPT: "Zkus tohle rozepsat do delšího textu (ideální je prodloužení o 100-160 znaků: Průměrný počet atomů v aminokyselině je 19.2. Pokud může v souboru typu *.pdb* být uloženo maximálně 99 999 atomů, tak jde uložit maximálně 5 200 průměrných aminokyselin v řetězci. Z tohoto důvodu vznikl novější formát mmCIF/PDBx, který nemá pevně daný rozsah a tím lze do něj uložit libovolně velké proteinové struktury.", [příspěvek vytvořený umělou inteligencí]. 2024. URL: `https://www.openai.com/chatgpt` (cit. 02. 08. 2024).

[102] Liisa Holm a Chris Sander. "Dali: a network tool for protein structure comparison". In: *Trends in biochemical sciences* **20**.*11* (1995), s. 478–480.

[103] Jean-Francois Gibrat, Thomas Madej a Stephen H Bryant. "Surprising similarities in structure comparison". In: *Current opinion in structural biology* **6**.*3* (1996), s. 377–385.

[104] Ilya N Shindyalov a Philip E Bourne. "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path." In: *Protein engineering* **11**.*9* (1998), s. 739–747. DOI: `11.9.739`.

[105] Yuzhen Ye a Adam Godzik. "Flexible structure alignment by chaining aligned fragment pairs allowing twists". In: *Bioinformatics* **19**.*suppl_2* (2003), s. ii246–ii255.

[106] Runze Dong et al. "mTM-align: a server for fast protein structure database search and multiple protein structure alignment". In: *Nucleic acids research* **46**.*W1* (2018), W380–W386. DOI: `gky430`.

[107] Zi Liu et al. "TM-search: An Efficient and Effective Tool for Protein Structure Database Search". In: *Journal of Chemical Information and Modeling* (2024).

[108] Tomonori Fukutani et al. "G-RMSD: Root mean square deviation based method for three-dimensional molecular similarity determination". In: *Bulletin of the Chemical Society of Japan* **94**.*2* (2021), s. 655–665.

[109] Marcos R Betancourt a Jeffrey Skolnick. "Universal similarity measure for comparing protein structures". In: *Biopolymers: Original Research on Biomolecules* **59**.*5* (2001), s. 305–309.

[110] Yang Zhang a Jeffrey Skolnick. "Scoring function for automated assessment of protein structure template quality". In: *Proteins: Structure, Function, and Bioinformatics* **57**.*4* (2004), s. 702–710.

[111] John Jumper et al. "Highly accurate protein structure prediction with AlphaFold". In: *Nature* **596**.*7873* (2021), s. 583–589. DOI: `10.1038/s41586-021-03819-2`.

[112] Alexander Rives et al. “Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences”. In: *Proceedings of the National Academy of Sciences* **118**.*15* (2021). DOI: pnas.2016239118.

[113] Minkyung Baek et al. “Accurate prediction of protein structures and interactions using a three-track neural network”. In: *Science* **373**.*6557* (2021), s. 871–876. DOI: science.abj8754.

[114] Ruidong Wu et al. “High-resolution de novo structure prediction from primary sequence”. In: *bioRxiv* (2022). DOI: 2022.07.21.500999.

[115] Duolin Wang et al. “S-PLM: Structure-aware Protein Language Model via Contrastive Learning between Sequence and Structure”. In: *bioRxiv* (2024). DOI: 10.1101/2023.08.06.552203.

[116] Jeliazko R Jeliazkov, Diego del Alamo a Joel D Karpiak. “ESMFold hallucinates native-like protein sequences”. In: *bioRxiv* (2023), s. 2023–05.

[117] Mihaly Varadi et al. “AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models”. In: *Nucleic acids research* **50**.*D1* (2022), s. D439–D444.

[118] Peter Buneman et al. “Curated databases”. In: *Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. 2008, s. 1–12.

[119] Amos Bairoch a Rolf Apweiler. “The SWISS-PROT protein sequence data bank and its new supplement TREMBL”. In: *Nucleic acids research* **24**.*1* (1996), s. 21–25.

[120] OpenAI. *Odpověď na dotaz: úprava textu o nerelevanci GO vrstev.* Poskytnuto službou ChatGPT od OpenAI. PROMPT: ”Dokážeš to učesat? (rozsekat na víc vět a napsat čitelněji?): Tato varianta je vybrána z důvodu nerelevance vzdálenosti jako měřítka kvality odhadu funkce, protože přiřazení GO anotace k proteinu reflektuje hloubku vědeckého poznání v konkrétním tématické oblasti a dalších informací, které jsou o daném proteinu známé v době anotovaní.”, zdroj myšlenky: [121], [příspěvek vytvořený umělou inteligencí]. 2024. URL: https://www.openai.com/chatgpt (cit. 19. 07. 2024).

[121] Ptáček Jakub a Balážová Faltejsková Kateřina. *Diskuze nad relevancí vrstev GO*. Praha, 5.3.2024. [osobní komunikace].

[122] David J Lipman a William R Pearson. “Rapid and sensitive protein similarity searches”. In: *Science* **227**.*4693* (1985), s. 1435–1441. DOI: science.2983426.

[123] David Binns et al. “QuickGO: a web-based tool for Gene Ontology searching”. In: *Bioinformatics* **25**.*22* (2009), s. 3045–3046.

[124] Rebecca Jackson et al. “OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies”. In: *Database* **2021** (2021). DOI: baab069.

[125] Georges Teto et al. “Gag P2/NC and pol genetic diversity, polymorphism, and drug resistance mutations in HIV-1 CRF02_AG-and non-CRF02_AG-infected patients in Yaounde, Cameroon”. In: *Scientific reports* **7**.*1* (2017), s. 14136.

[126] Raquel Blanco, Luis Carrasco a Iván Ventoso. “Cell killing by HIV-1 protease”. In: *Journal of Biological Chemistry* **278**.*2* (2003), s. 1086–1093.

[127] Wesley I Sundquist a Hans-Georg Kräusslich. “HIV-1 assembly, budding, and maturation”. In: *Cold Spring Harbor perspectives in medicine* **2**.*7* (2012), a006924.

[128] OpenAI. *Generated detailed BibTeX citation for the book ”Protein Function Prediction”*. Poskytnuto službou ChatGPT od OpenAI. Původní citace knihy: Kihara, Daisuke. ”Protein Function Prediction”. Springer, 2017. - Tato citace také vygenerována ChatGPT, [příspěvek vytvořený umělou inteligencí]. 2024. URL: https://www.openai.com/chatgpt (cit. 13. 07. 2024).

Příloha A

Soupis aminokyselin

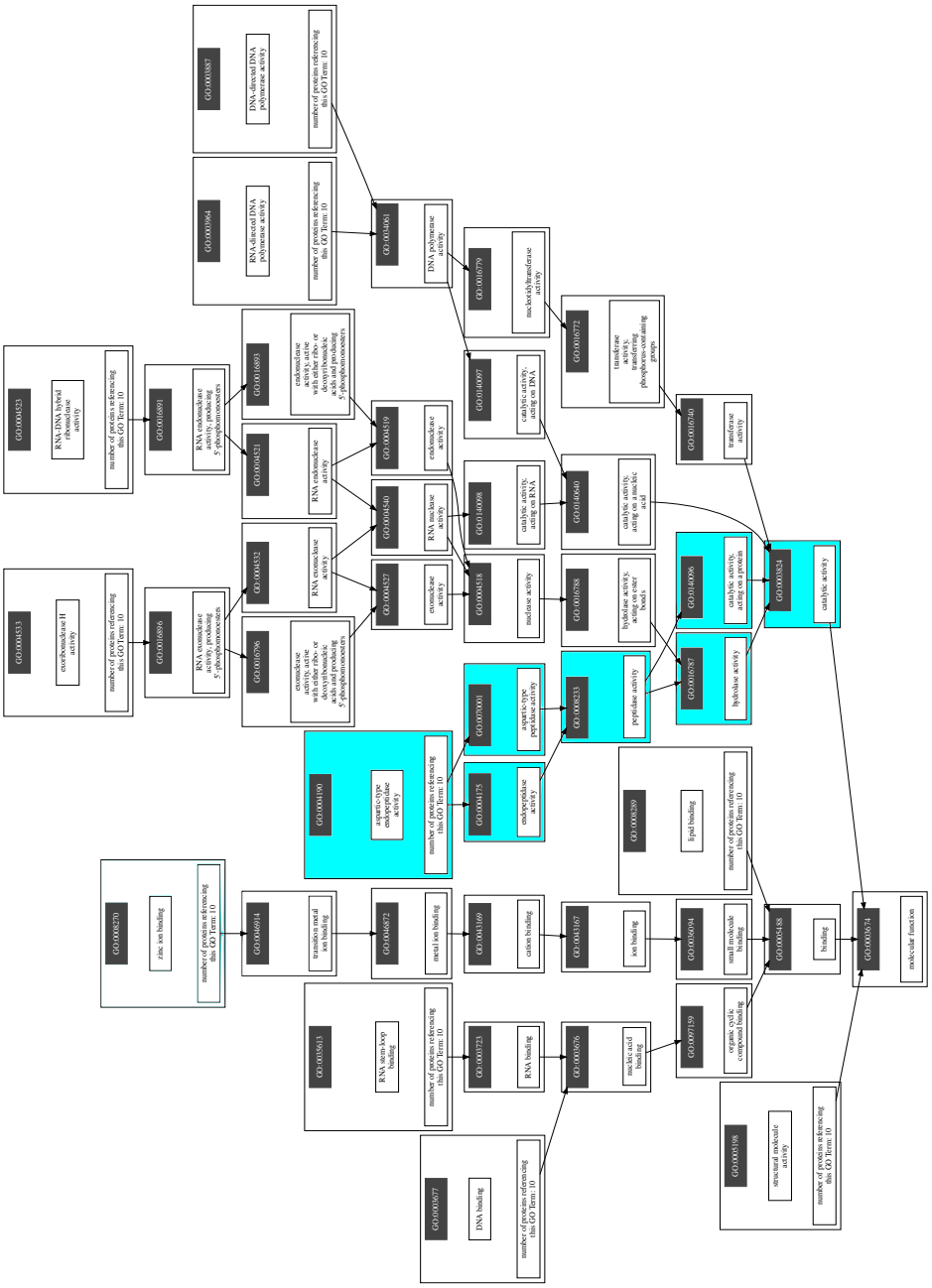
Zkratka	Název	Kodóny	Vlastnosti	Vzorec
A	Alanin	GCU, GCG, GCA, GCC	hydrofobní, malá	
C	Cystein	UGU, UGC	hydrofobní, malá, polární	
D	Asparagová kyselina	GAU, GAC	nabitá, malá	
E	Glutamová kyselina	GAA, GAG	nabitá	
F	Fenylalanin	UUU, UUC	aromatický, hydrofobní	
G	Glycin	GGG, GGA, GGU, GGC	hydrofobní, malá	
H	Histidin	CAC, CAU	aromatický, hydrofobní, malá	
I	Isoleucin	AUU, AUC, AUA	hydrofobní	
K	Lysin	AAA, AAG	hydrofobní, nabitá	
L	Leucin	UUA, UUG, CUU, CUC, CUA, CUG	hydrofobní	
M	Methionin	AUG	hydrofobní	
N	Asparagin	AAU, AAC	polární, malá	
P	Prolin	CCC, CCG, CCA, CCU	malá	
Q	Glutamin	CAA, CAG	polární	
R	Arginin	CGG, CGC, CGA, CGU, AGA, AGG	nabitá	
S	Serin	UCC, UCU, UCA, UCG, AGU, AGC	polární, malá	
T	Threonin	ACA, ACC, ACU, ACG	hydrofobní, malá, polární	
V	Valin	GUG, GUU, GUA, GUC	hydrofobní, malá	
W	Tryptophan	UGG	hydrofobní, polární, aromatická	
Y	Tyrosin	UAU, UAC	hydrofobní, polární, aromatická	

Tabulka A.1 Soupis 20 základních aminokyselin s uvedenou jednopísmennou zkratkou a odpovídajícími kodóny v genetickém kódu.

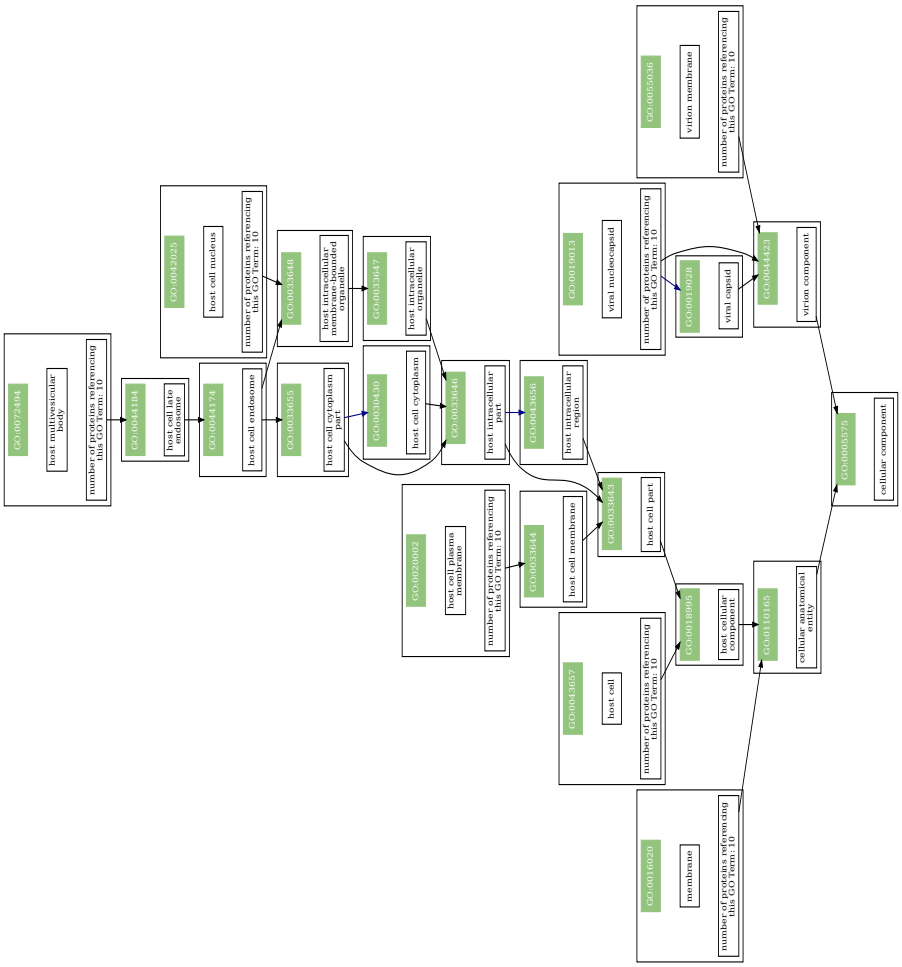
kodony v mRNA se skládají z adeninu (A), guaninu (G), cytosinu (C) a uracilu (U).

Příloha B

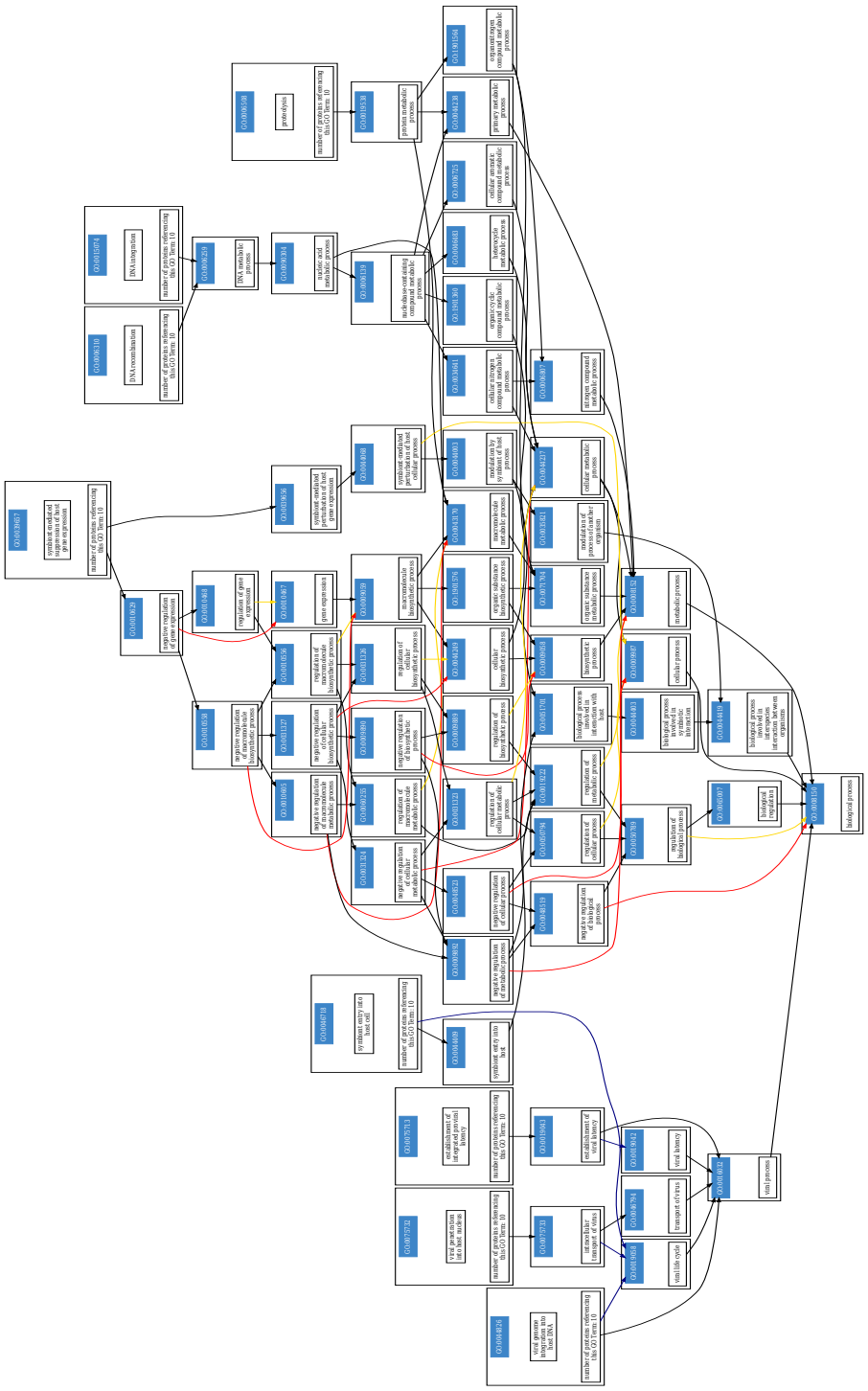
Výstupy z programu GOLizard



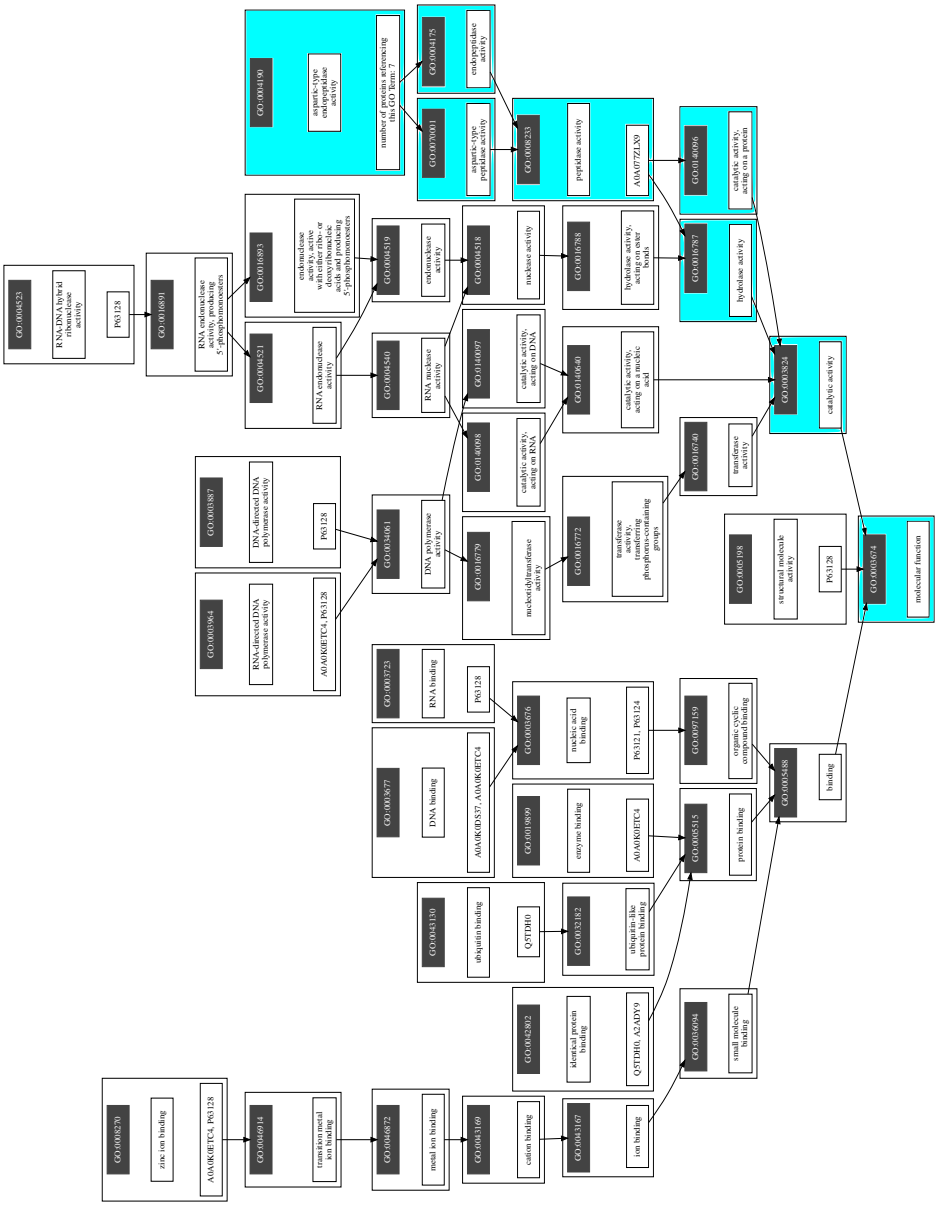
Obrázek B.1 Výstup programu GOLizard v GO oblasti molekulární funkce.
Jedná se o variantu kdy byly vyhledány sekvenčně podobné proteiny (BLAST) a nebyla provedena klastrovací analýza. Je zobrazen podgraf GO anotací odpovídajících nalezeným proteinům. Vstupní data představovala z 38% mutovaná sekvence HIV-proteázy.



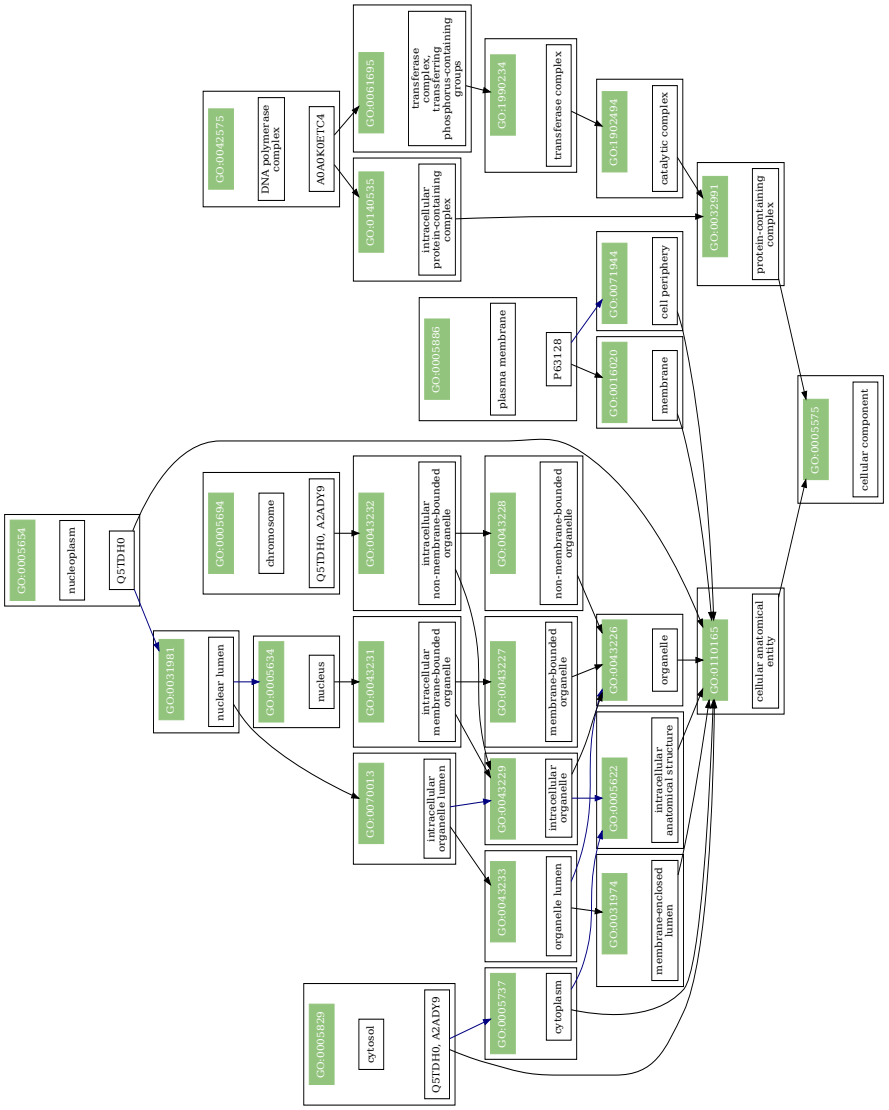
Obrázek B.2 Výstup programu GOLizard v GO oblasti buněčná komponenta. Jedná se o variantu kdy byly vyhledány sekvenčně podobné proteiny (BLAST) a nebyla provedena klastrovací analýza. Je zobrazen podgraf GO anotací odpovídajících nalezeným proteinům. Vstupní data představovala z 38% mutovaná sekvence HIV-proteázy.



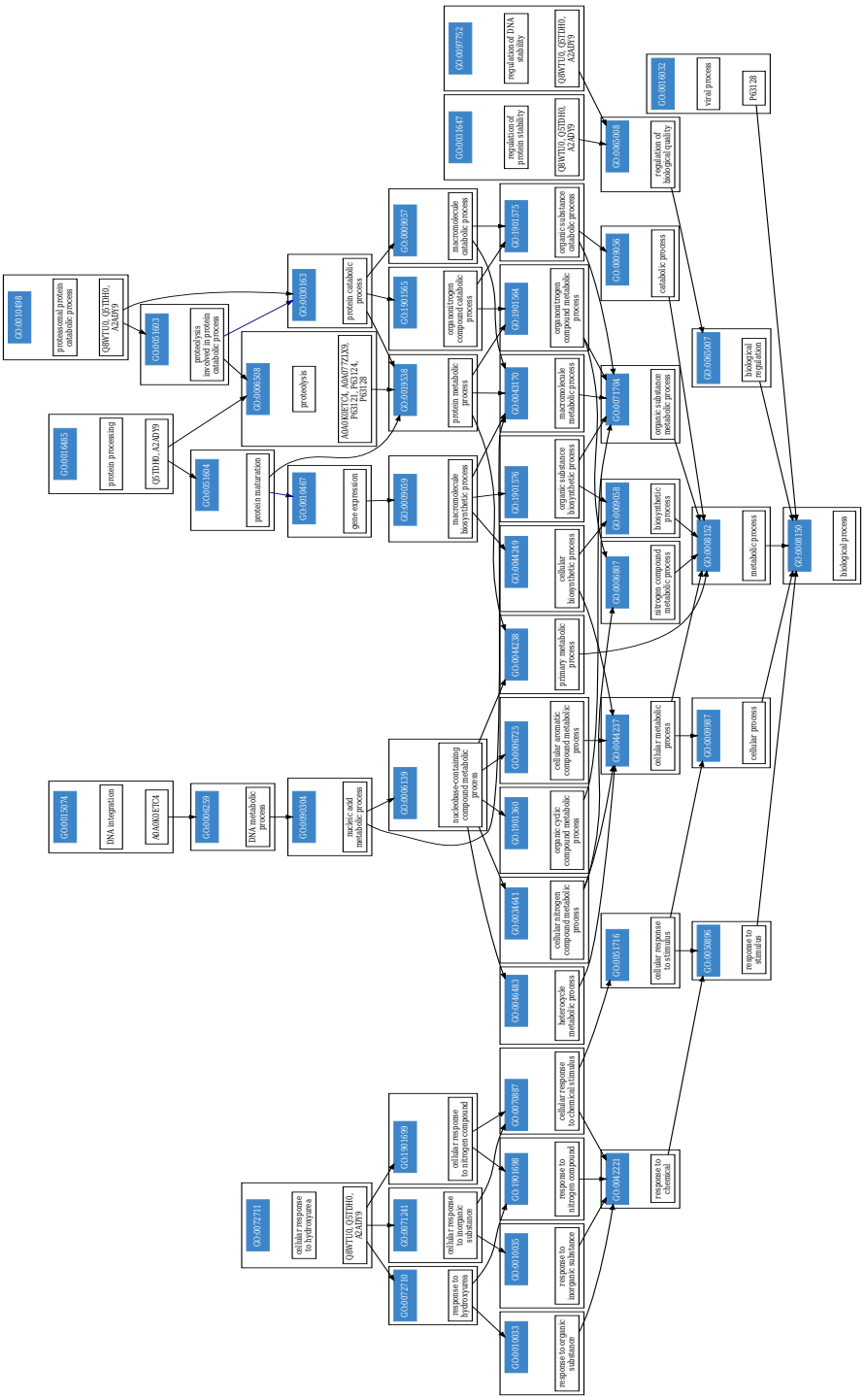
Obrázek B.3 Výstup programu GOLizard v GO oblasti biologický proces.
Jedná se o variantu kdy byly vyhledány sekvenčně podobné proteiny (BLAST) a nebyla provedena klastrovací analýza. Je zobrazen podgraf GO anotací odpovídajících nalezeným proteinům. Vstupní data představovala z 38% mutovaná sekvence HIV-proteázy.



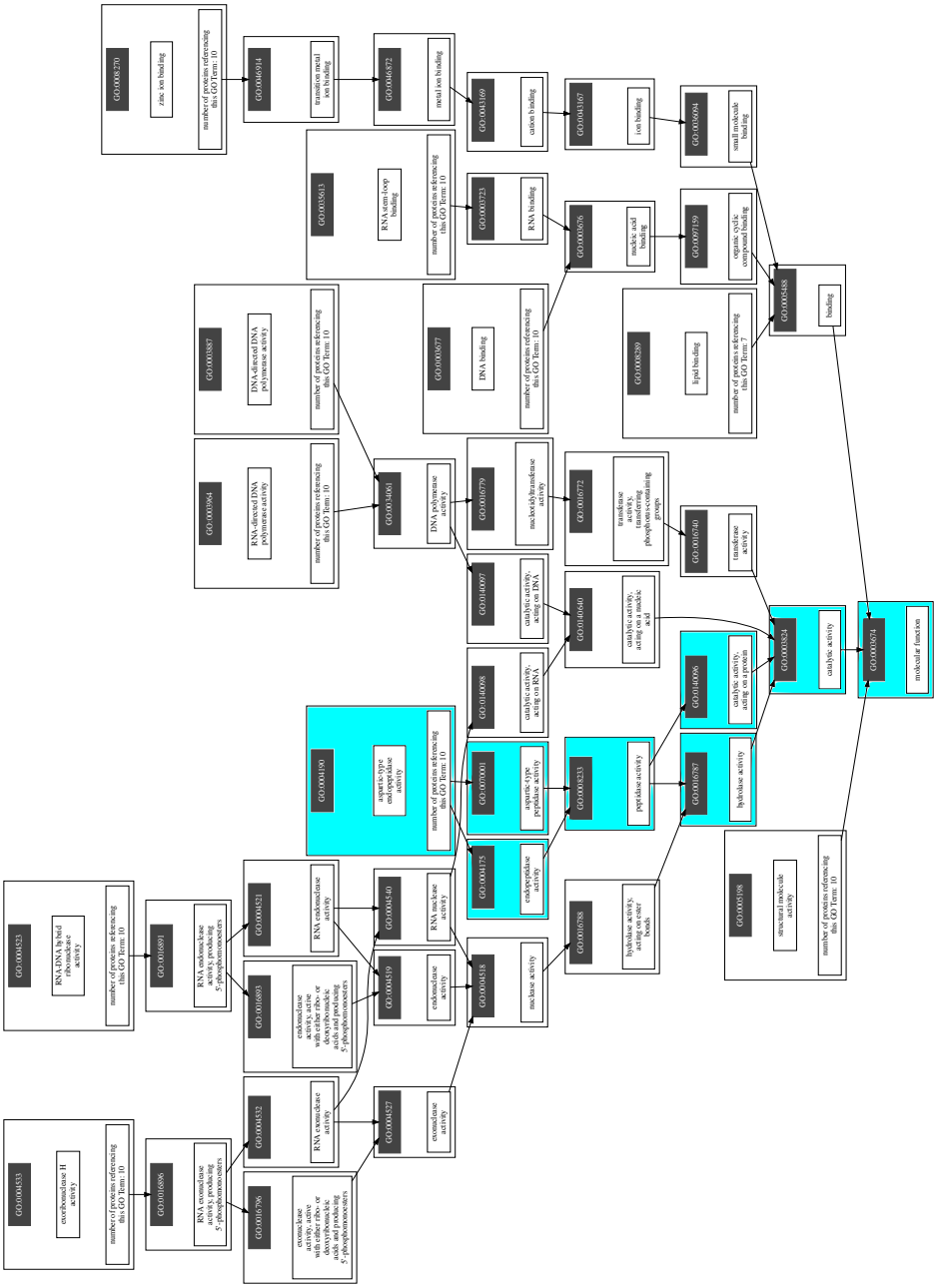
Obrázek B.4 Výstup programu GOLizard v GO oblasti molekulární funkce.
Jedná se o variantu kdy byly vyhledány strukturně podobné proteiny (FoldSeek) a nebyla provedena klastrovací analýza. Je zobrazen podgraf GO anotací odpovídajících nalezeným proteinům. Vstupní data představovala z 38% mutovaná sekvence HIV-proteázy.



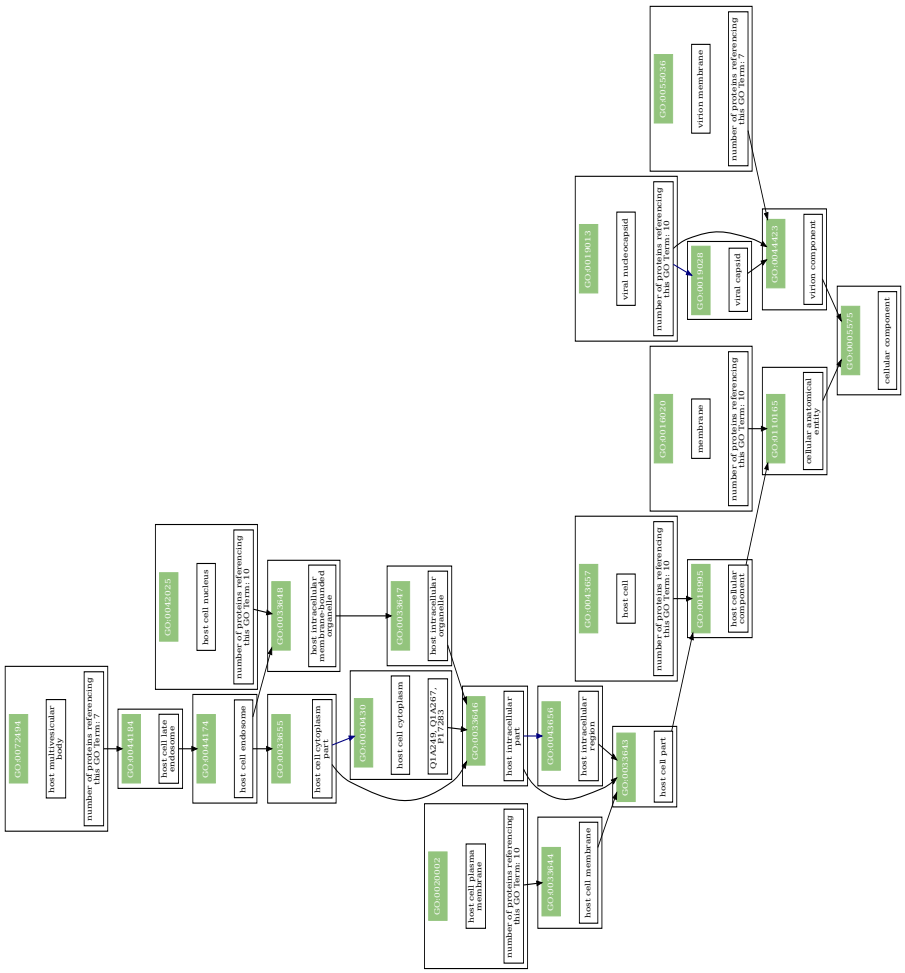
Obrázek B.5 Výstup programu GOLizard v GO oblasti buněčná komponenta. Jedná se o variantu kdy byly vyhledány strukturně podobné proteiny (FoldSeek) a nebyla provedena klastrovací analýza. Je zobrazen podgraf GO anotací odpovídajících nalezeným proteinům. Vstupní data představovala z 38% mutovaná sekvence HIV-proteázy.



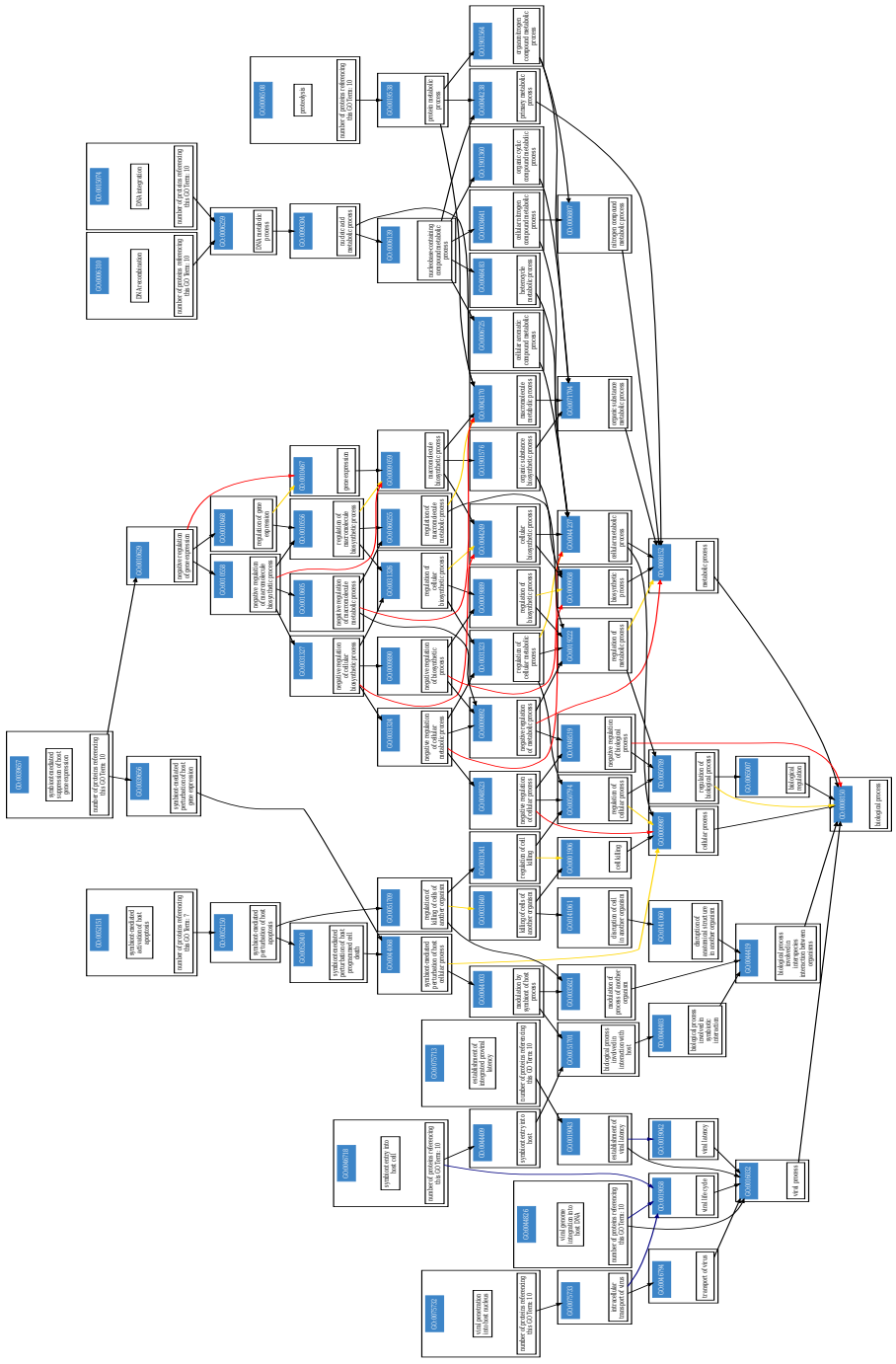
Obrázek B.6 Výstup programu GOLizard v GO oblasti biologický proces.
Jedná se o variantu kdy byly vyhledány strukturně podobné proteiny (FoldSeek) a nebyla provedena klastrovací analýza. Je zobrazen podgraf GO anotací odpovídajících nalezeným proteinům. Vstupní data představovala z 38% mutovaná sekvence HIV-proteázy.



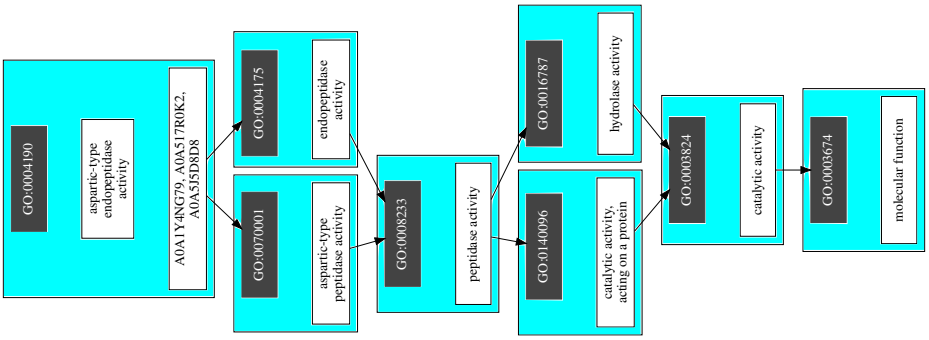
Obrázek B.7 Výstup programu GOLizard v GO oblasti molekulární funkce.
Jedná se o variantu kdy byly vyhledány sekvenčně podobné proteiny (BLAST) a následně bylo provedeno klastrovaní nalezených proteinů podle sekvencí se sekvenční identitou dvojic alespoň 90%. Je zobrazen podgraf GO anotací odpovídajících nalezeným proteinům. Vstupní data představovala z 38% mutovaná sekvence HIV-proteázy.



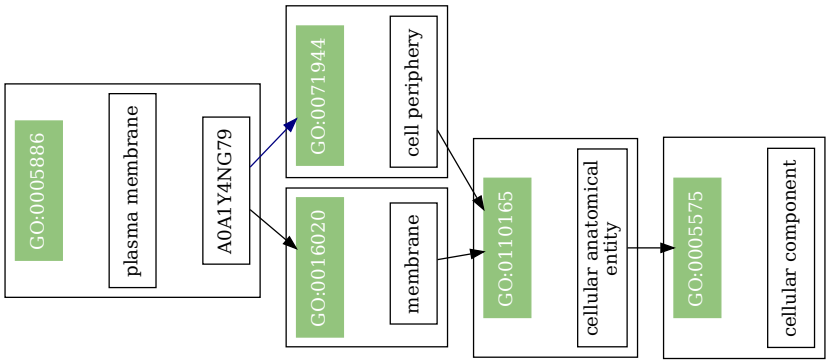
Obrázek B.8 Výstup programu GOLizard v GO oblasti buněčná komponenta.
Jedná se o variantu kdy byly vyhledány sekvenčně podobné proteiny (BLAST) a následně bylo provedeno klastrovaní nalezených proteinů podle sekvencí se sekvenční identitou dvojic alespoň 90%. Je zobrazen podgraf GO anotací odpovídajících nalezeným proteinům. Vstupní data představovala z 38% mutovaná sekvence HIV-proteázy.



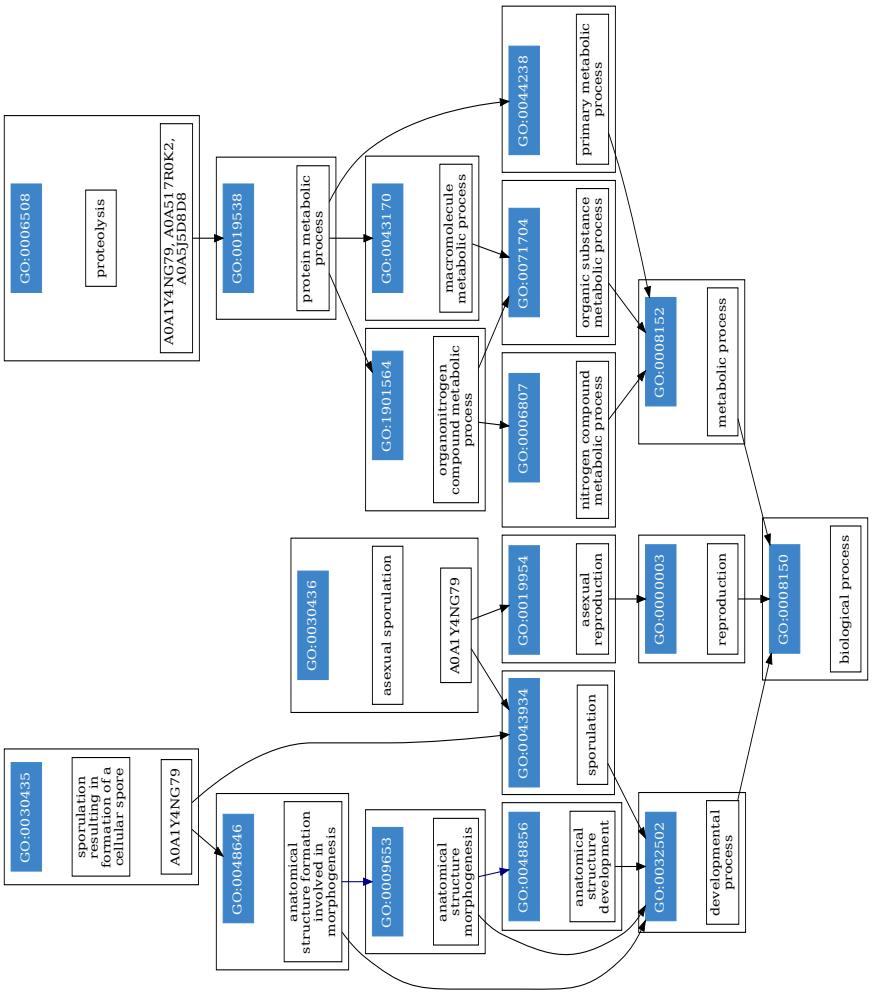
Obrázek B.9 Výstup programu GOLizard v GO oblasti biologický proces.
Jedná se o variantu kdy byly vyhledány sekvenčně podobné proteiny (BLAST) a následně bylo provedeno klastrovaní nalezených proteinů podle sekvencí se sekvenční identitou dvojic alespoň 90%. Je zobrazen podgraf GO anotací odpovídajících nalezeným proteinům. Vstupní data představovala z 38% mutovaná sekvence HIV-proteázy.



Obrázek B.10 Výstup programu GOLizard v GO oblasti molekulární funkce.
Jedná se o variantu kdy byly vyhledány strukturně podobné proteiny (FoldSeek) a následně bylo provedeno klastrovaní nalezených proteinů podle sekvencí se sekvenční identitou dvojic alespoň 90%. Je zobrazen podgraf GO anotací odpovídajících nalezeným proteinům. Vstupní data představovala z 38% mutovaná sekvence HIV-proteázy.



Obrázek B.11 Výstup programu GOLizard v GO oblasti buněčná komponenta. Jedná se o variantu kdy byly vyhledány strukturně podobné proteiny (FoldSeek) a následně bylo provedeno klastrovaní nalezených proteinů podle sekvencí se sekvenční identitou dvojic alespoň 90%. Je zobrazen podgraf GO anotací odpovídajících nalezeným proteinům. Vstupní data představovala z 38% mutovaná sekvence HIV-proteázy.



Obrázek B.12 Výstup programu GOLizard v GO oblasti biologický proces.
Jedná se o variantu kdy byly vyhledány strukturně podobné proteiny (FoldSeek) a následně bylo provedeno klastrovaní nalezených proteinů podle sekvencí se sekvenční identitou dvojic alespoň 90%. Je zobrazen podgraf GO anotací odpovídajících nalezeným proteinům. Vstupní data představovala z 38% mutovaná sekvence HIV-proteázy.