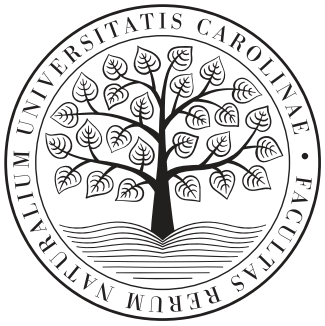


PŘÍRODOVĚDECKÁ
FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Aneta Tranová

Kritické porovnání vybraných databází protein-proteinových interakcí

Katedra buněčné biologie

Vedoucí bakalářské práce: prof. RNDr. Fatima Cvrčková, Dr.
Studijní program: Bioinformatika
Studijní obor: Bioinformatika

Praha 2024

Prohlašuji, že jsem tuto bakalářskou práci vypracovala samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Velké jazykové modely byly využity především ke zlepšení srozumitelnosti určitých vět. Tato práce 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

Poděkování: Chtěla bych upřímně poděkovat své vedoucí práce prof. RNDr. Fatimě Cvrčkové, Dr. za odborné vedení, ochotu, trpělivost a cenné rady, které mi pomohly při zpracování této práce. Rodině za jejich neustálou podporu, spolužákům Aničce, Marovi, Tomovi a Tomášovi za všechnu pomoc a motivaci pokračovat v mém studiu, a v neposlední řadě kamarádům Marki a Honzovi za jejich přátelství a podporu, ať už šlo o cokoliv.

Název práce: Kritické porovnání vybraných databází protein-proteinových interakcí

Autor: Aneta Tranová

Vedoucí bakalářské práce: prof. RNDr. Fatima Cvrčková, Dr.

Abstrakt: Tato práce zkoumá rozdíly a tok dat mezi jedenácti databázemi protein-proteinových interakcí (PPI). Řeší nedostatek povědomí uživatelů o nezávislosti databází a metodách zpracování dat. Analyzuje metody sběru dat, typy poskytovaných informací, pokrytí organismů a vzájemnou závislost databází. Byla zjištěna významná variabilita v počtech proteinů a interakcí, přístupech k validaci a pokrytí druhů. Tato práce poskytuje grafickou mapu toku dat a průniku druhového polrytí, zdůrazňuje nutnost konzultace více databází. Práce by měla pomoct čtenáři se zorientovat v PPI databázích a vybírat vhodné zdroje pro jejich potřeby.

Klíčová slova: protein-protein interakce, PPI, databáze proteinových interakcí, databáze, bioinformatika

Title: Critical comparison of protein-protein interaction databases

Author: Aneta Tranová

Supervisor: prof. RNDr. Fatima Cvrčková, Dr.

Abstract: This thesis explores the differences and data flow among eleven protein-protein interaction (PPI) databases. It addresses the lack of user awareness regarding the independence of databases and data processing methods. The study analyzes data collection methods, types of information provided, organism coverage, and the mutual dependencies between databases. Significant variability was found in the number of proteins and interactions, validation approaches, and species coverage. The thesis provides a graphical map of data flow and species overlap, emphasizing the need to consult multiple databases. The work aims to help readers navigate PPI databases and select appropriate sources for their needs.

Keywords: Protein-protein interactions, PPI, protein interaction databases, databases, bioinformatics

Obsah

Úvod

Protein-proteinové interakce jsou zásadní pro mnoho biologických procesech a jsou klíčové pro správné fungování buněk. Analýza těchto interakcí je nezbytná pro hlubší pochopení různých molekulárních mechanismů, což je prospěšné pro oblasti biomedicíny, farmakologii a bioinformatiku. V posledních desetiletích vzniklo mnoho databází, které data o proteinových interakcích shromažďují a zpřístupňují je veřejnosti. Navzdory široké dostupnosti těchto databází často mezi uživateli nepanuje dobrý přehled o nezávislosti jednotlivých dat a rozdílech ve způsobech, jakými databáze zpracovává svá data. Různé metodologie pro sběr a validaci dat mohou vést k rozdílným úrovním kvality a spolehlivosti informací. Cílem této práce je porovnat vybrané databáze protein-proteinových interakcí a zmapovat tok dat mezi nimi. Tato práce by měla pomoci uživatelům se lépe orientovat v dostupných zdrojích a vybrat vhodnou databázi pro jejich specifické výzkumné potřeby.

1. Obecná charakteristika protein-protein interakcí

1.1 Protein-proteinové interakce

1.1.1 Definice a význam protein-proteinových interakcí

Protein-proteinovými interakcemi v širším smyslu rozumíme jakékoli fyzické nebo funkční vztahy mezi dvěma či více proteiny. Dohromady pak soubor interakcí tvoří komplexní síť, která reguluje různé biologické procesy nezbytné pro správné fungování organismů. (1) Tyto interakce jsou klíčové pro buněčnou signalizaci, imunitní reakce a stabilitu proteinových komplexů. (2) Když se proteiny fyzicky váží a setrvávají spolu, tvoří homodimery (ze stejných proteinů) nebo heterodimery, případně heterooligomery (z různých proteinů). Interakce však zahrnují i vztahy, jako jsou enzym-substrátové interakce (například u proteinkináz), vztahy mezi motorem a cytoskeletálními kolejemi a mnoho dalších (i skutečnost, že dva proteiny jsou např. zmíněny ve stejné publikaci, lze v jistém smyslu pokládat za vztah). Pochopení těchto interakcí bývá obtížné, mimo jiné i kvůli složitým tvarům a chemickým vlastnostem proteinů. (3)

PPI jsou důležité pro mnoho biologických funkcí, jako je katalýza reakcí, různé procesy pro zpracování DNA, přenos signálů a transport molekul. (4) Proteiny jsou dynamické povahy a mění tvar a konformaci v závislosti na prostředí a podmínky, kde se nacházejí, což ztěžuje předpověď i detekci jejich interakcí. Další výzvou jsou například proteiny, které podléhají posttranslačním modifikacím (PTM) a tím mění místa jejich interakcí. Buněčné prostředí a dočasnost interakcí jsou dalším faktorem ztěžující jejich zachycení. (5) I přes všechny tyto výzvy je pochopení PPI zásadní pro odhalení a porozumění složitých biologických mechanismů.

1.1.2 Metody detekce PPI

Tradičně dělíme metody pro detekci PPI na in vitro a in vivo. (6) In vitro metody detekují a pozorují proteinové interakce v kontrolovaných laboratorních podmínkách a poskytují přímé důkazy o dané interakci, ale často vykazují falešně pozitivní výsledky. (7) Do tohoto typu detekce PPI patří například různé metody purifikace proteinových komplexů, jako je třeba tandemová afinitní purifikace, což je metoda pro čištění proteinů, která využívá dvě po sobě jdoucí afinitní purifikační kroky, umožňuje tak izolaci cílového proteinu, ale i právě detekci PPI. Tato metoda se kombinuje s analýzou hmotnostní spektrometrie. (8) Dále se zde řadí i afinitní chromatografie, která je velmi citlivá a může tak zachytit i slabé interakce. Dalšími metodami jsou například ko-imunoprecipitace (Co-IP), proteinové microarray nebo povrchová plasmonová resonance (SPR). (9)

In vivo metody detekují PPI v rámci živých organismů, což umožňuje pozorovat, jak proteiny interagují ve svém přirozeném prostředí, nebo aspoň v prostředí přirozenému blízkém. Patří sem kvasinkový dvouhybridní systém (Y2H), který je pro identifikaci PPI nejčastěji používán, je jednoduchý a není drahý, jen vyžadu-

je, aby interagující proteiny byly lokalizovány v jádře buňky kvasinky. Existuje také mnoho různých variací této metody pro různé typy PPI. (10) Řadí se zde i syntetická letalita, která identifikuje funkční interakce mezi proteiny, aniž by byl nutný jejich přímý fyzický kontakt. (11)

Existuje však mnoho dalších metod pro detekci PPI in vivo. (12) Různé mikroskopické techniky, jako je Förster resonanc energy transfer (FRET) (13) a bimolecular fluorescence complementation (BiFC), umožňují detekci kolokalizace a interakce proteinů. (14) Proximity labelling techniky, jako je BioID a APEX, umožňují identifikaci proteinů v těsné blízkosti označeného proteinu v živých buňkách. (15) Tyto pokročilé techniky poskytují složitější pohled na proteinové interakce v buněčném prostředí.

Žádné z těchto metod nejsou stoprocentní a mohou přinášet falešně pozitivní a negativní výsledky.

1.1.3 Predikce PPI in silico

Způsoby predikce PPI, které můžeme označovat jako in silico metody, využívají výpočetní techniky k analýze již existujících biologických (a literárních) dat. Tyto metody PPI přímo nedetekují, ale agregují dostupná data a mohou také odvozovat potenciální interakce na základě dostupných dat. Jelikož dříve zmíněné experimentální metody jsou často drahé, časové náročné a mohou přinášet falešně pozitivní nebo negativní výsledky, staly se tyto výpočetní metody pro předpověď PPI nezbytnými. (16)

Podle studie Hu a kol (2021) (17) lze modely pro predikci PPI rozdělit na dva hlavní typy - buď založené na sítích (network-based models) nebo integrované modely (integrated models). Výpočetní modely založené na sítích využívají data o struktuře a vzorcích propojení v rámci proteinových interakčních sítí. Díky tomu, že se zvyšuje pokrytí různých interaktomů, jelikož více PPI je experimentálně identifikováno, dokážou tyto modely odhalit existující vzorce interakcí mezi proteiny a na základě toho předpovídat chybějící interakce. Existuje několik klíčových přístupu pro predikci:

(a) Společní sousedé: PPI se předpovídají na základě společných interagujících proteinů, pokud protein A a protein B sdílejí několik takových sousedů, je pravděpodobné, že spolu A a B navzájem také interagují

(b) Síťové cesty (network path): pokud jsou dva proteiny propojeny navzájem krátkými cestami skrze jiné proteiny, je pravděpodobné, že spolu interagují

(c) Globální struktura sítě: v úvahu se bere topologie celé proteinové sítě, její tvar a propojení, a hledají se vzorce na základě kterých by se daly přepovědět další interakce

(d) Geometrický embedding (geometric embedding): tento přístup transformuje proteinovou síť do 3D geometrického prostoru, kde je vztah je mezi proteiny je reprezentován její vzdáleností, proteiny blízko sebe spolu pravděpodobně interagují

Integrované modely využívají dostupných biologických informací o proteinech. Extrahují různé vlastnosti z proteinových sekvencí, struktur, genomických dat a termínů z gene ontology (GO). Zde je několik významných přístupů:

(a) Modely založené na sekvenci (sequence-based): predikují PPI na základě aminokyselinové sekvence a to díky naučeným vzorcům (využívá například sekvenční podobnosti nebo koevoluční analýzy)

(b) Modely založené na struktuře (structure-based): určují jak dobře struktury dvou proteinů k sobě sedí tím, že se podívá na jejich 3D struktury

(c) Modely založené na genomice (genomic-based): predikují PPI na základě genomických vlastností jako je genová fúze, pořadí genů a fylogenetický profil

(d) Modely založené na GO (GO-based): pokud jsou dva proteiny podobně klasifikovány v rámci Gene Ontology (GO) terminologie, mají pravděpodobně podobnou funkci

Text mining

Text mining je metoda, která automatizuje proces získávání užitečných informací z velkého množství textu. V kontextu predikce PPI identifikuje významné společné výskyty názvů genů ve vědeckých článcích. Pokročilé metody využívají zpracování přirozeného jazyka (NLP) k vytváření sémantických sítí, které pomáhají predikovat PPI tím, že chápou vztahy a kontexty, ve kterých se názvy proteinů vyskytují. (18)

1.1.4 Typy protein-proteinových interakcí

PPI lze rozdělit na základě různých vlastností. Jedním z hlavních kritérií je stabilita reakcí. Stabilní interakce tvoří dlouhodobé a pevné proteinové komplexy, zatímco tranzientní interakce jsou krátkodobé a často se vyskytují při signalizaci nebo metabolických procesech. (19) PPI mohou být dále klasifikovány podle toho, zda jde o interakce přímé nebo nepřímé. Přímé interakce znamenají, že dochází k fyzickému kontaktu mezi proteiny, nepřímé interakce jsou zprostředkovány jinými molekulami nebo prostřednictvím série intermediárních kroků. (20) Za nepřímé interakce PPI můžeme považovat například funkční asociace, protože proteiny spolupracují na dosažení specifických biologických funkcí a to na základě nepřímých vazeb. (21) Dále lze PPI rozdělit podle funkce (např. enzym-substrát, signalizace,...), struktury (např. interakce mezi doménami/motivy) nebo biologického kontextu (např. intracelulární vs. intercelulární interakce).

Jako ortologické označujeme interakce mezi proteiny, které jsou ortology. Ortologické proteiny jsou homologní proteiny ve dvou nebo více druzích, které vznikly v důsledku speciace a obvykle mají podobné funkce v těchto různých organismech (22), na rozdíl od paralogů, které jsou produkty genových duplikací (23).

1.2 Databáze PPI

Databáze PPI shromažďují data o tom, jak proteiny navzájem interagují. Tyto databáze můžeme rozdělit na tři hlavní typy: primární databáze, integrované databáze a predikční databáze. Primární databáze PPI shromažďují a uchovávají experimentálně ověřené interakce proteinů z vědecké literatury. Jejich klíčovou vlastností je vysoká kvalita dat. Integrované databáze (někdy označované i jako metadatabáze) využívají data z primárních i jiných databází a kombinují je

s daty z dalších databází, kontrolovaných slovníků a literatury, což poskytuje integrovaný a komplexní pohled na PPI. Vzhledem k jejich povaze budou vždy méně aktuální než původní primární databáze. (24) Poslední typem jsou predikční databáze kombinující experimentální data s výpočetními metodami predikce PPI, čímž rozšiřují spektrum možných interakcí včetně těch, které dosud nebyly experimentálně potvrzeny. (25)

1.2.1 Integrace externích zdrojů pro PPI databáze

Databáze interakcí zpravidla čerpají z různých dalších specializovaných databází, což umožňuje komplexnější a podrobnější analýzu PPI. Jsou to například databáze drah, jako KEGG (26) nebo Reactome (27), které poskytují informace o konkrétních biologických procesech, ve kterých se daná interakce může vyskytovat a tím objasnit její biologický význam. Databáze funkcí a anotací proteinů jsou také klíčové ve výzkumu PPI. UniProt (28) slouží jako úložiště pro sekvenční a funkční informace o proteinech a poskytuje tak podrobnější pohled na jednotlivé proteiny. Tento zdroj doplňuje databáze Gene Ontology (29), která nabízí strukturovaný rámec pro reprezentaci genů a jejich atributů. Některé PPI databáze integrují data z databází nemocí a fenotypů, zvláště lidských (OMIM (30) a ClinVar (31)) nebo strukturních databází (PDB (32)). Možností je spousta a záleží, na co se konkrétní databáze PPI zaměřuje.

1.2.2 IMEx Consortium

IMEx Consortium (International Molecular Exchange) (33) je iniciativa zaměřená na poskytování vysoce kvalitních, standardizovaných dat o PPI. Jde o spolupráci několika veřejných poskytovatelů interakčních dat, které sdílejí úsilí při kurátorování literatury a poskytují neredundantní sadu proteinových interakcí prostřednictvím společného webového rozhraní. Byla vyvinuta společná pravidla pro kurátování a zaveden jednotný standard pro kurátorované záznamy. Aktivními členy konsorcia je několik významných databází(34), které jsou nalezeny v Tabulce 1.1.

Vše až na poslední tři jsou databáze PPI; UniProt[1], SIB[2] a EMBL-EBI[3] sice nejsou databáze PPI, ale poskytují data nebo nástroje relevantní pro PPI. Celý IMEx dataset je přístupný z webového rozhraní databáze IntAct, kde jsou data centralizovaná, nebo na portálu mentha. Data jsou skórovaná pomocí implementace MI skóre. (33) (více o MI skóre v 1.2.4 Validace dat, Kvalitativní kontrola)

[1]UniProt Consortium, "UniProt", https://www.uniprot.org
[2]Swiss Institute of Bioinformatics, https://www.sib.swiss
[3]European Bioinformatics Institute, https://www.ebi.ac.uk

Název	Klíčové vlastnosti
DIP (35)	poskytuje kurátorovaná PPI data z experimentálních i výpočetních zdrojů
IntAct (36)	poskytuje manuálně kurátorovaná data o molekulárních interakcí a nástroje pro jejich analýzu
MINT (37)	poskytuje vysoce kvalitní kurátorovaná data o protein interakcí z vědecké literatury
IID (38)	čerpá z několika zdrojů a poskytuje PPI data
MatrixDB (39)	zaměřuje se na kurátorované interakce mezi extracelulární matrix a receptory
InnateDB (40)	predikuje a kurátorsky zpracovává interakce a dráhy specifické pro nespecifický imunitní systém savců
UniProt (28)	poskytuje data o sekvencích a funkcích proteinů
SIB (41)	organizace poskytující služby a nástroje v oblasti bioinformatiky, včetně databází a softwarových platforem
EMBL-EBI (42)	nabízí bioinformatické služby a databáze primárně pro výzkum molekulární biologie

Tabulka 1.1: Aktivní členové IMEx konsorcia

1.2.3 Sběr dat

Data pro databáze PPI by měla být pečlivě získaná, aby byla zaručena spolehlivost a úplnost informací. U primárních databází jsou data obvykle shromažďována z vědecké literatury odborníky, kteří manuálně a pečlivě extrahují experimentálně ověřené interakce. (viz 1.2.4 Validace dat, Manuální kurace).

Integrované databáze pro sběr dat používají kombinaci automatizovaných metod, manuální kurace a integrují data z více zdrojů. Automatizované nástroje, jako jsou text mining a zpracování přirozeného jazyka (NLP), prohledávají vědeckou literaturu pro PPI interakce, které následně bývají přezkoumávány odborníky. Výzkumníci mohou také zasílat svá experimentální data, která jsou opět kurátorsky ověřena. Také se používají kontrolované slovníky za účelem standardizace dat.

Než se predikční databáze dostanou k předpovídání potenciálních PPI, často začínají integrací experimentálně ověřených interakcí z primárních a integrovaných databází, jejichž data by měla poskytovat spolehlivý základ pro další analýzy a predikce.

1.2.4 Validace dat

Pro zajištění přesnosti a spolehlivostí dat o PPI, je nutné je nějakým způsobem validovat. Následující metody jsou běžně používány pro validaci dat v databázích PPI.

Manuální kurace

Jedná se o proces zahrnující ruční přezkoumání vědecké literatury odborníky, kteří potvrzují správnost hlášených proteinových interakcí. Tito odborníci musí

vědecké články pečlivě přečíst, interpretovat jejich výsledky a extrahovat informace o proteinových interakcí. Informace o PPI jsou převedena z volně psaného textu do strukturovaného formátu, což se dělá pomocí kontrolovaných slovníků, které standardizují názvy proteinů, typy interakcí a použitých experimentálních metod. Takto získána data jsou velmi přesná, ale náročná na práci a čas. Tento způsob validace je typický pro primární databáze. (43; 44)

Experimentální ověření

Experimentální ověření zahrnuje provádění laboratorních experimentů k potvrzení konkrétních proteinových interakcí. Tento proces je důležitý v případech, kdy původní data byla získaná automatizovanými metodami nebo jsou-li pochybnosti o jejich přesnosti. Mohou být použité různé biochemické metody, které se používají i pro detekci PPI (viz 1.1.2 Detekce PPI) nebo biofyzikální metody, které poskytují kvantitativní údaje o interakcí, jako jsou síla nebo kinetka vazby (např. povrchová plasmonová resonance (SPR) nebo izotermická titrační kalorimetrie (ITC)).(10; 7)

Výpočetní validace

Výpočetní validace používá různé výpočetní metody k předpovědi a ověření správnosti PPI na základě dat o nich. Využívá algoritmy a analytické nástroje pro identifikaci pravděpodobných interakcí a ke zvyšování přesnosti dat v PPI databázích. Tyto metody jsou probírány v 1.1.3 Způsoby predikce PPI, používají se tedy k predikci interakce, což může být doplňujícím důkazem dané interakce (v kombinaci s například experimentálními důkazy).

Kvalitativní kontrola

Do toho typu validace dat v databázích PPI lze zařadit kontrola redundance a konzistence dat. Co se týče kontroly kvality jednotlivých interakcí, tak to lze určit pomocí přiřazení skóre důvěryhodnosti (nebo skóre spolehlivosti). Každá databáze k výpočtu tohoto skóre může přistupovat jinak. Obvykle se skóre počítá na základě dostupných důkazů o interakci. Důkazy mohou být různého typu, například experimentální důkazy, kdy je interakce potvrzena různými experimentálními metodami (pokud jich je více, tak je to tzv. ortogonální testování), anebo mohou třeba zohledňovat genomický kontext daných proteinů a tak dál. Záleží na databázi, jak s těmito důkazy zachází - může dávat různým typům důkazům různé váhy pro výpočet a výsledný výpočet skóre se také může lišit. Obecně platí, že čím více nezávislých důkazů o dané interakci existuje, tím vyšší je skóre důvěryhodnosti a tím je interakce považována za spolehlivější.(45; 46)

MI skóre

Jde o skórovací systém pro posuzování PPI na základě pečlivě shromážděných dat, vytvořený IMEx konsorciem. (33) Bere v úvahu hlavní tři faktory:

(a) Metoda použitá k detekci interakce: Jelikož různé experimentální metody se liší svou spolehlivostí, mají různou váhu pro výpočet skóre (například afinitní purifikace následovaná hmotnostní spektrometrií poskytuje přímé důkazy o interakci, a proto má vyšší váhu než méně přímé metody jako je Y2H).

(b) Typ interakce: Přímé vazby nebo fyzické asociace jsou považovány za spolehlivější než ko-lokalizace, která může pouze znamenat, že dva proteiny se

nacházejí ve stejném buněčném prostoru.

(c) Počet publikací uvádějících interakci: Interakce, které byly zmíněny v několika studiích s různými metodami, dostávají vyšší skóre, protože existuje více nezávislých důkazů podporujících jejich platnost.

Metoda použitá k detekci interakce: Jelikož různé experimentální metody se liší svou spolehlivostí, mají různou váhu pro výpočet skóre (například afinitní purifikace následovaná hmotnostní spektrometrií poskytuje přímé důkazy o interakci, a proto má vyšší váhu než méně přímé metody jako je Y2H). Typ interakce: Přímé vazby nebo fyzické asociace jsou považovány za spolehlivější než ko-lokalizace, která může pouze znamenat, že dva proteiny se nacházejí ve stejném buněčném prostoru. Počet publikací uvádějících interakci: Interakce, které byly zmíněny v několika studiích s různými metodami, dostávají vyšší skóre, protože existuje více nezávislých důkazů podporujících jejich platnost.

Výsledné skóre složeno z několika důkazů se normalizuje na škále od 0 do 1. Skóre mezi 0,45 a 1 značí střední spolehlivost, zatímco skóre mezi 0,6 a 1 značí spolehlivost vyšší. Tento skórovací systém používají zdroje jako UniProt(28) a IntAct(36). (47)

1.2.5 Kontrolované slovníky

V oblasti databází PPI jsou kontrolované slovníky (controlled dictionaries) klíčové pro udržování konzistence a přesnosti dat. Pomáhají data normalizovat, což usnadňuje porovnávání a integraci informací z různých zdrojů.(48)

Poskytují jednotný framework pro pojmenování proteinů, popis typů interakcí a specifikaci experimentálních metod. Například pro standardizaci názvů a identifikátorů genů a proteinů se často používá databáze UniProt(28). Synonyma pro tentýž gen nebo protein jsou díky slovníkům sjednocena, což redukuje zmatek a redundanci. Typy proteinových interakcí a experimentální metody jsou popisovány pomocí standardizovaných termínů z ontologií, jako jsou Molecular Interaction Ontology (49) a Experimental Factor Ontology (50).

1.2.6 Formáty dat

Databáze PPI kromě grafického webového rozhraní často nabízí i stažení jejich datasetů, kde jsou všechna data o PPI redukovaná na čistý text. Pro sdílení a analýzu takových dat se používá několik standardizovaných formátů. Jedním z nejrozšířenějších je formát MITAB (Molecular Interaction TAB-delimited), který navržen tak, aby byl snadno čitelný pro člověka i počítač. Jednotlivé vlastnosti interakce jsou odděleny tabulátorem. Umožňuje ukládat širokou škálu informací o PPI včetně identifikátorů proteinů, metody detekce interakce, experimentálních podmínek a dalších relevantních údajů. Dalším známým formátem je PSI-MI XML (Proteomics Standards Initiative Molecular Interaction eXtensible Markup Language), který je ve strukturovaném XML formátu, ale je hůř čitelný pro člověka. [4] Databáze PPI často nabízejí své datasety v těchto formátech, ale některé

[4]Proteomics Standards Initiative, "Molecular Interactions,", https://www.psidev.info/molecular-interactions

mají i své individuální, které jsou pravděpodobně vhodnější pro jejich vlastní účely (např. IID).

1.2.7 Vizualizační nástroje pro PPI

Vizualizace biologických dat je významná pro pochopení proteinových komplexů. Vizualizační nástroje tato data obvykle zpracovávají tak, že z jednotlivé proteiny představují vrcholy a hrany mezi nimi značí interakci. Mezi takové nástroje patří například Cytoscape (51) anebo StringDB(52). Existuje však mnoho dalších vizualizačních nástrojů, kde každý s daty zachází trochu jinak. Studie, která se tímto tématem zabývá, poskytuje srovnání jednotlivých nástrojů a jejich přístupů Jeanquartier a kol., 2015 (53).

2. Vybrané databáze a hodnocené charakteristiky

2.1 Výběr databází

Pro výběr databází, kterými se následující kapitola bude zabývat, nebyla použita objektivní metodologie a nepokouším se ani o vyčerpávající pokrytí tématu, jako tomu bylo například v recentní studii Bajpai AK a kol., 2020 (54). Selekce konkrétních databází byla subjektivní, přihlížela jsem k citovanost databáze, všeobecné známosti mezi kolegy, častému výskytem mezi zdroji jiných databází a citovanosti v jiných komparativních studiích. Byly zahrnuté všechny tři typy databází - primární, integrované i predikční. Pro jednoznačné určení primární databáze byla použita platforma Pathguide[1], která poskytuje seznam zdrojů týkajících se biologických drah a molekulárních interakcí. Dále rozlišuje zdroje dat databází na primární a sekundární. Zde chápu predikční databázi jako takovou, která navíc používá výpočetní modely k předpovědi PPI.

Zaměřuji se zde především (i když nikoli výhradně) na funkční databáze a na databáze se širším druhovým pokrytím. Nediskutuji zde weby zaměřené pouze na vizualizaci PPI ani databáze specializující se na velmi konkrétní typy PPI - například proteiny extracelulární matrix (např. MatrixDB(39)), PPI v buněčných organelách (např. ComPPI(55)) nebo mitochondriální proteiny (např. MitoInteractome(56)). (57) Snažím se zahrnout zejména databáze, které jsou uznávané a široce používané ve vědecké komunitě. Proto podrobněji diskutuji 11 databází, z toho pět primárních: BioGRID, DIP, HuRI, IntAct a MINT; čtyři integrované: APID, HINT, mentha, PICKLE; a dvě predikční: IID a STRING.

2.2 Kritéria hodnocení

V následující kapitole jsou popsány konkrétní databáze. Pro každou databázi je vytvořena karta, která ji charakterizuje. V datových zdrojích jde hlavně o zdroje PPI, proto jsou kurzívou rozlišeny ty, které poskytují primárně jiná data než interakční (například strukturní jako PBD) nebo samy nejsou PPI databází (např. IMEx konsorcium). Dále je jsou rozlišeny neaktivní zdroje, které databáze uvádí, ty jsou vypsány v závorce. V typech interakcí je uvedeno, jestli se databáze zabývá i jinými než PPI. Druhové pokrytí řeší, zda databáze pokrývá aspoň jednoho člena z následujících skupin: prokaryota, houby, rostliny, bezobratlí a obratlovci. Každá databáze pokrývá PPI u člověka. Následně je vypsán přesný nebo aspoň přibližný počet pokrytých organismů. Dále je uveden počet proteinů a interakcí. Tyto informace se mohou lišit v závislosti na tom, jak často databáze aktualizuje svůj PPI dataset, nicméně všechny informace jsou aktuální k datu 27.6.2024. V neposlední řadě je řešeno, jestli je kurátorování manuální či automatické a jakým způsobem se validují data, která jsou do datasetu databáze přidány. Na závěr jaká je frekvence aktualizace databáze a její poslední verze (ze které tato práce vychází). Některé informace nebyly dohledatelné, ty jsou označeny jako „NA“.

[1]Pathguide, "Pathway Resource List,"http://www.pathguide.org

3. Analýza vybraných databází

3.1 Popis jednotlivých databází

3.1.1 Primární databáze

BioGRID

BioGRID (Biological General Repository for Interaction Datasets) je databáze sloužící k uchování a vyhledávaní protein-proteinových, protein-DNA, genetických a chemických interakcí. Pojem „interakce" je v kontextu této databáze brán jako přímé fyzické spojení dvou proteinů, společný výskyt ve stabilním komplexu a genetická interakce.(58) Tyto interakce jsou shromažďovány a kurátorsky zpracovávány především z recenzovaných vědeckých publikací. Zdroje literatury zahrnují přední vědecké časopisy a databáze, jako je například PubMed. BioGRID byl vytvořen v roce 2006 s cílem kurátorsky zpracovat veškerá dostupná data o biologických interakcí v modelovém organismu kvasinky Saccharomyces cerevisiae, následně se databáze rozšířila o všechny hlavní modelové organismy a člověka a také o dalších méně známých druhů.(59) Není aktivním členem IMEx, ale je v něm označen jakožto „Observer".(60)

Název databáze	BioGRID (Biological General Repository for Interaction Datasets)
Webová stránka	https://thebiogrid.org/
Zdroje dat	kurátorovaná literatura
Typy interakcí	protein-protein, protein-DNA, genetické, chemické, PTM
Species coverage	prokaryota, houby, rostliny, bezobratlí a obratlovci
Počet organismů	83
Počet interakcí	více než 2,7 milionu
Počet proteinů	89 642
Způsob kurace	manuální
Validační metoda	publikace potvrzující interakci primárními experimentálními výsledky
Frekv. aktualizace	4 týdny, poslední verze: BIOGRID-4.4.235

Tabulka 3.1: Vlastnosti databáze BioGRID

Tato databáze klade důraz na to, aby všechny důkazy ke kterékoli interakci v databázi byly odvozeny z experimentálních výsledků z primární literatury. Proto BioGRID, jakožto vysoce spolehlivá databáze interakcí, neobsahuje žádná data, která nebyla experimentálně podložena a také data předpovězená na základě výpočetních metod.(59) Právě díky pečlivému zpracování jsou data často používaná jako zlatý standard pro výpočetní studie (například jako trénovací nebo validační množina pro strojové učení). (54)

Mimo databází s vědeckou literaturou jako je již zmíněný PubMed, BioGRID spolupracuje také s databázemi modelových organismů (MODs) jako jsou například SGD, PomBase, FlyBase, Wormbase, a Bio-Analytic Resource for Plant Biology (BAR) a další.(59) Díky spolupráci s dalšími databázemi BioGRID rozšiřuje své zdroje dat, zvyšuje přesnost informací a umožňuje komplexnější analýzy

napříč různými organismy. Nicméně tato databáze stále data primárně sbírá a zpracovává sama a vyvinula si vlastní postupy pro jejich validaci, proto partnerské PPI databáze nejsou uvedené ve zdrojích dat.

DIP

Databáze DIP (Database of Interacting Proteins) integruje různé experimentální důkazy o PPI do jedné online platformy.(61) Jakožto člen IMEx konsorcia DIP spolupracuje s dalšími hlavními databázemi na standardizaci a sdílení dat o PPI na globální úrovni.(33) Kromě podrobného popisu protein-proteinových interakcí je DIP cenný pro pochopení funkcí a vztahů mezi proteiny, analýzu vlastností sítí interagujících proteinů, benchmarkování predikcí protein-proteinových interakcí a studium jejich vývoje. Primárně slouží DIP jako klíčový benchmark pro ověření výkonnosti nových metod predikce PPI.(62) Interakce je zde definována tak, že dva aminokyselinové řetězce byly experimentálně identifikovány jako vzájemně se vážící.(63)

Název databáze	DIP (Database of Interacting Proteins)
Webová stránka	https://dip.doe-mbi.ucla.edu/dip/
Zdroje dat	kurátorovaná literatura, *IMEx*
Typy interakcí	protein-protein
Species coverage	prokaryota, houby, rostliny, bezobratlí a obratlovci
Počet organismů	834
Počet interakcí	81 923
Počet proteinů	28 850
Způsob kurace	manuální, automatická
Validační metoda	kontrola dvěma kurátory, automatizované testy (35), hodnocení kvality podle PVM a EPR metod(64; 61)
Frekv. aktualizace	NA, poslední verze: 5.2.2017

Tabulka 3.2: Vlastnosti databáze DIP

HuRI

Databáze The Human Reference Interactome (HuRI) poskytuje komplexní mapu binárních protein-proteinových interakcí u člověka a je také součástí širšího projektu Human Interactome Project, který je zaměřen na mapování interakčních sítí nezbytné pro lidské buněčné procesy. HuRI využívá vysokokapacitní screeningy, kde jsou použity tři varianty Y2H, což zvyšuje citlivost a taky tím mohou být detekovány i komplementární PPI sady. Dále využívá rozsáhlou sbírku otevřených čtecích rámců (ORF) známou jako lidský ORFeome v9.1, která pokrývá přibližně 90 % lidského proteinkódujícího genomu. Díky této sbírce je možné systematicky prozkoumat téměř celý lidský proteom.(65) Aby byla zajištěna přesnost detekovaných interakcí, jsou výsledky z Y2H ověřovány pomocí dvou dalších testů: Mammalian Protein-Protein Interaction Trap (MAPPIT)(66) a Gaussia Princeps Complementation Assay (GPCA)(67). Tyto nezávislé testy potvrzují zjištěné interakce. (65)

Název databáze	HuRI (The Human Reference Interactome)
Webová stránka	http://www.interactome-atlas.org/
Zdroje dat	kurátorovaná literatura, experimentální metody
Typy interakcí	protein-protein (fyzické)
Species coverage	Homo sapiens
Počet organismů	1
Počet interakcí	64 006
Počet proteinů	9 094
Způsob kurace	manuální
Validační metoda	ortogonální testy, komplementární testy, skórování
Frekv. aktualizace	NA (poslední aktualizace nejspíš v roce 2020)

Tabulka 3.3: Vlastnosti databáze HuRI

IntAct

IntAct je PPI databáze vyvinuta v EMBL-EBI a nabízí několik nástrojů pro analýzu interakcí. Je aktivním členem IMEx konsorcia.(33) Všechna data jsou převážně odvozena z vědecké literatury nebo přímých vkladů dat od odborníků. Tato databáze obsahuje více než milion binárních interakcí, které byly zpracovány ve spolupráci s členy konsorcia IMEx a podle jeho standardů. Mimo binární obsahuje i n-ární interakce, na kterých se podílí více než dvě molekuly (ty se pro komputační a jiné účely mohou rozložit na několik binárních interakcí). (36) IntAct ohodnocuje spolehlivost svých dat na základě MI skóre, které je dále normalizováno a váženo na základě počtu nezávislých důkazů interakce a souvisejících experimentálních metod, které byly použity k pozorování dané interakce. (33)

Název databáze	IntAct
Webová stránka	https://www.ebi.ac.uk/intact/
Zdroje dat	kurátorovaná literatura, (MBINFO), *AgBase, HPIDB3.0, IMEx, Rappsilber Laboratory*
Typy interakcí	protein-protein (fyzické), protein-DNA, protein-RNA
Species coverage	prokaryota, houby, rostliny, bezobratlí a obratlovci
Počet organismů	3671
Počet interakcí	1 572 071 (všech binárních), 1 175 906 PPI (k 26.6.2024)
Počet proteinů	124 275
Způsob kurace	manuální
Validační metoda	MI skóre
Frekv. aktualizace	nepravidelná, poslední verze: 23.5.2024

Tabulka 3.4: Vlastnosti databáze IntAct

MINT

Databáze MINT se specializuje na ukládání informací o PPI, přičemž se obzvlášť zaměřuje na experimentálně ověřené fyzické interakce. Tato databáze udržuje vysokou kvalitu dat s důrazem na experiméntálně ověřené interakce, s přesným dokumentováním zdrojů a metod, proto zde nejsou zahrnuty žádné interakce, které byly předpovězeny výpočetními metodami.(37) Pro zvýšení efektivity kurace a optimalizaci využití zdrojů došlo k fúzi MINT s databází IntAct. MINT je nyní plně integrován do infrastruktury IntAct a jeho data jsou dostupná přímo na

platformě IntAct. Obě tyto databáze podporují iniciativu IMEX a její standardy, což zlepšuje sdílení a integraci dat napříč výzkumnými platformami.(68)

Název databáze	MINT (The Molecular Interaction Database)
Webová stránka	https://mint.bio.uniroma2.it/
Zdroje dat	kurátorovaná literatura, IntAct, *IMEx*
Typy interakcí	protein-protein (fyzické)
Species coverage	prokaryota, houby, rostliny, bezobratlí a obratlovci
Počet organismů	674
Počet interakcí	139 457
Počet proteinů	27 756
Způsob kurace	manuální
Validační metoda	MI skóre
Frek. aktualizace	NA

Tabulka 3.5: Vlastnosti databáze MINT

3.1.2 Integrované databáze

APID

APID (Agile Protein Interaction Data Analyzer) shromažďuje data o PPI z pěti primárních PPI databází a také informace o jejich 3D struktuře z PBD. Všechny proteiny mapuje na identifikátory UniProtKB(28). Zaměřuje se výhradně na experimentálně ověřené PPI. Interakce jsou kategorizovány podle počtu experimentů, metod a publikací, které danou interakci ověřují. Interakce musí být podpořeny aspoň dvěma experimenty nebo publikacemi. Navíc poskytuje interaktomy, které jsou mapovány na jejich proteomy.(69)

Název databáze	APID (Agile Protein Interaction Data Analyzer)
Webová stránka	http://apid.dep.usal.es:8080/APID/
Zdroje dat	IntAct, (HPRD), BioGRID, DIP, BioPlex,*UniProt, PBD, PBDsum*
Typy interakcí	protein-protein (fyzické)
Species coverage	prokaryota, houby, rostliny, bezobratlí a obratlovci
Počet organismů	více než 1 000
Počet interakcí	NA (v roce 2016: 678 441)
Počet proteinů	NA (v roce 2016: 90 379)
Způsob kurace	manuální a automatický
Validační metoda	skórování na základě počtu
Frekv. aktualizace	3 měsíce, poslední verze: březen 2023

Tabulka 3.6: Vlastnosti databáze APID

HINT

Databáze HINT je pečlivá kurátorská kompilace vysoce kvalitních PPI, čerpaných z hlavních osmi zdrojů interactomů zmíněných níže. Interakce jsou systematicky a manuálně filtrovány s cílem odstranit interakce nízké kvality a chybné záznamy a aktualizace probíhá každou noc. HINT přistupuje k výběru dat s vysokou přísností, zaměřuje se výhradně na interakce nejvyšší kvality, což zahrnuje vysokoprůtokové experimenty a interakce z maloměřítkových studií, které byly ve

vědecké literatuře zmíněny minimálně dvakrát. Navíc jsou do databáze zahrnuty experimenty, které prošly ortogonálním testováním. Experimenty bez jakéhokoliv ověření nejsou akceptovány. Tento striktní výběrový proces zajišťuje, že jsou všechny interakce experimentálně ověřené a omezené pouze na fyzické interakce. (70)

Název databáze	HINT (High-quality Interactomes)
Webová stránka	http://hint.yulab.org/
Zdroje dat	BioGRID, MINT, (iRefWeb), DIP, IntAct, (HPRD), (MIPS), *PBD*
Typy interakcí	protein-protein (fyzické)
Species coverage	prokaryota, houby, rostliny, bezobratlí a obratlovci
Počet organismů	10
Počet interakcí	264 054
Počet proteinů	-
Způsob kurace	manuální
Validační metoda	ortogonální testování, interakce zmíněné aspoň ve dvou různých vědeckých studií
Frekv. aktualizace	nepravidelná, poslední verze: červen 2024

Tabulka 3.7: Vlastnosti databáze HINT

mentha

Databáze mentha poskytuje experimentálně určených PPI a čerpá z dat, která byla manuálně kurátorována z databází PPI dodržující standardy konsorcia IMEx a kontrolována slovníky PSI-MI. Zaměřuje se pouze na fyzické interakce a vylučuje interakce, které byly odvozené, čímž zajišťuje přesnost dat. Každý záznam obsahuje odkazy na původní články a zachovává anotace, čímž uživatelům umožňuje hlubší pochopení kontextu interakcí. Z dat o PPI vytváří interaktomy pro různé organismy, která jsou týdně aktualizovány.(71) Interakce jsou hodnoceny skórem spolehlivosti na základě typu interakce a použité metody pro detekci PPI a to podle metodologie MINT skóre, které je popsáno publikaci MINT databáze z roku 2012(37).

Název databáze	mentha
Webová stránka	https://mentha.uniroma2.it/
Zdroje dat	MINT, IntAct, DIP, *MatrixDB*, BioGRID
Typy interakcí	protein-protein
Species coverage	prokaryota, houby, rostliny, bezobratlí a obratlovci
Počet organismů	1 044
Počet interakcí	741 337
Počet proteinů	90 905
Způsob kurace	manuální
Validační metoda	MINT skóre
Frekv. aktualizace	týdně

Tabulka 3.8: Vlastnosti databáze mentha

PICKLE

PICKLE (Protein InteraCtion KnowLedgebasE) je metadatabáze, která obsahuje přímé PPI z proteomů člověka a myši a syntetizuje data z veřejně dostupných PPI databází. Jako referenční soubor pro integraci PPI dat využívá recenzovaný kompletní proteom člověka z UniProt/Swiss-Prot (RHCP - reviewed human complete proteome). Taktéž využívá recenzovaný kompletní proteom myši definovaný UniProt/Swiss-Prot jako referenční proteinovou sadu. PICKLE mimo interakce mezi lidmi (PPI(h-h)) a mezi myšmi (PPI(m-m)) zahrnuje i interakce mezi myšími a lidskými genetickými entitami navzájem (PPI(m-h)). Informace o PPI těží z databází jako jsou IntAct, BioGRID a DIP.(72)

Název databáze	PICKLE (Protein InteraCtion KnowLedgebasE)
Webová stránka	http://www.pickle.gr/
Zdroje dat	PPI databáze: BioGRID, IntAct, (HPRD), DIP biologické databáze: *UniProt, GenBank, Ensembl, European Nucleotide Archive, Gene Ontology Consortium, HUPO Proteomics Standards Initiative, Mouse Genome Informatics*
Typy interakcí	protein-protein (fyzické, ortologní)
Species coverage	Homo sapiens, Mus musculus
Počet organismů	2
Počet interakcí	PPI(h-h) : 218 025 PPI(m-m) : 13 140 PPI(m-h) : 6 211
Počet proteinů	h : 16 420 m: 5 666 m-h: 5 068
Způsob kurace	manuální a automatická
Validační metoda	křížová validace mezi jinými PPI databázemi
Frekv. aktualizace	nepravidelná, poslední verze: PICKLE 3.3, 1.10.2021

Tabulka 3.9: Vlastnosti databáze PICKLE

3.1.3 Predikční databáze

IID

Databáze IID poskytuje kontextově specifické sítě pro 18 druhů, zahrnující člověka, modelové organismy a domestikované druhy. Tato databáze se vyznačuje tím, že anotuje PPI s důrazem na podmínky, ve kterých interakce probíhají (např. vývojové stádium, tkáň, ...) a zahrnuje údaje o konzervaci napříč druhy, směrovosti interakce, trvání či přítomnosti ve větších proteinových komplexech. IID integruje experimentálně detekované PPI z deseti databází a mimo to zahrnuje i ortologní PPI a vysoce spolehlivé PPI předpovězené podle nejmodernějších výpočetních metod (jak je uvedeno např. v Rhodes (Nat. Biotech., 2005)(73), Elefsinioti (Mol. Cell Proteomics, 2011)(74), Zhang (Nature, 2012)(75) a Kotlyar (Nat. Methods, 2015)(76)). (77)

Název databáze	IID (Integrated Interactions Database)
Webová stránka	http://iid.ophid.utoronto.ca/
Zdroje dat	(BCI), (BIND), BioGRID, DIP, (HPRD), *InnateDB, MatrixDB*, MINT
Typy interakcí	protein-protein
Species coverage	prokaryota, houby, rostliny, bezobratlí a obratlovci
Počet organismů	18
Počet interakcí	7 369 019
Počet proteinů	NA
Způsob kurace	manuální a automatická
Validační metoda	NA
Frekv. aktualizace	nepravidelná, poslední verze: květen 2021

Tabulka 3.10: Vlastnosti databáze IID

STRING

STRING databáze (Search Tool for Retrieval of Interacting Genes/Proteins) je jedna z nejpoužívanějších a nejrozsáhlejší databází protein-proteinových interakcí. Jde o predikční databázi, která se zabývá jak fyzickými (přímými) interakcemi, tak také funkčními (nepřímými) asociacemi. STRING definuje fyzickou interakci jako přímý kontakt mezi dvěma proteiny, typicky identifikované experimentálně, a definuje funkční asociaci jako vztah dvou neidentických proteinů, kde každý pochází z jiného protein-kódujícího lokusu. Všechny tyto interakce jsou odvozeny z dat získaných hned několika způsoby a to jsou(78):

- experimentální data
- predikce podle genomického kontextu
- text mining vědecké literatury
- predikce podle ko-exprese anebo také z již známých komplexů a dráh, které jsou k dispozici v kurátorsky moderovaných zdrojích

Dále jsou interakce pečlivě hodnoceny, je jim přiřazené skóre věrohodnosti a poté jsou informace o interakcích automaticky přeneseny na méně prozkoumané organismy. K vyhodnocení každé interakce přispívá sedm typů důkazů, neboli důkazových kanálů („evidence channels"), které jsou na sobě nezávislé. Tři kanály jsou založeny na genomickém kontextu:

(a) kanál sousedství (neighbourhood channel): sleduje, jak blízko jsou geny na chromozomu

(b) fúzní kanál (fusion channel): zkoumá evoluční historii proteinů a sleduje, zda jsou dvě proteinové domény v některých organismech spojeny do jednoho proteinu, což naznačuje jejich spolupráci

(c) kanál společného výskytu (co-occurrence channel): se zaměřuje na přítomnost genů napříč různými organismy, což může naznačovat jejich společnou funkci

Další kanály zahrnují společnou expresi (co-expression channel) a experimentální data (experiments channel), která vycházejí z genomických a laboratorních

experimentů (data jsou importována ze zdrojů jako IMEx Consortium anebo BioGRID). Kanál znalostí (database channel) využívá manuálně schválené databáze (například KEGG a Reactome) a text mining kanál (text-mining channel) analyzuje vědecké články, což funguje základě hlubokého učení, kde je použit biomedický jazykový model RoBERTa(79). (52)

Všechny důkazy o interakcích jsou nejprve podrobeny procesu benchmarkingu, což znamená, že jsou porovnávány vyhodnoceny vzhledem k předem stanoveným standardům nebo referenčním datům. STRING přenáší asociace mezi organismy pomocí COG a Protein-mode metod. Účel tohoto kroku je rozšířit známé a již predikované PPI z modelového organismu na odpovídající páry v jiných organismech a také potvrdit, že daná asociace či interakce je aplikovatelná na cílový organismus. Skóre z jednotlivých důkazů jsou integrována do kombinovaného skóre pomocí naivního Bayesova klasifikátoru. Uživatelé mohou vidět skóre jednotlivých kanálů u každé interakce a pochopit tak výslednou predikci. Podrobněji o výpočtu skóre lze najít v publikaci Von Mering a kol., 2005(80).

Název databáze	STRING (Search Tool for Retrieval of Interacting Genes/Proteins)
Webová stránka	https://string-db.org/
Zdroje dat	experimentální a chemická data: DIP, BioGRID, (HPRD), IntAct, MINT, *PDB* kurátorsky moderovaná data: *BioCarta, BioCyc, Gene Ontology, KEGG, Reactome*
Typy interakcí	protein-protein (fyzické a funkční)
Species coverage	prokaryota, houby, rostliny, bezobratlí a obratlovci
Počet organismů	12 535
Počet interakcí	víc než 27 miliard z toho přes 332 milionů se skórem >= 0,9
Počet proteinů	59 309 604
Způsob kurace	kombinace manuální a automatizované kurace
Validační metoda	benchmarking
Frekv. aktualizace	čtvrtletní

Tabulka 3.11: Vlastnosti databáze STRING

3.1.4 Neaktivní databáze

Mnoho databází ve svých zdrojích uvádí databáze, které již nejsou aktivní. Avšak v době svého fungování byly významný zdrojem pro dnešní fungující databáze a pro vědeckou komunitu obecně, jelikož jsou ve vědeckých publikacích často citovány. Jsou to například datbáze v tabulce 3.12.

Již neaktivní PPI databáze jsou samy o sobě zajímavým tématem. Bylo by užitečné zjistit, kam se jejich data poděla, zda byla přelita do jiné aktivní databáze, nebo zda zanikla. Například u MiMI se zdá, že data vplynula do NCIBI[1](81). Tuto otázku systematicky neřeším, ale považuji ji za hodnou dalšího zkoumání.

[1]NCIBI, "National Center for Integrative Biomedical Informatics,"http://www.ncibi.org

Název	**Klíčové vlastnosti**
Primární databáze	
BIND	databáze molekulárních interakcí (mezi proteiny, DNA, RNA, sacharidy a dalšími malými molekulami), komplexů a drah(82), bývalý člen IMEx konsorcia(34)
HPRD	databáze PPI u člověka, zahrnuje informace důležité pro funkci lidských proteinů ve zdraví a nemoci(83)
Integrované databáze	
MiMI	databáze pokytující integrovaná data o molekulárních interakcích(84)

Tabulka 3.12: Příklady neaktivních databáze

3.2 Porovnání databází

Tabulka 3.13 shrnuje klíčové znaky vybraných databází. Jako PPI zdroje jsou vypsány pouze PPI databáze. Pokud je druhové pokrytí označeno jako „široké", znamená to, že obsahuje aspoň jednoho člena ze těchto pěti skupin: prokaryota, houby, rostliny, bezobratlí a obratlovci. Dále jsou zde shrnuty i informace o počtu proteinů a interakcí. Zahrnuty byly i znaky o jejich datasetech, jelikož pro některé z následujících analýz bylo zapotřebí je stáhnout, tudíž je nutné vědět s jakou verzí bylo zacházeno. STRING databáze nabízí ke stažení celý svůj dataset, ale kvůli jeho značné velikosti to pro tuto práci nebylo proveditelné.

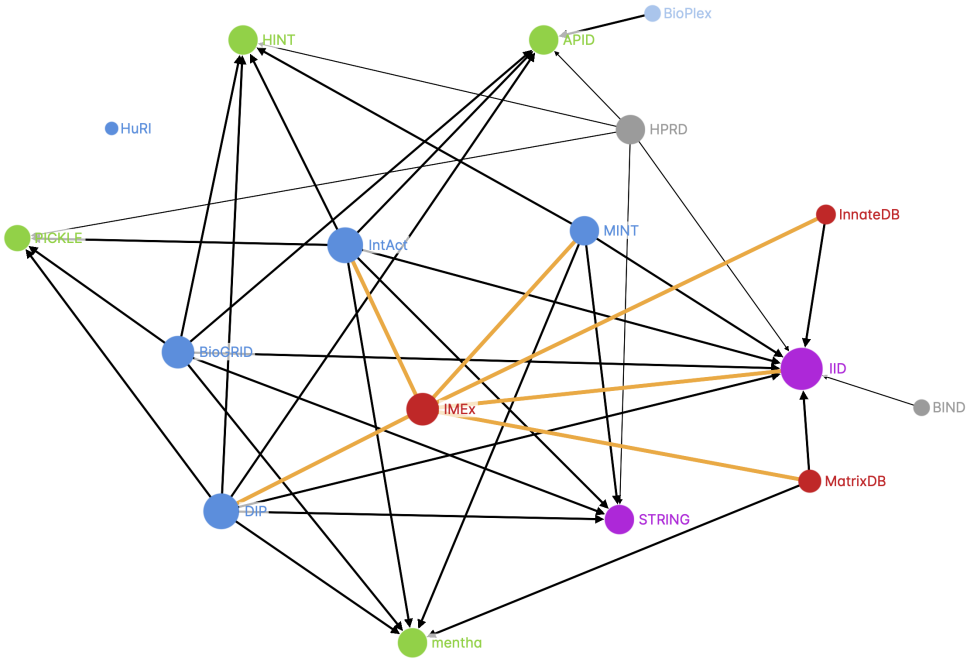
Tabulka 3.13: Souhrn znaků jednotlivých databází.

Databáze	**PPI zdroje**	**Typ interakcí**	**Druhové pokrytí**	**Počet proteinů**	**Počet interakcí**	**Možnost stažení**	**MITAB kompatibilita**	**Stažená verze**
Primární databáze								
BioGRID	kurátorovaná literatura	PPI + další	široké (83)	89 642	>2,7 milionů*	ano	ano	Release 4.4.235 (červenec 2024)
DIP	kurátorovaná literatura	PPI	široké (834)	28 850	81 923	ano	ne***	5.2.2017
HuRI	kurátorovaná literatura, experimentální metody	PPI	Homo sapiens	9 094	64 006	ano	ne	HI-union (2020)
IntAct	kurátorovaná literatura	PPI + další	široké (3671)	124 275	844 973*	ano	ano	23.5.2024
MINT	kurátorovaná literatura	PPI	široké (674)	27 756	139 457	ano	ano	15.7.2024
Integrované databáze								
APID	IntAct, (HPRD), BioGRID, DIP, BioPlex	PPI	široké (>1000)	NA	NA	ano**	ano	březen 2021
HINT	BioGRID, MINT, DIP, IntAct, (HPRD)	PPI	široké (10)	NA	264 054	ano	ne	červen 2024
mentha	MINT, IntAct, DIP, MatrixDB, BioGRID	PPI	široké (1044)	90 905	741 337	ano	ne	15.7.2024
PICKLE	BioGRID, IntAct, (HPRD), DIP	PPI	Homo sapiens, Mus musculus	22 086	237 376	ano**	ne	PICKLE 3.3 (říjen 2021)
Predikční databáze								
IID	(BIND), BioGRID, DIP, (HPRD), InnateDB, IntAct, MatrixDB, MINT	PPI	široké (18)	NA	7 369 019	ano**	ne	květen 2021
STRING	DIP, BioGRID, (HPRD), IntAct, MINT	PPI + další	široké (12535)	59 309 604	>27 miliard*	ano	ne	NA

*Vysvětlivky: *zahrnuje i jiné než PPI interakce, **pouze po jednotlivých organismech, *** kompatibilní pouze datasety pár jednotlivých organismů*

3.2.1 Toky dat mezi databázemi

Z grafu 3.1 lze vidět, že z vybraných databází je DIP nejčastějším poskytovatelem dat. PPI databáze, která data získává z nejvíc zdrojů je IID i po odečtení nefunkčních databází. IntAct a BioGRID poskytují data stejným databázím, jediný rozdíl je, že IntAct je aktivním členem IMEx konsorcia narozdíl od BioGRIDu. Databáze HuRI svá data nikomu neposkytuje a sama je od nikoho nečerpá. MINT je hned po HuRI druhým nejmenším poskytovatelem svých dat. Z nefunkčních databází byla význačná HPRD databáze, která svá data sdílela jiným předním databázím.



Obrázek 3.1: Graf reprezentující datový tok mezi vybranými databázemi. *Modře jsou označeny primární databáze, zeleně integrované, fialově predikční a šedě nefunkční. Červeně jsou označeny zdroje dat, které nejsou databáze PPI. Oranžové hrany vyznačují aktivní členy IMEx konsorcia.*

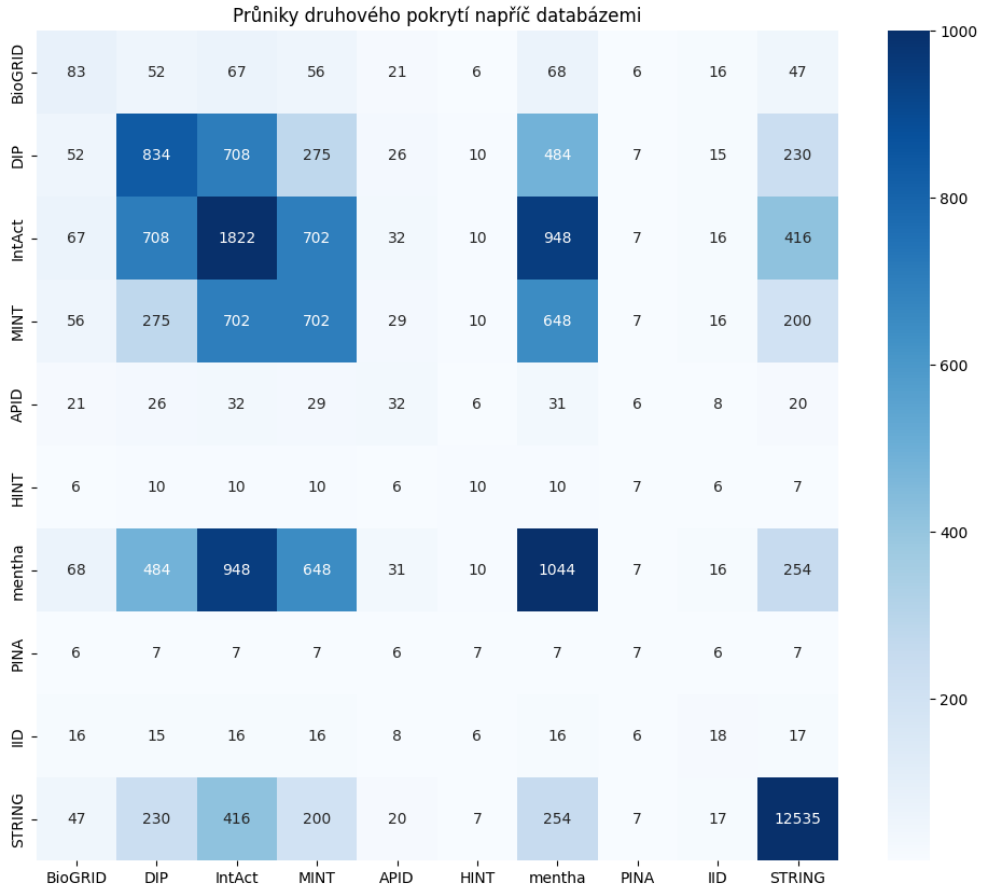
3.2.2 Překryv druhového pokrytí

Pro tuto analýzu byly vyřazeny databáze HuRI a PICKLE, neboť jejich druhové pokrytí je velmi malé a jejich řádky a sloupce by ukazovaly stejná čísla (Homo sapiens a Mus musculus jsou ve všech zmíněných databázích). Byly staženy jednotlivé datasety interakcí z každé databáze až na STRING, kde se pracovalo pouze se seznamem organismů. U databáze APID nebyl stažen kompletní dataset, protože není k dispozici. Místo toho byla stažena pouze podmnožina organismů, které mají zaznamenáno více než 500 interakcí, nicméně udávají, že databáze obsahuje přes 1000 organismů. IID dataset neobsahuje taxidy (taxonomické identifikátory) organismů, proto byly vyvozeny z názvu objevujících na jejich stránkách. Ostatně ze zbylých datasetů byl pro komparativní analýzu extrahován seznam interaktujících taxidů.

Podle vygenerované heat mapy na obrázku 3.2 je druhové pokrytí mezi primárními databázemi poměrně vysoké (kromě BioGRIDu). Z integrovaných databází je velmi význačná mentha, která s každou z databází má značný překryv. Obecně má největší překryv IntAct a za ním mentha. Přestože STRING má výrazně vysoký celkový počet organismů, jeho průniky s ostatními databázemi nejsou až tak velké.

U databáze HINT s menším druhovým pokrytím se často vyskytuje stoprocentní překryv. Co se týče ostatních databází, stoprocentní průnik je pouze mezi databázemi MINT a IntAct, protože celý MINT dataset je v IntActu integrovaný.

U IntActu si lze všimnout, že jeho celkový počet organismů je podstatně menší než uvádí ve svých statistikách. Je možné, že některé varianty a poddruhy jednoho organismu v datasetu klasifikuje jakožto stejný organismus, ale počítá zvlášť ve svých statistikách, nicméně přesný důvod není znám.



Obrázek 3.2: Heatmapa zachycující průniky druhového pokrytí mezi databázemi. *Tmavší pole označují vyšší překryv organismů. Na diagonále je celkový počet organismů v databázi.*

3.2.3 Porovnání výsledků hledání pro konkrétní PPI

Porovnání anotací PPI v primárních databázích pro konkrétní dvojici proteinů

V Tabulce 3.14 byly vybrány konkrétní PPI a následně hledány v primárních databázích. Cílem bylo zjistit, jak se mezi nimi mohou lišit metody detekce a také kurátorovaná literatura, ze kterých byly PPI odvozeny. PPI byly vybrány libovolně, jde o proteiny interagující v organismu Homo Sapiens. Význačná je databáze BioGRID, která má velký počet kurátorsky zpracovaných článku detekujících PPI a také unikátní metody detekce. Tato tabulka taktéž ilustruje jak různé databáze pojmenovávají metody detekce, příklad DIP některé metody značí jako jednoduše „experimentální" nebo IntAct označuje „coip" jako coimmunoprecipitation. Lze vidět, že každá databáze s publikacemi zachází trochu jinak a anotuje různé články nezávisle na ostatních.

Tabulka 3.14: Srovnání kurátorovaných výsledků pro konkrétní PPI mezi primárními databázemi

Interagující proteiny		**BioGRID**		**DIP**		**HuRI**		**MINT**		**IntAct**	
		Metoda detekce	PubMed ID	Metoda detekce	PubMed ID	Metoda detekce	PubMed ID	Metoda detekce	PubMed ID	Metoda detekce	PubMed ID
Q13158 (FADD)	Q15628 (TRADD)	affinity capture-MS, Y2H	30561431, 21724995, 12911633, 8565075, 12796506, 24686082, 16919273, 32296183, 12107169	experimentální, anti tag coimmunoprecipitation	8565075, 23955153	3 varianty Y2H, 9krát opakováno	32296183	two hybrid, coimmunoprecipitation	8565075	two hybrid, two hybrid prey pooling, two hybrid array	11112409, 32296183
P19438 (TNFRSF1A)	Q15628 (TRADD)	affinity capture-MS, affinity capture-western, two hybrid	28514442, 33961781, 30420664, 18939944, 24758719, 22297296, 32968279, 7758105, *+dalších 20*	experimentální, anti tag coimmunoprecipitation	9129204, 23955153, 19641494	3 varianty Y2H, 9krát opakováno	32296183	pull down, coimmunoprecipitation	8621670	coip, anti bait coip, anti tag coip, pull down, two hybrid	11684708, 16611992, 19524513, 19641494, 19781631, 33961781, 8621670, 8943045, 23955153, 7758105, 8565075
P46109 (CRKL)	P00519 (ABL1)	affinity capture-MS, affinity capture-western, reconstituted complex, two hybrid	33961781, 8978305, 19823681, 9820532, 14604282, 16955467, 18835194, 7926767, 8978305, 8524328, 9710592	protein kinase assay, small scale experiment	22286129	3 varianty Y2H, 9krát opakováno	32296183	anti bait coimmunoprecipitation, pull down	16443220	protein kinase assay, anti bait coip, pull down, anti tag coip	22286129, 16443220, 33961781, 24412932

Porovnání popisů téže interakce v integrovaných databázích

Integrované databáze v Tabulce 3.15 byly dotázány na stejné PPI jako primární databáze v Tabulce 3.14. Cílem byly porovnat jak se databáze liší v tom co a kolik toho od primárních převezmou do svého datasetu. Byl sledován celkový počet důkazů a následně jaký podíl byl převzat z jaké databáze. Integrovaný důkaz zde chápu jako důkaz o interakci převzatý z kurátorsky moderované databáze. Je nutné podotknout, že v potaz byly brány i nepřímé důkazy o interakci. HINT ve zdrojí neuvádí databáze, ale přímo směruje na vědecké publikace. V tomto případě lze vidět, že každá integruje různé množství z různých zdrojů a že se počty důkazů mohou dost lišit.

Tabulka 3.15: Srovnání počtu a zdrojů integrovaných důkazů

Interagující proteiny		**APID**		**HINT**		**mentha**		**PICKLE**	
		Počet integrovaných důkazů	Zdroje	Počet integrovaných důkazů	Zdroje	Počet integrovaných důkazů	Zdroje	Počet integrovaných důkazů	Zdroje
Q13158 (FADD)	Q15628 (TRADD)	15	BioGRID (9), DIP (1), HPRD (1), IntAct (4)	22	-	11	BioGRID (7), DIP (1), IntAct (3)	13	BioGRID (10), DIP (1), HPRD (1), IntAct (5)
P19438 (TNFRSF1A)	Q15628 (TRADD)	42	BioGRID (23), BioPlex (1), DIP (1), HPRD (1), IntAct (16)	53	-	33	BioGRID (20), DIP (2), IntAct (8), MINT (3)	30	BioGRID (24), DIP (2), HPRD (1), IntAct (9), MINT (3), UniProt (1)
P46109 (CRKL)	P00519 (ABL1)	21	BioGRID (13), BioPlex (1), DIP (1), HPRD (2), IntAct (4)	18	-	16	BioGRID (13), IntAct (1), MINT (2)	18	BioGRID (13), DIP (1), HPRD (1), IntAct (1), MINT (2)

Porovnání popisů téže interakce v predikčních databázích

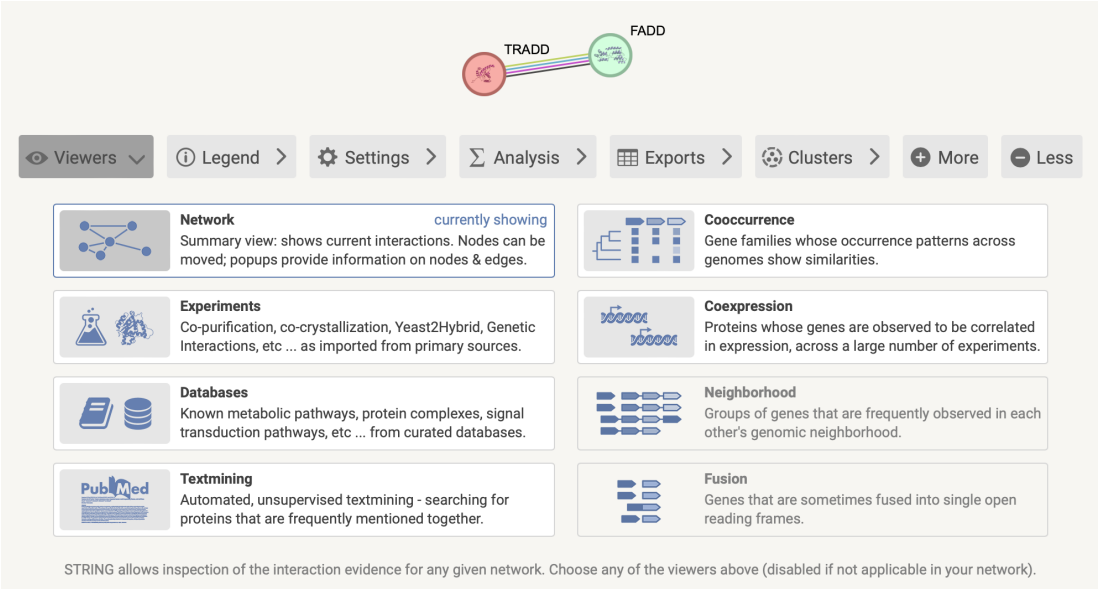
Vybrané predikční databáze jsou od sebe značně odlišné a s důkazy pro PPI zachází každá jinak. Pro ilustraci nebudu v této části porovnávat výsledky všech tří PPI jako v předchozích případech, ale pouze s první PPI mezi proteiny FADD a TRADD.

V databázi IID nalezneme typ důkazů podporující PPI - experimentální, predikční nebo ortologický. Následující Obrázek 3.3 ukazuje výsledek hledání. Interakci mezi FADD a TRADD podporují typy experimentální a predikční a stejně jako u integrovaných databází nalezneme metody detekce interakce, publikace ze kterých vychází (celkově 15 publikací) a následně i zdrojové databáze, od kterých důkazy převzala. Ve zdrojových databází uvádí i sebe, což naznačuje, že svojí predikcí přispívá do důkazů interakce. Nicméně uživatel nenalezne více informací o jednotlivých důkazech, např. co bylo převzato z jaké databáze nebo popis predikce interakce.

UniProt1	UniProt2	symbol1	symbol2	evidence type	methods	PMIDs	DBs
Q13158	Q15628	FADD	TRADD	exp|pred	affinity chromatography technology|anti tag coimmunoprecipitation|experimental interaction detection|two hybrid|two hybrid array|two hybrid prey pooling approach|validated two hybrid|virotrap	10911999|11112409|11464215|12107169|12796506|12911633|16919273|21183682|21724995|23955153|24686082|30561431|32296183|8565075|8681376	bci|biogrid|dip|hprd|iid|innatedb|intact

Obrázek 3.3: Výsledek hledání FADD-TRADD interakce v databází IID.

STRING poskytuje pro stejnou interakci mnohem detailnější informace. Jak bylo dříve zmíněno, nabízí 7 důkazových kanálu a uživatel se může každý zvlášť prozkoumat a zjistit podrobnosti o interakci v daných oblastech. Co se týče experimentálního kanálu, který integruje experimentální důkazy z jiných databází, tak databáze uvádí 15 publikací i se zdrojovými databázemi.(85)



Obrázek 3.4: FADD-TRADD interakce a její dostupné důkazové kanály ve STRINGu.

Diskuze

Tato práce se zaměřuje na jedenácti vybraných databází protein-proteinových interakcí. Byly popsány základní pojmy a aspekty takových databází. Dále byly vybrány konkrétní PPI databáze na základě subjektivních kritérií a byly popsány jejich vybrané charakteristiky. Cílem bylo zjistit, jak každá databáze sbírá svá data, jaký typ dat poskytuje, jaké organismy pokrývá a jiné relevantní údaje, zejména nakolik a jakým způsobem jsou databáze na sobě vzájemně obsahově závislé.

Při tvorbě této práce jsem narazila i na databáze, které již nejsou aktivní. Mnohé z těchto databází byly v minulosti klíčovými zdroji informací, ale jejich obsah již není dostupný. Zpětné dohledání jejich obsahu je velmi obtížné.

Vybrané databáze se mezi sebou významně liší ve všech hodnocených parametrech. Například počet proteinů a interakcí se značně liší, jelikož některé databáze, jako například BioGRID, se zaměřují na detailnější a pečlivě ověřené informace, ale zato v menším objemu, zatímco jiné, jako je STRING obsahují rozsáhlé množství dat a pokrývají velké množství interakcí. Každá databáze má různou míru druhového pokrytí, což vyplývá z jejich zaměření a cílů. Některé se specializují pouze na modelové organismy, což umožňuje poskytovat kvalitní a detailní informace o konkrétních organismech, zatímco jiné databáze se mohou snažit pokrýt co nejširší spektrum a to někdy na úkor kvality dat. Způsoby, jakými jednotlivé databáze moderují a validují svá data, se taktéž liší.

Z výsledků se zdá, že mezi obashově nejbohatšími a současně nejaktuálnějšími primární databázemi patří BioGRID a IntAct a to díky vysokému počtu proteinů a interakcí, širšímu druhovému pokrytí, časté aktualizaci a integraci jejich záznamů do jiných databází. DIP již delší dobu nebyl aktualizován a MINT je plně integrován do databáze IntAct. HuRI je omezen čistě na Homo sapiens, ale zato zaručuje kvalitu jejich dat.

U vybraných integrovaných databází je těžší vyvodit nějaký závěr, každá má své přednosti a každá má jiné zaměření a cíle. Nejaktuálnější jsou určitě HINT a mentha. Co se týče predikčních databází, tak také se navzájem dost liší a to ve všech kritériích. IID je zaměřen na užší pokrytí organismů, zatímco u STRINGu se zdá, že snaží pokrýt co se dá a od toho se odvíjí další rozdíly v hodnocených kritériích (např. počty proteinů). Samozřejmě jejich výpočetní metody pro predikci PPI se také liší.

Součástí práce je též grafická mapa toku dat mezi vybranými databázemi, kde také bylo vyobrazeno i IMEx konsorcium a jeho aktivní členové. Následně se práce zabývala průniky druhového pokrytí napříč databázemi, kde byla podtržena integrace dat z MINT do IntActu a kde také bylo ukázáno jak se mezi sebou databáze v tomto aspektu značně liší.

V poslední části byly pro několik vybraných PPI porovnány metody detekce a počty literátních zdrojů které jednotlivé primární databáze evidují ke konkrétní PPI. Dále se stejné PPI hledaly v integrovaných databází a byl sledován počet důkazů převzatých z jiných databází a z jakých konkrétních databází to bylo. Za důležité lze považovat zjištění, že ani integrované databáze využívající tytéž primární zdroje neodkazují vždy ke stejnému souboru primárních pozorování. Nakonec byly ilustrovány rozdíly výsledků hledání mezi vybranými predikčními

databázemi.

V souhrnu lze uživatelům doporučit, aby se při vyhledávání dat o PPI neomezovali na jedinou databázi. Tato práce by mohla posloužit jako vodítko pro orientaci nových uživatelů ve složitém světě databází protein-proteinových interakcí.

Seznam použité literatury

[1] N. Safari-Alighiarloo, M. Taghizadeh, M. Rezaei-Tavirani, B. Goliaei, and A. A. Peyvandi, "Protein-protein interaction networks (ppi) and complex diseases," *Gastroenterology and Hepatology from bed to bench*, vol. 7, no. 1, p. 17, 2014.

[2] A. Bauch and G. Superti-Furga, "Charting protein complexes, signaling pathways, and networks in the immune system," *Immunological reviews*, vol. 210, no. 1, pp. 187–207, 2006.

[3] I. M. Nooren and J. M. Thornton, "Diversity of protein–protein interactions," *The EMBO journal*, vol. 22, no. 14, pp. 3486–3492, 2003.

[4] P. Li, L. Wang, and L.-j. Di, "Applications of protein fragment complementation assays for analyzing biomolecular interactions and biochemical networks in living cells," *Journal of proteome research*, vol. 18, no. 8, pp. 2987–2998, 2019.

[5] O. Keskin, N. Tuncbag, and A. Gursoy, "Predicting protein–protein interactions from the molecular to the proteome level," *Chemical reviews*, vol. 116, no. 8, pp. 4884–4909, 2016.

[6] K. Kuroda, M. Kato, J. Mima, and M. Ueda, "Systems for the detection and analysis of protein–protein interactions," *Applied microbiology and biotechnology*, vol. 71, pp. 127–136, 2006.

[7] C. Von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork, "Comparative assessment of large-scale data sets of protein–protein interactions," *Nature*, vol. 417, no. 6887, pp. 399–403, 2002.

[8] W. Bian, H. Jiang, S. Feng, J. Chen, W. Wang, and X. Li, "Protocol for establishing a protein-protein interaction network using tandem affinity purification followed by mass spectrometry in mammalian cells," *STAR protocols*, vol. 3, no. 3, p. 101569, 2022.

[9] M. Zhou, Q. Li, and R. Wang, "Current experimental methods for characterizing protein–protein interactions," *ChemMedChem*, vol. 11, no. 8, pp. 738–756, 2016.

[10] K. Pichlerova and J. Hanes, "Technologies for the identification and validation of protein-protein interactions," *Gen. Physiol. Biophys*, vol. 40, pp. 495–522, 2021.

[11] H. C. Reinhardt, H. Jiang, M. T. Hemann, and M. B. Yaffe, "Exploiting synthetic lethal interactions for targeted cancer therapy," *Cell cycle*, vol. 8, no. 19, pp. 3112–3119, 2009.

[12] Y.-Y. Wang, W. Li, B.-C. Ye, and X.-B. Bi, "Chemical and biological strategies for profiling protein-protein interactions in living cells," *Chemistry–An Asian Journal*, vol. 18, no. 14, p. e202300226, 2023.

[13] Y. Sun, N. M. Hays, A. Periasamy, M. W. Davidson, and R. N. Day, "Monitoring protein interactions in living cells with fluorescence lifetime imaging microscopy," *Methods in enzymology*, vol. 504, pp. 371–391, 2012.

[14] M. Morell, S. Ventura, and F. X. Avilés, "Protein complementation assays: approaches for the in vivo analysis of protein interactions," *FEBS letters*, vol. 583, no. 11, pp. 1684–1691, 2009.

[15] W. Qin, K. F. Cho, P. E. Cavanagh, and A. Y. Ting, "Deciphering molecular interactions by proximity labeling," *Nature methods*, vol. 18, no. 2, pp. 133–143, 2021.

[16] L. Skrabanek, H. K. Saini, G. D. Bader, and A. J. Enright, "Computational prediction of protein–protein interactions," *Molecular biotechnology*, vol. 38, pp. 1–17, 2008.

[17] L. Hu, X. Wang, Y.-A. Huang, P. Hu, and Z.-H. You, "A survey on computational models for predicting protein–protein interactions," *Briefings in bioinformatics*, vol. 22, no. 5, p. bbab036, 2021.

[18] K. Raja, J. Natarajan, F. Kuusisto, J. Steill, I. Ross, J. Thomson, and R. Stewart, "Automated extraction and visualization of protein–protein interaction networks and beyond: A text-mining protocol," *Protein-Protein Interaction Networks: Methods and Protocols*, pp. 13–34, 2020.

[19] S. Lalonde, D. W. Ehrhardt, D. Loqué, J. Chen, S. Y. Rhee, and W. B. Frommer, "Molecular and cellular approaches for the detection of protein–protein interactions: latest techniques and current limitations," *The Plant Journal*, vol. 53, no. 4, pp. 610–635, 2008.

[20] Z. Shaukat, S. Aiman, C.-H. Li *et al.*, "Protein-protein interactions: Methods, databases, and applications in virus-host study," *World Journal of Virology*, vol. 10, no. 6, p. 288, 2021.

[21] H. N. Chua, W.-K. Sung, and L. Wong, "Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions," *Bioinformatics*, vol. 22, no. 13, pp. 1623–1630, 2006.

[22] A. Kuzniar, R. C. van Ham, S. Pongor, and J. A. Leunissen, "The quest for orthologs: finding the corresponding gene across genomes," *Trends in Genetics*, vol. 24, no. 11, pp. 539–551, 2008.

[23] E. V. Koonin, "Orthologs, paralogs, and evolutionary genomics," *Annu. Rev. Genet.*, vol. 39, no. 1, pp. 309–338, 2005.

[24] B. Lehne and T. Schlitt, "Protein-protein interaction databases: keeping up with growing interactomes," *Human genomics*, vol. 3, pp. 1–7, 2009.

[25] N. Safari-Alighiarloo, M. Taghizadeh, and M. Rezaei tavirani, "Protein-protein interaction databases: An overall view on interactome organization the nature of protein-protein interactions data," *International journal of analytical, pharmaceutical and biomedical sciences*, vol. 4, pp. 15–23, 01 2015.

[26] M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.

[27] M. Milacic, D. Beavers, P. Conley, C. Gong, M. Gillespie, J. Griss, R. Haw, B. Jassal, L. Matthews, B. May *et al.*, "The reactome pathway knowledgebase 2024," *Nucleic acids research*, vol. 52, no. D1, pp. D672–D678, 2024.

[28] "Uniprot: the universal protein knowledgebase in 2023," *Nucleic acids research*, vol. 51, no. D1, pp. D523–D531, 2023.

[29] S. A. Aleksander, J. Balhoff, S. Carbon, J. M. Cherry, H. J. Drabkin, D. Ebert, M. Feuermann, P. Gaudet, N. L. Harris *et al.*, "The gene ontology knowledgebase in 2023," *Genetics*, vol. 224, no. 1, p. iyad031, 2023.

[30] J. Amberger, C. Bocchini, and A. Hamosh, "A new face and new challenges for online mendelian inheritance in man (omim®)," *Human mutation*, vol. 32, no. 5, pp. 564–567, 2011.

[31] M. J. Landrum, J. M. Lee, G. R. Riley, W. Jang, W. S. Rubinstein, D. M. Church, and D. R. Maglott, "Clinvar: public archive of relationships among sequence variation and human phenotype," *Nucleic acids research*, vol. 42, no. D1, pp. D980–D985, 2014.

[32] "Protein data bank: the single global archive for 3d macromolecular structure data," *Nucleic acids research*, vol. 47, no. D1, pp. D520–D528, 2019.

[33] P. Porras, E. Barrera, A. Bridge, N. Del-Toro, G. Cesareni, M. Duesbury, H. Hermjakob, M. Iannuccelli, I. Jurisica, M. Kotlyar *et al.*, "Towards a unified open access dataset of molecular interactions," *Nature communications*, vol. 11, no. 1, p. 6144, 2020.

[34] S. Orchard, S. Kerrien, S. Abbani, B. Aranda, J. Bhate, S. Bidwell, A. Bridge, L. Briganti, F. S. Brinkman, G. Cesareni *et al.*, "Protein interaction data curation: the international molecular exchange (imex) consortium," *Nature methods*, vol. 9, no. 4, pp. 345–350, 2012.

[35] I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S.-M. Kim, and D. Eisenberg, "Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions," *Nucleic acids research*, vol. 30, no. 1, pp. 303–305, 2002.

[36] N. Del Toro, A. Shrivastava, E. Ragueneau, B. Meldal, C. Combe, E. Barrera, L. Perfetto, K. How, P. Ratan, G. Shirodkar *et al.*, "The intact database: efficient access to fine-grained molecular interaction data," *Nucleic acids research*, vol. 50, no. D1, pp. D648–D653, 2022.

[37] L. Licata, L. Briganti, D. Peluso, L. Perfetto, M. Iannuccelli, E. Galeota, F. Sacco, A. Palma, A. P. Nardozza, E. Santonico *et al.*, "Mint, the molecular interaction database: 2012 update," *Nucleic acids research*, vol. 40, no. D1, pp. D857–D861, 2012.

[38] C. Pastrello, M. Kotlyar, and I. Jurisica, "Informed use of protein–protein interaction data: A focus on the integrated interactions database (iid)," *Protein-Protein Interaction Networks: Methods and Protocols*, pp. 125–134, 2020.

[39] O. Clerc, M. Deniaud, S. D. Vallet, A. Naba, A. Rivet, S. Perez, N. Thierry-Mieg, and S. Ricard-Blum, "Matrixdb: integration of new data with a focus on glycosaminoglycan interactions," *Nucleic acids research*, vol. 47, no. D1, pp. D376–D381, 2019.

[40] K. Breuer, A. K. Foroushani, M. R. Laird, C. Chen, A. Sribnaia, R. Lo, G. L. Winsor, R. E. Hancock, F. S. Brinkman, and D. J. Lynn, "Innatedb: systems biology of innate immunity and beyond—recent updates and continuing curation," *Nucleic acids research*, vol. 41, no. D1, pp. D1228–D1233, 2013.

[41] "The sib swiss institute of bioinformatics semantic web of data," *Nucleic Acids Research*, vol. 52, no. D1, pp. D44–D51, 2024.

[42] M. Thakur, A. Buniello, C. Brooksbank, K. T. Gurwitz, M. Hall, M. Hartley, D. G. Hulcoop, A. R. Leach, D. Marques, M. Martin *et al.*, "Embl's european bioinformatics institute (embl-ebi) in 2023," *Nucleic Acids Research*, vol. 52, no. D1, pp. D10–D17, 2024.

[43] D. Zhou and Y. He, "Extracting interactions between proteins from the literature," *Journal of biomedical informatics*, vol. 41, no. 2, pp. 393–407, 2008.

[44] M. Krallinger, M. Vazquez, F. Leitner, D. Salgado, A. Chatr-Aryamontri, A. Winter, L. Perfetto, L. Briganti, L. Licata, M. Iannuccelli *et al.*, "The protein-protein interaction tasks of biocreative iii: classification/ranking of articles and linking bio-ontology concepts to full text," *BMC bioinformatics*, vol. 12, pp. 1–31, 2011.

[45] J.-W. Chang, Y.-Q. Zhou, M. T. Ul Qamar, L.-L. Chen, and Y.-D. Ding, "Prediction of protein–protein interactions by evidence combining methods," *International journal of molecular sciences*, vol. 17, no. 11, p. 1946, 2016.

[46] A. Kamburov, U. Stelzl, and R. Herwig, "Intscore: a web tool for confidence scoring of biological interactions," *Nucleic acids research*, vol. 40, no. W1, pp. W140–W146, 2012.

[47] J. M. Villaveces, R. C. Jimenez, P. Porras, N. del Toro, M. Duesbury, M. Dumousseau, S. Orchard, H. Choi, P. Ping, N. Zong *et al.*, "Merging and scoring molecular interactions utilising existing community standards: tools, use-cases and a case study," *Database*, vol. 2015, p. bau131, 2015.

[48] Columbia University Libraries, "Controlled vocabulary in databases," https://library.cumc.columbia.edu/kb/controlled-vocabulary-databases, navštíveno 1.8.2024.

[49] H. Hermjakob, L. Montecchi-Palazzi, G. Bader, J. Wojcik, L. Salwinski, A. Ceol, S. Moore, S. Orchard, U. Sarkans, C. Von Mering *et al.*, "The hupo psi's molecular interaction format—a community standard for the representation of protein interaction data," *Nature biotechnology*, vol. 22, no. 2, pp. 177–183, 2004.

[50] J. Malone, E. Holloway, T. Adamusiak, M. Kapushesky, J. Zheng, N. Kolesnikov, A. Zhukova, A. Brazma, and H. Parkinson, "Modeling sample variables with an experimental factor ontology," *Bioinformatics*, vol. 26, no. 8, pp. 1112–1118, 2010.

[51] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome research*, vol. 13, no. 11, pp. 2498–2504, 2003.

[52] D. Szklarczyk, R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, A. L. Gable, T. Fang, N. T. Doncheva, S. Pyysalo *et al.*, "The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest," *Nucleic acids research*, vol. 51, no. D1, pp. D638–D646, 2023.

[53] F. Jeanquartier, C. Jean-Quartier, and A. Holzinger, "Integrated web visualizations for protein-protein interaction databases," *BMC bioinformatics*, vol. 16, pp. 1–16, 2015.

[54] A. K. Bajpai, S. Davuluri, K. Tiwary, S. Narayanan, S. Oguru, K. Basavaraju, D. Dayalan, K. Thirumurugan, and K. K. Acharya, "Systematic comparison of the protein-protein interaction databases from a user's perspective," *Journal of Biomedical Informatics*, vol. 103, p. 103380, 2020.

[55] D. V. Veres, D. M. Gyurkó, B. Thaler, K. Z. Szalay, D. Fazekas, T. Korcsmáros, and P. Csermely, "Comppi: a cellular compartment-specific database for protein–protein interaction network analysis," *Nucleic acids research*, vol. 43, no. D1, pp. D485–D493, 2015.

[56] R. Reja, A. Venkatakrishnan, J. Lee, B.-C. Kim, J.-W. Ryu, S. Gong, J. Bhak, and D. Park, "Mitointeractome: mitochondrial protein interactome database, and its application in'aging network'analysis," in *BMC genomics*, vol. 10. Springer, 2009, pp. 1–8.

[57] A. K. Bajpai, S. Davuluri, K. Tiwary, S. Narayanan, S. Oguru, K. Basavaraju, D. Dayalan, K. Thirumurugan, and K. K. Acharya, "How helpful are the protein-protein interaction databases and which ones?" *bioRxiv*, p. 566372, 2019.

[58] BioGRID, "Experimental evidence codes," https://wiki.thebiogrid.org/doku.php/experimental$_s ystems$, 2024, navštíveno: 1.8.2024.

[59] R. Oughtred, J. Rust, C. Chang, B.-J. Breitkreutz, C. Stark, A. Willems, L. Boucher, G. Leung, N. Kolas, F. Zhang *et al.*, "The biogrid database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions," *Protein Science*, vol. 30, no. 1, pp. 187–200, 2021.

[60] I. Consortium, "About imex," http://www.imexconsortium.org/about/, 2024, navštíveno: 1.8.2024.

[61] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg, "The database of interacting proteins: 2004 update," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D449–D451, 2004.

[62] A. Sharma, R. Virk, M. Khurana, and R. Kaur, "Protein interaction databases: A review," *Research Journal of Life Sciences, Bioinformatics, Pharmaceutical and Chemical Sciences*, 2018.

[63] D. D. of Interacting Proteins), "Guide to the dip database," https://dip.doembi.ucla.edu/dip/Guide.cgi, 2024, navštíveno: 1.8.2024.

[64] C. M. Deane, Ł. Salwinski, I. Xenarios, and D. Eisenberg, "Protein interactions: two methods for assessment of the reliability of high throughput observations," *Molecular & Cellular Proteomics*, vol. 1, no. 5, pp. 349–356, 2002.

[65] K. Luck, D.-K. Kim, L. Lambourne, K. Spirohn, B. E. Begg, W. Bian, R. Brignall, T. Cafarelli, F. J. Campos-Laborie, B. Charloteaux *et al.*, "A reference map of the human binary protein interactome," *Nature*, vol. 580, no. 7803, pp. 402–408, 2020.

[66] S. Eyckerman, A. Verhee, J. V. der Heyden, I. Lemmens, X. V. Ostade, J. Vandekerckhove, and J. Tavernier, "Design and application of a cytokine-receptor-based interaction trap," *Nature cell biology*, vol. 3, no. 12, pp. 1114–1119, 2001.

[67] P. Cassonnet, C. Rolloy, G. Neveu, P.-O. Vidalain, T. Chantier, J. Pellet, L. Jones, M. Muller, C. Demeret, G. Gaud *et al.*, "Benchmarking a luciferase complementation assay for detecting protein complexes," *Nature methods*, vol. 8, no. 12, pp. 990–992, 2011.

[68] S. Orchard, M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N. H. Campbell, G. Chavali, C. Chen, N. Del-Toro *et al.*, "The mintact project—intact as a common curation platform for 11 molecular interaction databases," *Nucleic acids research*, vol. 42, no. D1, pp. D358–D363, 2014.

[69] D. Alonso-López, F. J. Campos-Laborie, M. A. Gutiérrez, L. Lambourne, M. A. Calderwood, M. Vidal, and J. De Las Rivas, "Apid database: redefining protein–protein interaction experimental evidences and binary interactomes," *Database*, vol. 2019, p. baz005, 2019.

[70] J. Das and H. Yu, "Hint: High-quality protein interactomes and their applications in understanding human disease," *BMC systems biology*, vol. 6, pp. 1–12, 2012.

[71] A. Calderone, L. Castagnoli, and G. Cesareni, "Mentha: a resource for browsing integrated protein-interaction networks," *Nature methods*, vol. 10, no. 8, pp. 690–691, 2013.

[72] G. N. Dimitrakopoulos, M. I. Klapa, and N. K. Moschonas, "Pickle 3.0: Enriching the human meta-database with the mouse protein interactome extended via mouse–human orthology," *Bioinformatics*, vol. 37, no. 1, pp. 145–146, 2021.

[73] D. R. Rhodes, S. A. Tomlins, S. Varambally, V. Mahavisno, T. Barrette, S. Kalyana-Sundaram, D. Ghosh, A. Pandey, and A. M. Chinnaiyan, "Probabilistic model of the human protein-protein interaction network," *Nature biotechnology*, vol. 23, no. 8, pp. 951–959, 2005.

[74] A. Elefsinioti, Ö. S. Saraç, A. Hegele, C. Plake, N. C. Hubner, I. Poser, M. Sarov, A. Hyman, M. Mann, M. Schroeder *et al.*, "Large-scale de novo prediction of physical protein-protein association," *Molecular & Cellular Proteomics*, vol. 10, no. 11, 2011.

[75] Q. C. Zhang, D. Petrey, L. Deng, L. Qiang, Y. Shi, C. A. Thu, B. Bisikirska, C. Lefebvre, D. Accili, T. Hunter *et al.*, "Structure-based prediction of protein–protein interactions on a genome-wide scale," *Nature*, vol. 490, no. 7421, pp. 556–560, 2012.

[76] M. Kotlyar, C. Pastrello, F. Pivetta, A. Lo Sardo, C. Cumbaa, H. Li, T. Naranian, Y. Niu, Z. Ding, F. Vafaee *et al.*, "In silico prediction of physical protein interactions and characterization of interactome orphans," *Nature methods*, vol. 12, no. 1, pp. 79–84, 2015.

[77] I. I. D. (IID), "About iid," http://iid.ophid.utoronto.ca, 2024, navštíveno: 1.8.2024.

[78] S. Consortium, "About string: Functional protein association networks," https://string-db.org/cgi/about?footer$_a$*ctive*$_s$*ubpage* = *content*, 2024, navštíveno: 1.8.2024.

[79] D. Szklarczyk, A. L. Gable, K. C. Nastou, D. Lyon, R. Kirsch, S. Pyysalo, N. T. Doncheva, M. Legeay, T. Fang, P. Bork *et al.*, "The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets," *Nucleic acids research*, vol. 49, no. D1, pp. D605–D612, 2021.

[80] C. Von Mering, L. J. Jensen, B. Snel, S. D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M. A. Huynen, and P. Bork, "String: known and predicted protein–protein associations, integrated and transferred across organisms," *Nucleic acids research*, vol. 33, no. suppl_1, pp. D433–D437, 2005.

[81] M. Jayapandian, A. Chapman, V. G. Tarcea, C. Yu, A. Elkiss, A. Ianni, B. Liu, A. Nandi, C. Santos, P. Andrews *et al.*, "Michigan molecular interactions (mimi): putting the jigsaw puzzle together," *Nucleic acids research*, vol. 35, no. suppl_1, pp. D566–D571, 2007.

[82] G. D. Bader, D. Betel, and C. W. Hogue, "Bind: the biomolecular interaction network database," *Nucleic acids research*, vol. 31, no. 1, pp. 248–250, 2003.

[83] S. Peri, J. D. Navarro, R. Amanchy, T. Z. Kristiansen, C. K. Jonnalagadda, V. Surendranath, V. Niranjan, B. Muthusamy, T. Gandhi, M. Gronborg *et al.*, "Development of human protein reference database as an initial platform for approaching systems biology in humans," *Genome research*, vol. 13, no. 10, pp. 2363–2371, 2003.

[84] V. G. Tarcea, T. Weymouth, A. Ade, A. Bookvich, J. Gao, V. Mahavisno, Z. Wright, A. Chapman, M. Jayapandian, A. Özgür *et al.*, "Michigan molecular interactions r2: from interacting proteins to pathways," *Nucleic acids research*, vol. 37, no. suppl_1, pp. D642–D646, 2009.

[85] S. Consortium, "Evidence for task berlipyc8ing," https://string-db.org/cgi/setevidence?taskId=beRLIPYC8ing&sessionId=bhUcXYFUmzRR&data_channel=experimental, 2024, navštíveno 7.8.2024.

Seznam obrázků

Seznam tabulek