9. Bioinformatika a proteiny II

David Potěšil

Core Facility – Proteomics

CEITEC-MU

Masaryk University

Kamenice 5, A26

phone: +420 54949 8426

email: david.potesil@ceitec.muni.cz

Obsah přednášky

5. Biologické sítě
6. Biologické sítě – biologické ontologie, KEGG
7. Biologické sítě – příklady použití
8. Vybrané on-line zdroje
9. Další vybrané aplikace
10. Několik zamyšlení závěrem
11. Příklad využití bioinformatických nástrojů

5. Biologické sítě

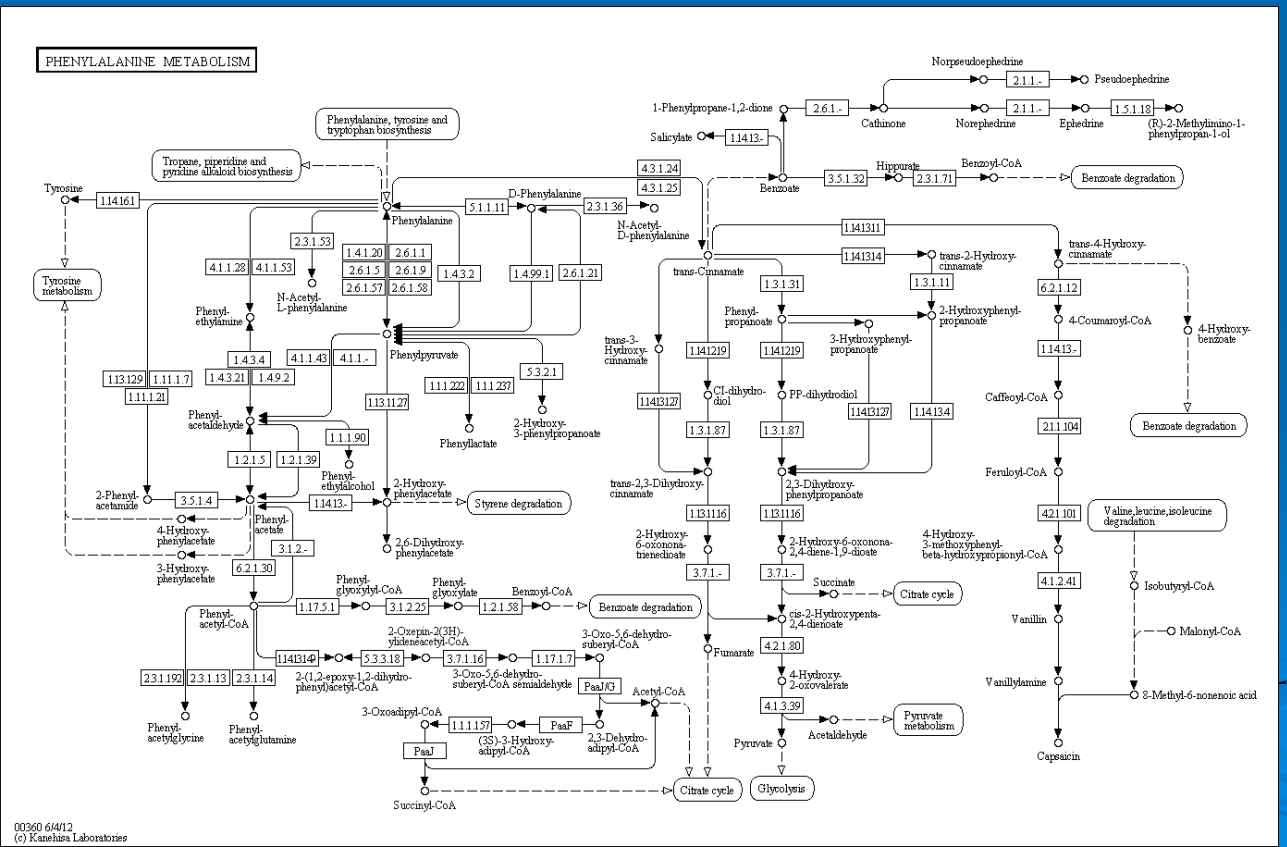
Biologické sítě

- **snaha o zachycení celého světa pomocí jeho jednotlivých složek (*nodes*) a vztahů mezi nimi (*edges*) – vytváření sítí (*networks*)**
  - prvopočátky již v 18. století...

- **biologická síť (**alt. sociální síť jako facebook, twitter**)**
  - **sada molekul** (alt. lidí)**, např. proteinů, geny, metabolity = *nodes***
  - **propojených pomocí definovaných, funkčních vztahů** (alt. přátelé)**; např. protein-protein interakce = *edges***

Biologické sítě – příklady

- **metabolické dráhy (*metabolic pathways*)**
  - spojují proteiny (*nodes*) skrze produkty a reaktanty (*edges*)
    - **produkt jednoho = substrát druhého**
  - **např. KEGG; WikiPathways**



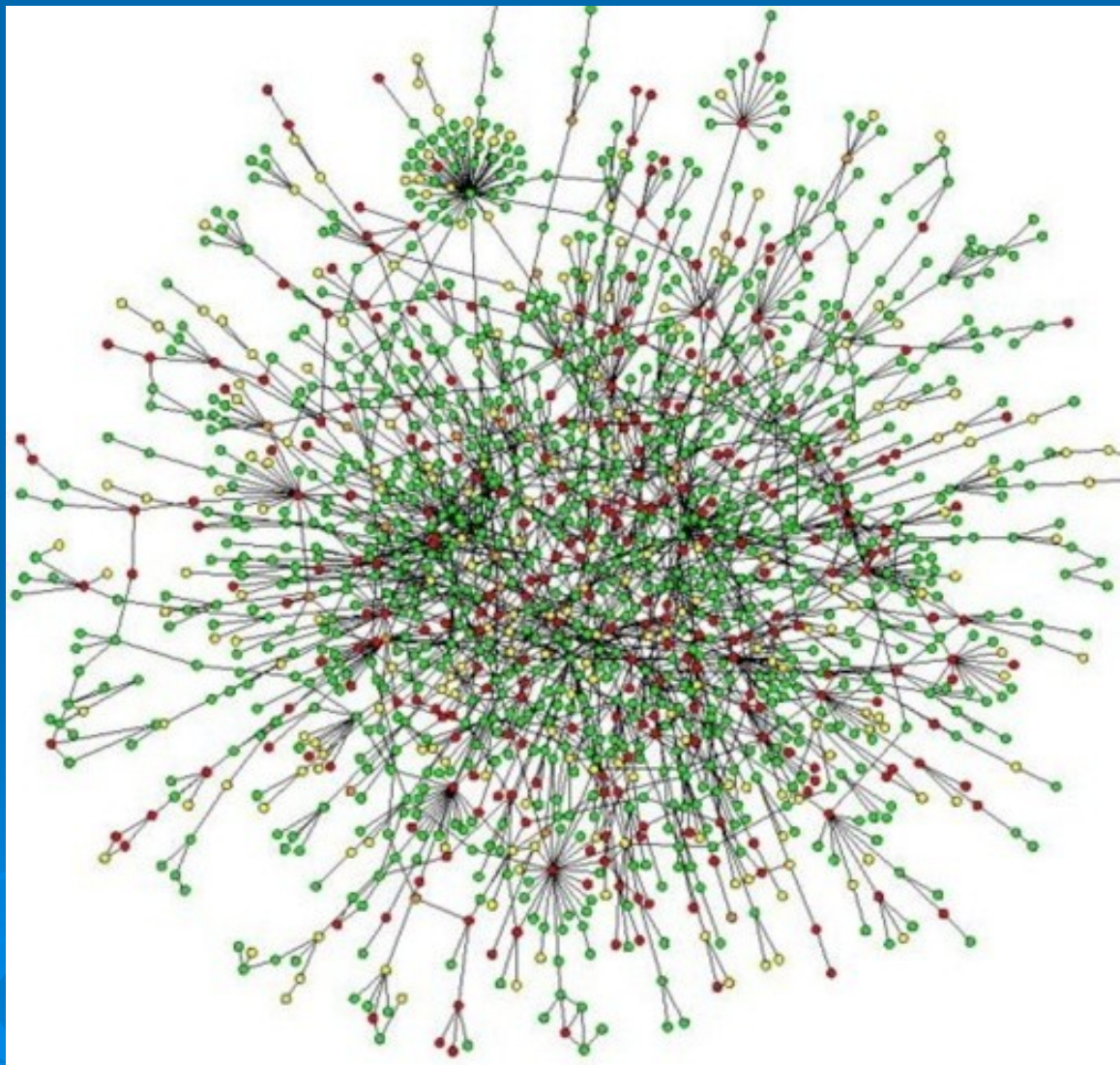
část metabolické sítě – metabolismus Phe (KEGG)

Biologické sítě – příklady (2)

- sítě regulace genů (*gene regulatory networks*; *DNA-protein interaction networks*)
  - edge = transkripční vztah mezi dvěma proteiny
  - jeden protein ovlivňuje expresi genu druhého proteinu(ů)

Biologické sítě – příklady (3)

- **protein-protein fyzické interakce – ze sítě samotné není přímá informace o významu dané interakce...**
  - ***nodes* – ?**
  - ***edges* – ?**
  - **příklady databází**
    - **STRING** (www.string-db.org)
    - MINT
    - DIP
    - BioGRID
    - ...



6. Biologické sítě

Biologické ontologie, KEGG, Reactome

Biologické ontologie

- **ontologie = systém kategorií (termínů; *terms*), do kterých jsou zařazeny jednotlivé informační jednotky, spolu s jejich vlastnostmi a vztahy**
- **biologické ontologie – příklady**
  - **proteiny (*gene products*) – genová ontologie (GO)**; funkce, lokalizace, ...
  - průběh buněčného dělení (*Cell Cycle Ontology*)
  - rostlinná ontologie (*Plant ontology*)

- **OLS – *Ontology Lookup Service***
  - **https://www.ebi.ac.uk/ols/index**
  - **jednotný přístup k více ontologiím**
  - **možnost procházet celé ontologie, případně vyhledávat termíny**

stále živý proces úprav/doplnění ontologií; není statické!

Genová ontologie (GO)

- **pravděpodobně nejvíce rozpracovaná biologická ontologie**
  - jak co do počtu termínů, tak co do počtu anotovaných položek (genů/prot.)
- **společné termíny pro všechny organizmy**

- **tři GO domény**
  - **buněčná komponenta (*cellular component*)**
    - informace o buněčné lokalizaci proteinu
  - **molekulární funkce (*molecular function*)**
    - informace o funkci proteinu
  - **biologický proces (*biological process*)**
    - informace o procesech, kterých se protein účastní

- **GO Slims (podmnožina GO termínů; organizmus, specifická aplikace, ...)**
- **http://www.geneontology.org/ + AmiGO2 prohlížeč (online)**

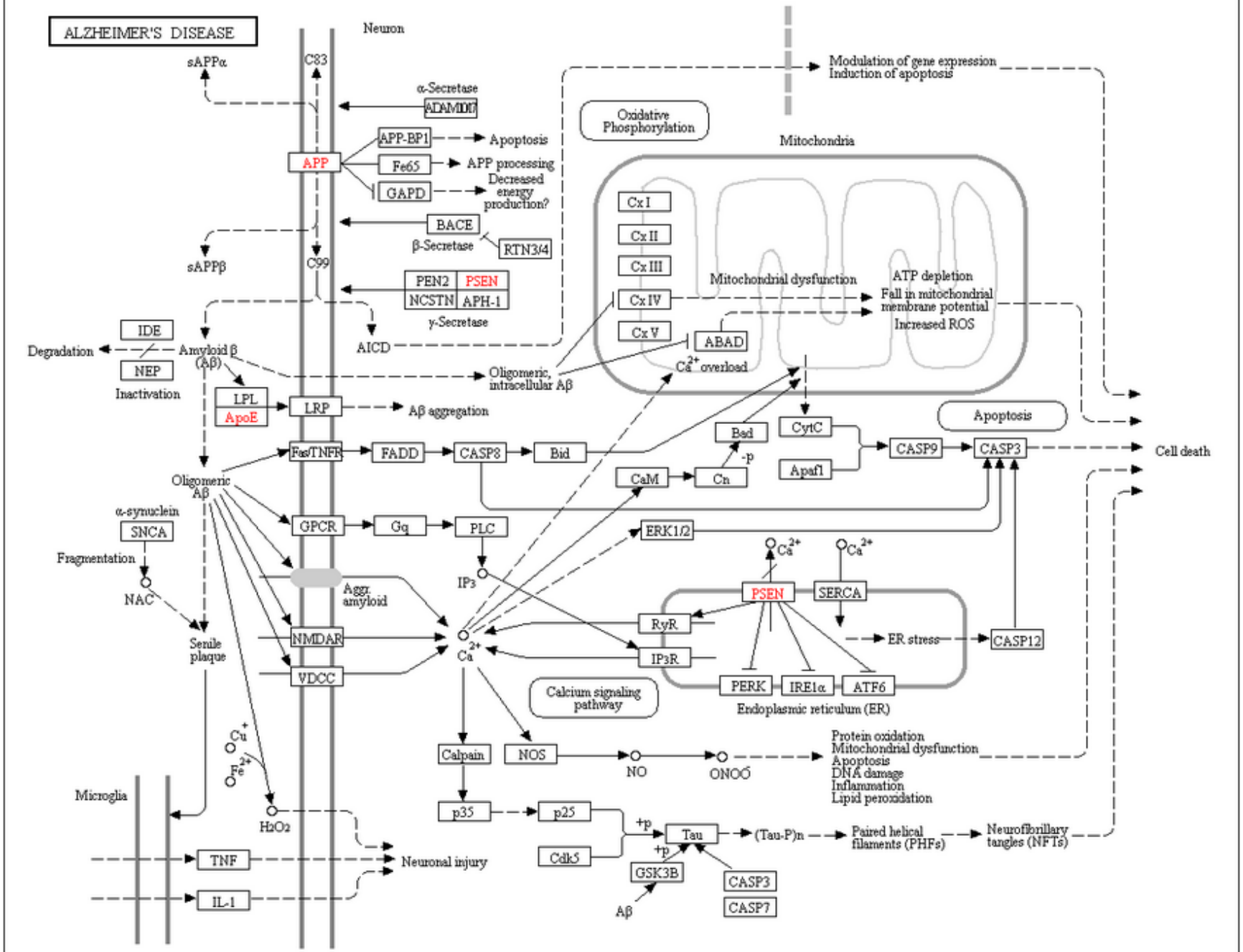
Genová ontologie (GO) (2)

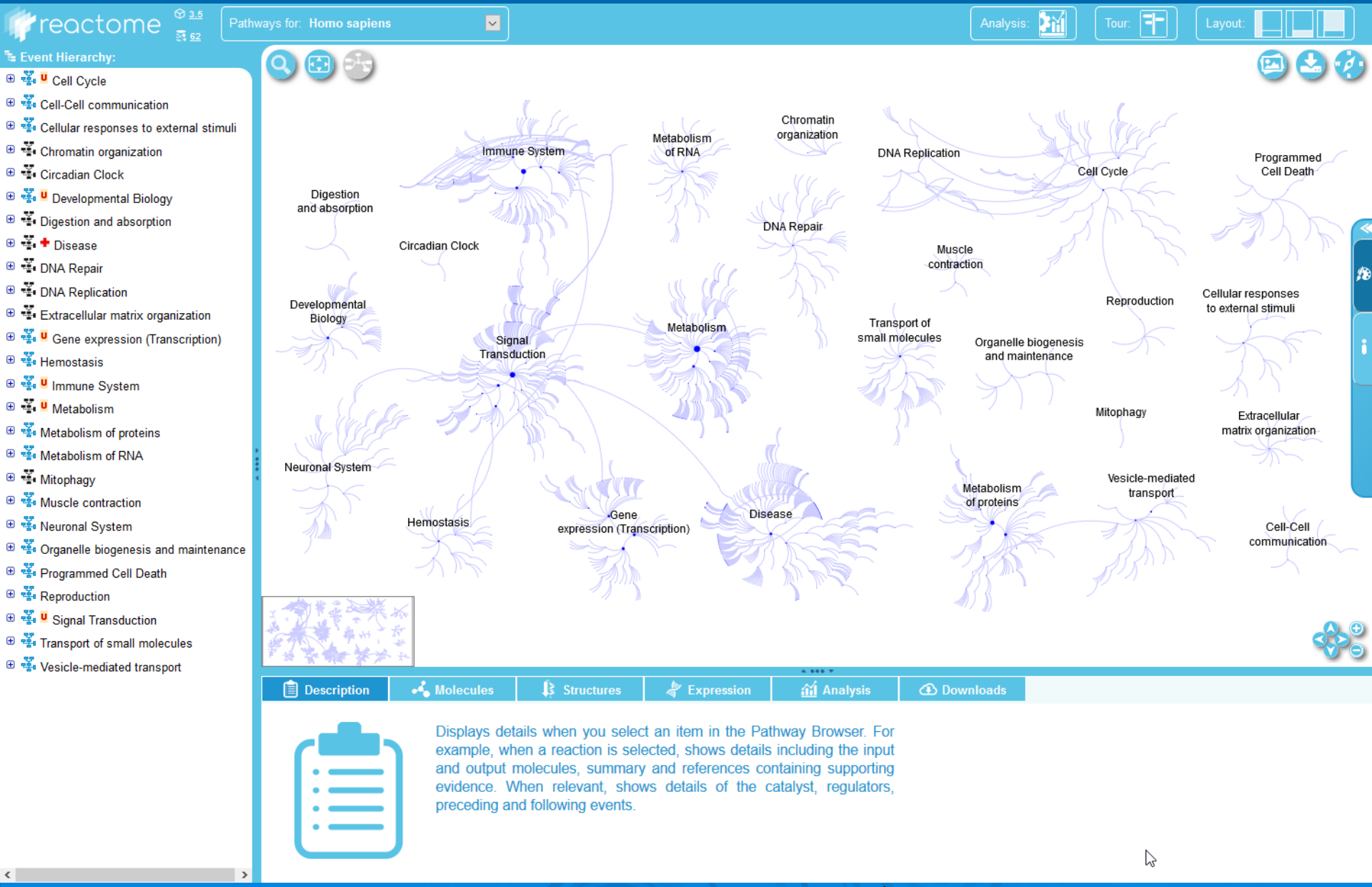
- **kde se berou data pro GO?**
  - **každá anotace obsahuje informaci o svém původu – *evidence code***
    - **http://geneontology.org/page/guide-go-evidence-codes**
  - **A) manuálně přiřazené správcem (*curator*)** (dále se dělí...)
    - ***experimental evidence codes* ⇒ z reálného experimentu**
    - ***High Throughput (HTP) evid. codes* ⇒ z *high throughput* analýz**
    - ***computational analysis evidence codes* ⇒ z *in silico* analýzy**
    - ***author statement evidence codes* ⇒ tvrzení autora + citace**
    - ***curatorial statement codes* ⇒ tvrzení správce, nepatří do žádné kategorie výše...**
  - **B) automaticky přiřazené (bez zásahu správce)**
    - ***automatically-assigned evidence code***
      - ***Inferred from Electronic Annotation* (IEA)**

KEGG, Reactome

- KEGG = *Kyoto Encyclopedia of Genes and Genomes*
- https://www.genome.jp/kegg/
- manuální katalogizace znalostí biologických systémů v počítačově zpracovatelné podobě
- čerpá z dosavadních znalostí v dané problematice
- z informací na nízké biologické úrovni nám umožní odvodit informace na vyšší biologické úrovni
  - například ze seznamu regulovaných genů/proteinů odvodí informaci o ovlivněných metabolických drahách

- obdobně i např. https://www.reactome.org
  - progresivnější vývoj
  - vizualizace dostupných drah v souhrnném diagramu


ALZHEIMER'S DISEASE
Neuron
sAPPα
C83
α-Secretase
ADAM10?
APP
APP-BP1
Apoptosis
Fe65
APP processing
GAPD
Decreased energy production?
BACE
β-Secretase
RTN3/4
sAPPβ
C99
PEN2
PSEN
NCSTN
APH-1
γ-Secretase
AICD
IDE
Degradation
Amyloid β (Aβ)
NEP
Inactivation
LPL
ApoE
LRP
Aβ aggregation
Oligomeric, intracellular Aβ
Modulation of gene expression
Induction of apoptosis
Oxidative Phosphorylation
Mitochondria
Cx I
Cx II
Cx III
Cx IV
Cx V
Mitochondrial dysfunction
ATP depletion
Fall in mitochondrial membrane potential
Increased ROS
ABAD
Ca2+ overload
Bad
-p
CytC
Apaf1
CASP9
CASP3
Apoptosis
Cell death
Fas/TNFR
FADD
CASP8
Bid
CaM
Cn
Oligomeric Aβ
α-synuclein
SNCA
Fragmentation
NAC
Senile plaque
GPCR
Gq
PLC
ERK1/2
Aggr. amyloid
IP3
Ca2+
PSEN
SERCA
RyR
IP3R
ER stress
CASP12
NMDAR
VDCC
PERK
IRE1α
ATF6
Endoplasmic reticulum (ER)
Calcium signaling pathway
Cu+
Fe2+
H2O2
Microglia
Calpain
NOS
NO
ONOO-
Protein oxidation
Mitochondrial dysfunction
Apoptosis
DNA damage
Inflammation
Lipid peroxidation
p35
p25
Cdk5
+p
Tau
(Tau-P)n
Paired helical filaments (PHFs)
Neurofibrillary tangles (NFTs)
GSK3B
Aβ
CASP3
CASP7
TNF
IL-1
Neuronal injury
05010 9/5/16
(c) Kanehisa Laboratories


reactome
3.5
62
Pathways for: Homo sapiens
Analysis:
Tour:
Layout:
Event Hierarchy:
Cell Cycle
Cell-Cell communication
Cellular responses to external stimuli
Chromatin organization
Circadian Clock
Developmental Biology
Digestion and absorption
Disease
DNA Repair
DNA Replication
Extracellular matrix organization
Gene expression (Transcription)
Hemostasis
Immune System
Metabolism
Metabolism of proteins
Metabolism of RNA
Mitophagy
Muscle contraction
Neuronal System
Organelle biogenesis and maintenance
Programmed Cell Death
Reproduction
Signal Transduction
Transport of small molecules
Vesicle-mediated transport
Immune System
Metabolism of RNA
Chromatin organization
DNA Replication
Cell Cycle
Programmed Cell Death
Digestion and absorption
Circadian Clock
DNA Repair
Muscle contraction
Developmental Biology
Signal Transduction
Metabolism
Transport of small molecules
Reproduction
Cellular responses to external stimuli
Organelle biogenesis and maintenance
Mitophagy
Extracellular matrix organization
Neuronal System
Hemostasis
Gene expression (Transcription)
Disease
Metabolism of proteins
Vesicle-mediated transport
Cell-Cell communication
Description
Molecules
Structures
Expression
Analysis
Downloads
Displays details when you select an item in the Pathway Browser. For example, when a reaction is selected, shows details including the input and output molecules, summary and references containing supporting evidence. When relevant, shows details of the catalyst, regulators, preceding and following events.

7. Biologické sítě

Příklady použití

Příklad 1: Vliv nízkomolekulární látky na rostlinu

- **identifikace sady ovlivněných proteinů**
- **jsou tyto proteiny zahrnuty v odpovídající metabolické dráze?** (KEGG)
  - fungoval experiment dle předpokladu?
- **jaké jiné metabolické dráhy byly „významně“ zastoupeny?** (KEGG)
  - objevili jsme i jiné, dosud nepotvrzené, ale související metabolické dráhy?
- **jsou známy proteinové komplexy mezi nalezenými proteiny?** (protein-protein interakční síť)
  - dokáží nám tyto pomoci při interpretaci vlivu látky na rostlinu?
- **je mezi proteiny zastoupeno více proteinů z konkrétního GO termínu?** (GO)
  - na základě daných GO termínů je možno odvodit souvislosti s funkcí či lokalizací probíhajících (i sekundárních) dějů

Příklad 2: Interakční partneři zvoleného proteinu

- **vidíme již známé interakční partnery?**
  - pozitivní kontrola průběhu experimentu
- **nově pozorované interakce**
  - **studium biologických vlastností možných interakčních partnerů** (GO termíny, metabolické dráhy, ...)
    - **zapadají tyto do již známých informací o funkci, lokalizaci aj. zvoleného proteinu?**
    - **jsou patrné souvislosti s lokalizací našeho proteinu?**

Příklad 3: Studium proteinu, se vztahem k onemocnění...

- **jsou pro tento protein známy proteinové interakce?**
  - u interakčních partnerů zvýšená pravděpodobnost, že se tyto proteiny aktivně nebo pasivně účastní daného onemocnění; GO analýza

- **je známa lokalizace proteinu v buňce?**
  - lokalizace může souviset s funkcí (konkrétní funkce proteinu často vázána na jeho buněčnou lokalizaci)

- **je známa úloha proteinu v některé metabolické dráze?**
  - možná úloha (i nepřímá, ovlivňující např. „jen“ dostupnost klíčového proteinu) dráhy v onemocnění – její proteinové i neproteinové komponenty

⇒ potencionální cíle dalšího studia a nové léčby

Příklad 4: „Zdraví versus nemocní“ – rozdílně exprimované proteiny

- **kterých metabolických drah se proteiny účastní?**
  - vysvětluje to důsledky, průběh, ... vlastní nemoci?

- **jsou rozdílné proteiny převážně lokalizované v některé z organel?**
  - má tato informace souvislost se vznikem/průběhem nemoci v konkrétní části buňky?

- **je mezi proteiny „často“ přítomen konkrétní GO termín?**
  - má tento termín souvislost se vznikem, průběhem, projevem onemocnění?

7. Biologické sítě

Analýza biologických sítí

Analýza sítí (*network analysis*) – na co si dát pozor?

- **falešně pozitivní i negativní informace v biologických sítích**
  - častěji falešně negativní – absence příslušných proteinů v sítích
- **mnoho dat v databázích z automatických analýz dostupných dat**
  - i přes kontrolu nemusí zcela odpovídat zdrojovým datům a skutečnosti
  - někdy lze vyloučit z analýzy (např. automaticky anotované GO...)
- **stále víme málo...**
  - důležitost sekvenčních a funkčních homologií u proteinů bez anotace
  - **rychlý vývoj v anotaci proteinů a vývoji bioinformatických nástrojů!**
- **volba vhodných otázek, na které nám biologické sítě dokážou dát odpověď**

Analýza sítí – jak se postavit k výstupům?

- **manuální validace výstupů**
- **ověření původních zdrojů**
- **pochybovat a ptát se**
- **nesnažit se proces analýzy a ověření výsledků urychlit**
- **experimentální ověření závěrů** (např. buněčné linie s mutantní formou genu)
  - **drahé a časově náročné ⇒ důkladné ověření předchozích kroků!**

8. Vybrané on-line zdroje

Universal Protein Resource (UniProt)

- http://www.uniprot.org
- **bohatá anotace proteinů s odkazy na specializované databáze/zdroje**
- **široké možnosti využití v databázi přítomných informací**
  - **převod (*mapping*) identifikátorů z různých databází** (např. UniProt → KEGG)
  - **tabulkový formát s vybranými informacemi o sadě proteinů** (stažení...)
  - **možný pohled ze strany určité taxonomie, nemoci, buněčné lokalizace...**
  - **informace o přítomnosti sady proteinů v metabolických drahách, GO**

Universal Protein Resource (UniProt) (2)

- **odkud bere proteinové sekvence?**
  - **většina (~95 %) z nukleotidových databází CDS (*coding sequences*)**
    - sekvence zadávány jednotlivými výzkumnými skupinami
    - EMBL-Bank/GenBank/DDBJ
      - pod *International Nucleotide Sequence Databases* (INSD)
  - **translace na proteinovou sekvenci**
  - **automatické zpracování za účelem anotace a klasifikace proteinů**
    - **na základě sekvenční homologie**
  - **takto zpracovaný protein je zaveden do UniProtKB/TrEMBL databáze**
  - **je-li protein vybrán pro manuální zpracování, provede správce (*curator*) jeho manuální zařazení do UniProtKB/SwissProt databáze**

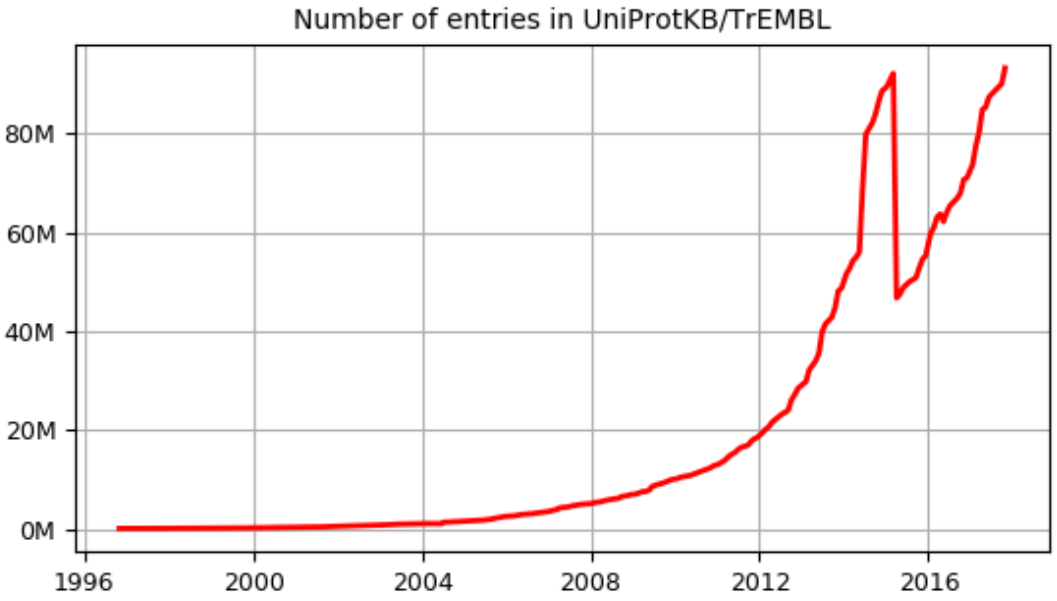
Universal Protein Resource (UniProt) (3)

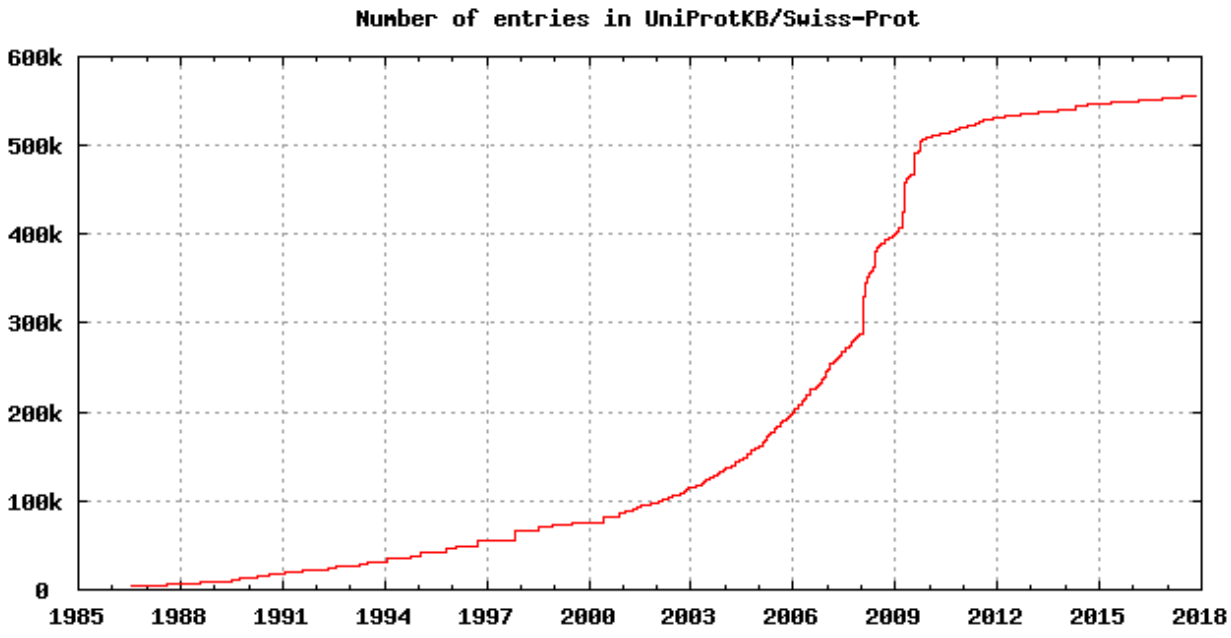
- UniProtKB/SwissProt – manuální zpracování (*curation*) správcem
  - kontrola sekvence – není-li v původní sekvenci chyba
  - sekvenční analýza – manuálně kontrolované predikce atd.
  - studium literárních zdrojů – dodány biologicky relevantní informace k proteinu na základě dostupných publikací; název genu, funkce proteinu, enz. aktivita, subc. lokalizace, přiřazení GO termínů k proteinu atd.
  - získání informací o proteinové rodině – zjištění případných členů proteinové rodiny a jejich společné zpracování
  - přidání zdrojů – z jakého konkr. zdroje pochází ta která informace; možnost ověření přítomných informací „u zdroje“
  - kontrola kvality, integrace, aktualizace – všechna manuálně přidaná data zkontrolována a zakomponována do nové verze SwissProt db.

TrEMBL/SwissProt

high-throughput

„závazek“





Universal Protein Resource (UniProt) (5)

- **typy proteinových setů v UniProtKB proteinové databázi**
  - **UniProtKB/TrEMBL – automaticky klasifikované a anotované**
    - i zde probíhají automaticky řízené opravy...
  - **UniProtKB/SwissProt – po manuální úpravě správcem (*curation*)**
  - ***(Complete) Proteome Set* – pro kompletně sekv. organizmy** (T+S)
    - **dnes již bez *complete* označení, rozděleno dle taxonomií**
  - ***Reference Proteome Set* – vybrané modelové organizmy** (T+S)
    - ***"... The approach adopted by UniProt to meet this challenge is to define a set of 'reference proteomes' which are 'landmarks' in proteome space."***
    - ***"Reference proteomes have been selected among all proteomes (manually and algorithmically, according to a number of criteria) to provide broad coverage of the tree of life."***

Universal Protein Resource (UniProt) (6)

- typy proteinových setů v UniProtKB proteinové databázi
  - UniRef – *UniProt Reference Clusters*
    - seskupené primární sekvence do klastrů na základě sekv. podobnosti
    - umožňuje skrýt „redundantní“ proteinové sekvence
    - UniRef100 – seskupeny záznamy se 100% identitou
    - UniRef90; UniRef50
      - snížení počtu sekvencí (o ~58 a 79%) – BLAST aj.
    - seskupováno dle kritérií – SwissProt, jméno, organizmus, délka
  - UniParc – databáze proteinových sekvencí
    - unikátní identifikátor pro každou primární sekvenci (UNI)
    - identifikátor se nikdy nemění, ani nemaže
    - vedle sekvence informace o zdrojové databázi, identifikátoru atd.

Universal Protein Resource (UniProt) (7)

- **typy proteinových setů v UniProtKB proteinové databázi (2)**
  - **UniSave – *UniProtKB Sequence/Annotation Version Archive***
    - **zpřístupňuje různé verze daného proteinu (*accession*)**
    - **relativně nové, prozatím „jen“ txt výstupní formát (**i json...**)**
      - **plus je možné jít na starší verze u daného proteinu**
    - „*The UniProtKB Sequence/Annotation Version Archive (UniSave) has the mission of providing freely to the scientific community a repository containing every version of every Swiss-Prot/TrEMBL entry in the UniProt Knowledge Base (UniProtKB). This is achieved by archiving, every release, the entry versions within the current release.*“
    - „*The primary usage of this service is to provide open access to all entry versions of all entries. In addition to viewing their content, one can also filter, download and compare versions.*“

PubMed

- http://www.ncbi.nlm.nih.gov/pubmed
- více orientovaná na genomová data, ale...
- *Protein Clusters* – obdoba UniRef
- RefSeq – obdoba SwissProt; méně informačně „hodnotné“; oproti SwissProt cca 4M RefSeq záznamů

- obdobně informace o jednotlivých organizmech, taxonomiích aj.
  - nenabízí tak široké možnosti filtrování a práce s proteinovými sekvencemi jako UniProt

- mimo to i indexace vědeckých publikací aj.

European Bioinformatics Institute (EBI)

- http://www.ebi.ac.uk/services
- opět sada bioinformatických nástrojů a databází pro studium proteinů a souvisejících informací
- např. zmiňované InterPro; GeneOntology.org; OLS; ...

Expasy

- http://expasy.org
- sada nástrojů pro práci s proteiny/geny
- převážně nástroje z dílny *Swiss Institute of BioInformatics* (SIB; http://www.isb-sib.ch/)
- původně pouze proteomický portál
- rozšířen (2011) o genomické, transkriptomické aj. informace a nástroje

bioinformatics.ca Links Directory

- http://bioinformatics.ca/links_directory/
- sady odkazů na různé kategorie on-line zdrojů

OMICtools

- http://omictools.com/; opět sada bioinformatických nástrojů

Pax-DB

- http://pax-db.org/#!home
- databáze abundancí jednotlivých proteinů v organizmech či jejich částech

DAVID

- https://david.ncifcrf.gov/
- extrakce biologického významu v seznamu genů/proteinů

CESNET

- https://www.cesnet.cz/
- infrastruktura pro virtuální stroje, osobní certifikáty, filesender, ownCloud aj.
- alternativně také nástroje od ÚVT MU
  - např. openstack – virtualizační prostředí

ELIXIR

- https://elixir-europe.org/
- *„ELIXIR is an intergovernmental organisation that brings together life science resources from across Europe. These resources include databases, software tools, training materials, cloud storage and supercomputers.“*

bio.tools

- https://bio.tools/
- *„Essential scientific and technical information about software tools, databases and services for bioinformatics and the life sciences.“*

BioStar

- https://www.biostars.org/
- **bioinformatické diskuzní a QA fórum**

StackOverflow

- https://stackoverflow.com/
- **QA fórum, skriptování**

Google, Yahoo, ...

- obecné, globální, vyhledávací servery
- „... Na googlu je všechno, jen je potřeba položit správnou otázku...“

9. Další vybrané aplikace

Microsoft Excel

- **flexibilní „tabulkový procesor“**
- **jediný *closed source* v tomto výběru... (**výhody, nevýhody...**)**
- **pro jednodušší až pokročilou analýzu dat**
- **záznam a úprava VBA maker (**skriptů**), pokročilejší mohou psát samostatné skripty/programy**
- **mnoho implementovaných funkcí („**=suma(A1:A10)“ to jen začíná...**)**
  - **klikněte na *Insert function* a prozkoumejte co je vše možné...**
- **pokud dostupné funkce nestačí, napište si vlastní... (**pomocí VBA**)**
- **včetně definování „proměnných“ (***Formulas – Name manager***)**

Galaxy server

- ***open-source*** **serverová aplikace umožňující pokročilou analýzu dat**
- **formou konkrétního sledu kroků provádějící jednotlivé operace** (např. filtrování datové matice)
- **často používané kroky lze uložit jako *workflow***
- **sdílení *workflow* i výsledků s kterýmkoli jiným uživatelem**
  - **reprodukovatelnost**
- **relativně velká a aktivní komunita udržující celý projekt**

- **veřejné servery, kde si můžete provést analýzu Vašich dat** (je možné mít i vlastní instanci na vlastním hardware...)
  - **např. https://usegalaxy.eu/**
  - **běží na superpočítačích, jednotlivé kroky se zařazují do fronty k zpracování**

KNIME

- https://www.knime.com/
- ***open-source* platforma pro analýzu dat**
- **stovky nodů, které lze propojovat do nelineárních celků (*workflows*)**
  - např. operace s datovými maticemi, statistiky, reportování, ...
- **umožňuje zpracovávat data také vlastními skripty** (R, Python, Java)
- **reprodukovatelné procesování dat formou vizuálních nodů**

Cytoscape

- https://cytoscape.org/
- *open-source* platforma pro vizualizaci interakčních sítí, biologických sítí a drah a integrování těchto sítí s anotacemi, expresními daty atd.
- modulární design
  - moduly = aplikace (*apps*) zajišťují specifickou funkci
  - http://apps.cytoscape.org/

Jupyter notebooky

- http://jupyter.org/
- skriptové prostředí pro snadnější sdílení Vaší práce
- něco na způsob „laboratorního deníku“ bioinformatika z jeho skriptové práce
- webová aplikace
- propojuje skriptové prostředí s vizualizací pomocí grafů atd.
- možno skriptovat v Python, R, Ruby, ...
- export do a otevření v prohlížeči ale i jako pdf...
- celý aplikovaný postup je zaznamenám a reprodukovatelný!

Aplikace běžící v kontejnerech (*containers*)

- např. Docker (https://www.docker.com/), Quay (https://www.openshift.com/products/quay)
- podobné virtuálním strojům, ale je odlišné využití hardware
- nutno mít nainstalovanou např. Docker aplikaci a s její pomocí si lokálně pustíte kontejner (*container*), který
  - obsahuje operační systém pro běh aplikace (Linux, ale už i Windows)
  - obsahuje konkrétní doinstalované balíčky, aplikace (jako python, R, Java, Blast, KNIME, ...) v konkrétní verzi
  - je nastaven pro okamžité použití
- mnoho volně dostupných aplikací/prostředí např. na Docker Hub

- BioContainers (https://biocontainers.pro/#/)
  - bioinformatické aplikace ve formě kontejnerů

Příklad reprodukovatelného procesování dat

- kombinace Docker kontejneru s nainstalovaným KNIME
  - https://github.com/OmicsWorkflows/KNIME_docker_vnc
  - https://hub.docker.com/r/cfprot/knime/
- přístup přes VNC
- práce uvnitř kontejneru pomocí KNIME workflow

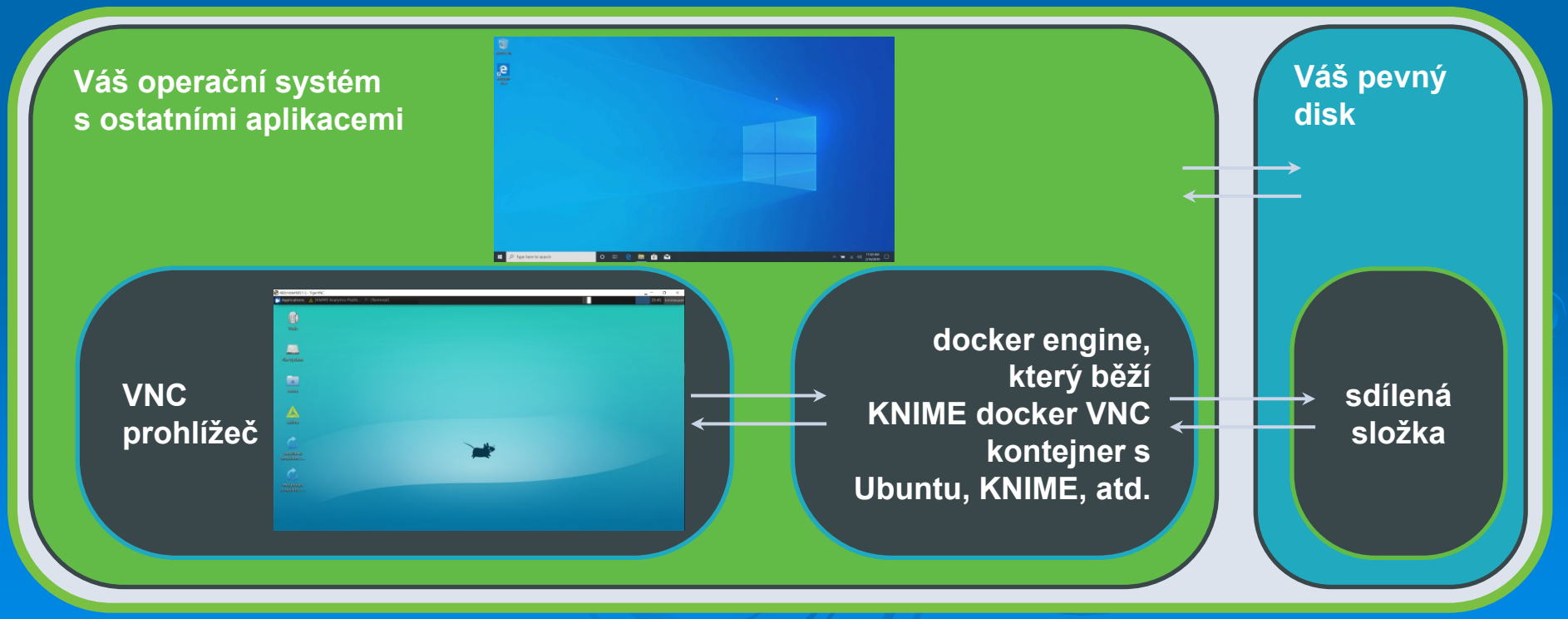
- KNIME workflow a kontejner, kde workflow vznikl umožňuje zreprodukovat totožný postup, nebo se jen podívat na mezikroky, konkrétní skripty atd...

- kontejner obsahuje konkrétní verze VŠECH aplikací
  - reprodukovatelné prostředí i procesování dat i po letech


Ubuntu (včetně grafického rozhraní, int. prohlížeče atd.)
KNIME Analytics Platform s dalšími moduly
KNIME dockerV
version 3.7.
python a vybrané balíčky pro použití v KNIME
SOFTWARE KONTEJNER (virtualní stroj)
R a vybrané balíčky pro použití v KNIME
Skripty pro konkrétní účely jako git
Vzdálený přístup přes VNC a sdílenou složku

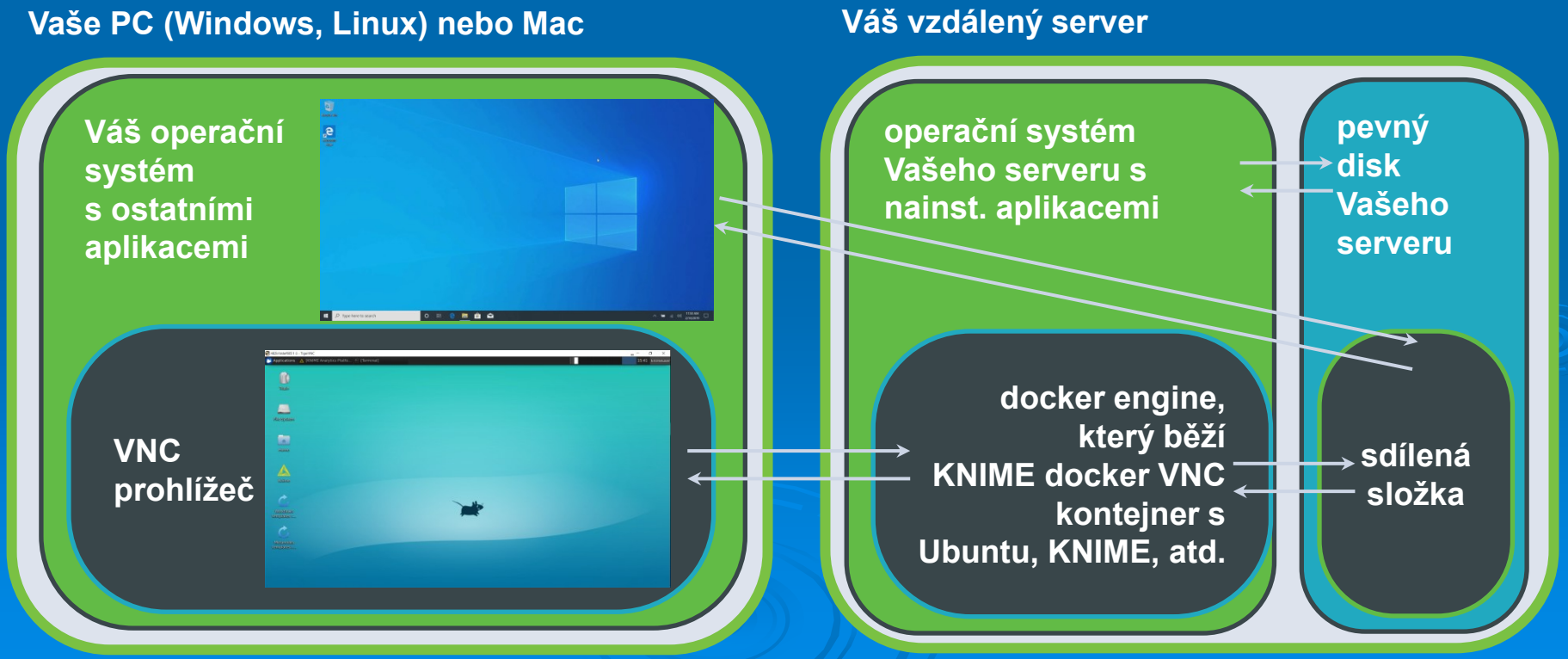
Použití softwarového kontejneru – lokální verze

- lokálně instalovaná docker aplikace vytvoří kontejner, kde běží další operační systém a další aplikace co jsou součástí *docker image*
- z lokálního počítače se pomocí VNC prohlížeče připojím dovnitř běžícího kontejneru a mohu pracovat jako na běžném počítači



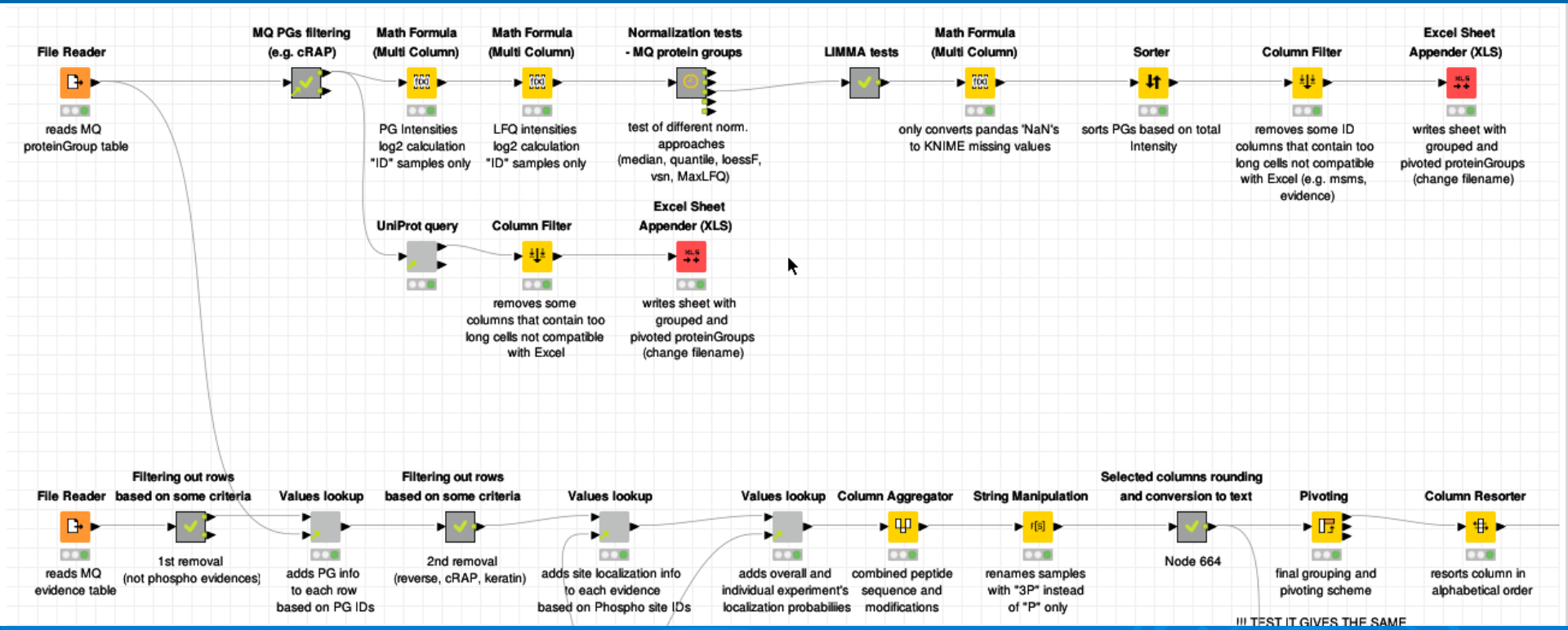
Použití softwarového kontejneru – serverová verze

- docker aplikace běží na serveru a na serveru se vytvoří i kontejner, kde běží další operační systém a další aplikace co jsou součástí *docker image*
- z lokálního počítače se pomocí VNC prohlížeče připojím dovnitř běžícího kontejneru a mohu pracovat jako na běžném počítači



Příklad reprodukovatelného procesování dat (2)

- KNIME workflow s komentáři k jednotlivým krokům...



https://www.jetbrains.com/

- **subjektivně velmi efektivní programovací studio**
- **výborné pro začátky s programováním (**našeptávání proměnných, funkcí, kontrola kvality kódu i pro stylistické stránce, ...**)**
- **např. pro programování v python (PyCharm), ale i jiných programech**
  - **https://www.jetbrains.com/pycharm-edu/learners/**

10. Několik zamyšlení závěrem

Rychlý vývoj bioinformatických aplikací/databází

- **vzniká hodně nástrojů/databází, které nejsou následně používané**
  - nepoužívané nástroje často dále nevyvíjené, neaktualizované (přítomnost chyb, které se objeví až při masivním používání...), používají zastaralé algoritmy, používají starší proteinové databáze...
- **význam „zavedených“ zdrojů bioinformatických nástrojů/databází (**UniProt, Pubmed, EBI**)**
  - např. anotace proteinů, vytváření biologických sítí – **lidské kapacity**
  - dlouholeté zkušenosti nutné k střednědobému **směřování vývoje**
- **důležitá grafická stránka programu/databáze a prvotní „jednoduchost“**
  - důležité pro rychlé „rozkoukání“, *user friendly* uživatelské prostředí
- **významná předchozí zkušenost s prací v aplikaci/s databází**
  - nové aplikace to nemají snadné...
  - důvod proč i nápadité nástroje mohou zůstat nepoužívány

Rychlý vývoj bioinformatických aplikací/databází (2)

- **bioinformatické aplikace/databáze není možné nevyvíjet/neaktualizovat**
  - při vytváření nástroje/databáze nutno počítat s udržitelností jeho vývoje...
- **veřejně dostupný kód (*open-source*) výhodou**
  - validace, vylepšení ostatními, ale i riziko „zneužití“
  - obdobně i pro natrénované modely (umělé neuronové sítě)
    - https://kipoi.org/
- **několik let (nejen v bioinformatice...) je velmi dlouhá doba**
  - aktualizace minimálně 1 ročně, optimálně měsíční, půlroční
  - i přes to mohou starší nástroje fungovat lépe než novější...
    - případně nic „lepšího“ není
  - důležité celosvětové reference a citovanost/používání (recentní) daného nástroje/databáze (např. i informace na Githubu, případně jiném *vcs*)

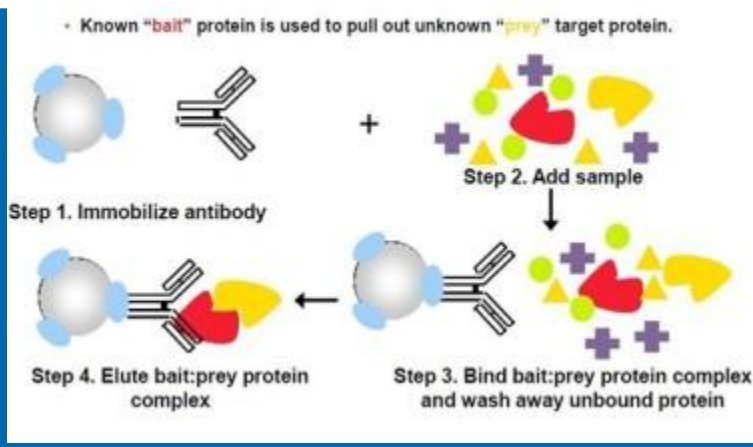
Rychlý vývoj bioinformatických aplikací/databází (3)

- **školící programy/workshopy/stáže v bioinformatických centrech**
  - EBI, SIB aj.
- **význam spoluprací – jeden tým často nedokáže pojmout celé spektrum použitých nástrojů, technologií, přístupů včetně interpretace výstupů**

11. Příklad využití bioinformatických nástrojů

Zavedení problému

- **studium proteinových komplexů vybraného proteinu**



https://www.creative-proteomics.com

- **imunoprecipitace proteinových komplexů (*IP experiment*)**
  - protilátka proti proteinu (*bait*), u kterého chceme zjistit jeho partnery
  - např. protilátka imobilizovaná např. na kuličkách (magnetické, v kolonkách)
  - nativní prostředí při experimentech – podmínky pro interakce jako *in vivo*
  - výstupem *pull-down* roztoky – proteiny vázající se na *bait* a nespecificky vázené proteiny
  - paralelně experimenty bez *bait* – negativní kontrola pro nespecificky vázané proteiny – *bead proteome*
  - minimálně 3 biologické replikáty, lépe 5 od vzorku i negativní kontroly

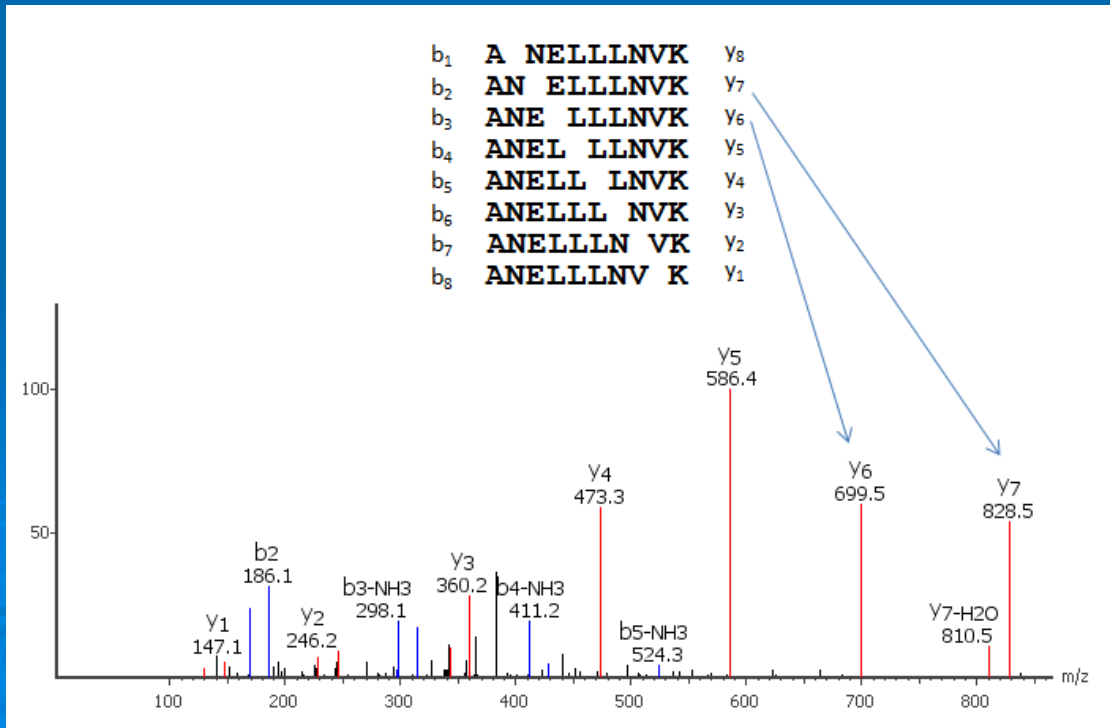
LC-MS/MS analýza *pull-down* vzorků

- **digesce proteinů** (x100) ⇒ **peptidy** (např. trypsinem; x1 000-x10 000)
- **LC-MS/MS analýza směsi peptidů**
  - **peptidy vstupují do hmotnostního spektrometru (MS) v pořadí rostoucí hydrofobicity** (LC separace)
- **MS zjistí MW peptidů a získá MS/MS spektra** (fragmentační spektrum vybraného peptidu)

např. peptid ANELLLNVK
(MW 1012,5917 Da**)**

1. **MW_{exp} = 1012,5923 Da**
   (0,6 ppm chyba)

2. změřené fragmentační (MS/MS) spektrum ⇒
(CID; *collision induced dissociation*)

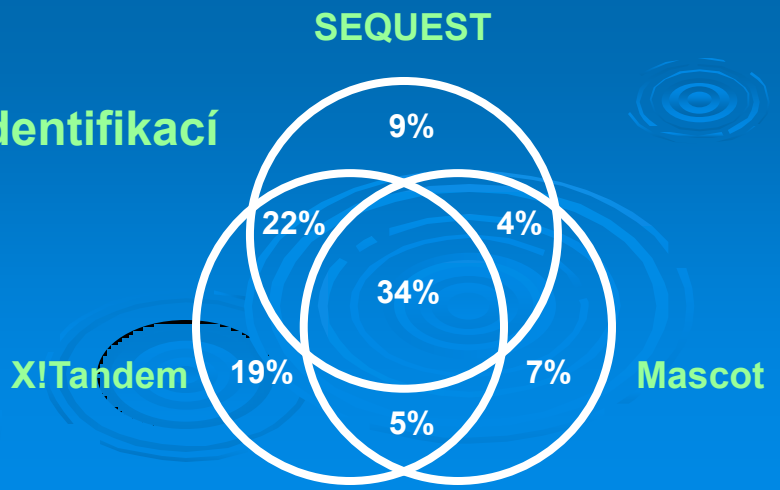


Zpracování LC-MS/MS dat (zde začíná bioinformatika)

- **LC-MS/MS data z analýz *pull-down* vzorků po digesci = MS/MS spektra**
- **řádově 10 000 – 1 000 000 MS/MS spekter**
- **identifikace peptidů**
  - **vycházíme z proteinové databáze, např. TAIR (*Arabidopsis thaliana*)**
  - ***in silico* se vytvoří seznam možných peptidů**
  - **>20 algoritmů pro automat. přiřazení MS spektra možným peptidům**
    (Sequest, Mascot, XTandem!, OMSSA, Phenyx, Andromeda, ...)
  - **jiný algoritmus ⇒ jiný přístup ⇒ různá citlivost ⇒ odlišné výsledky**

⇒ kombinace algoritmů

⇒ zvýšení počtu pozitivních identifikací

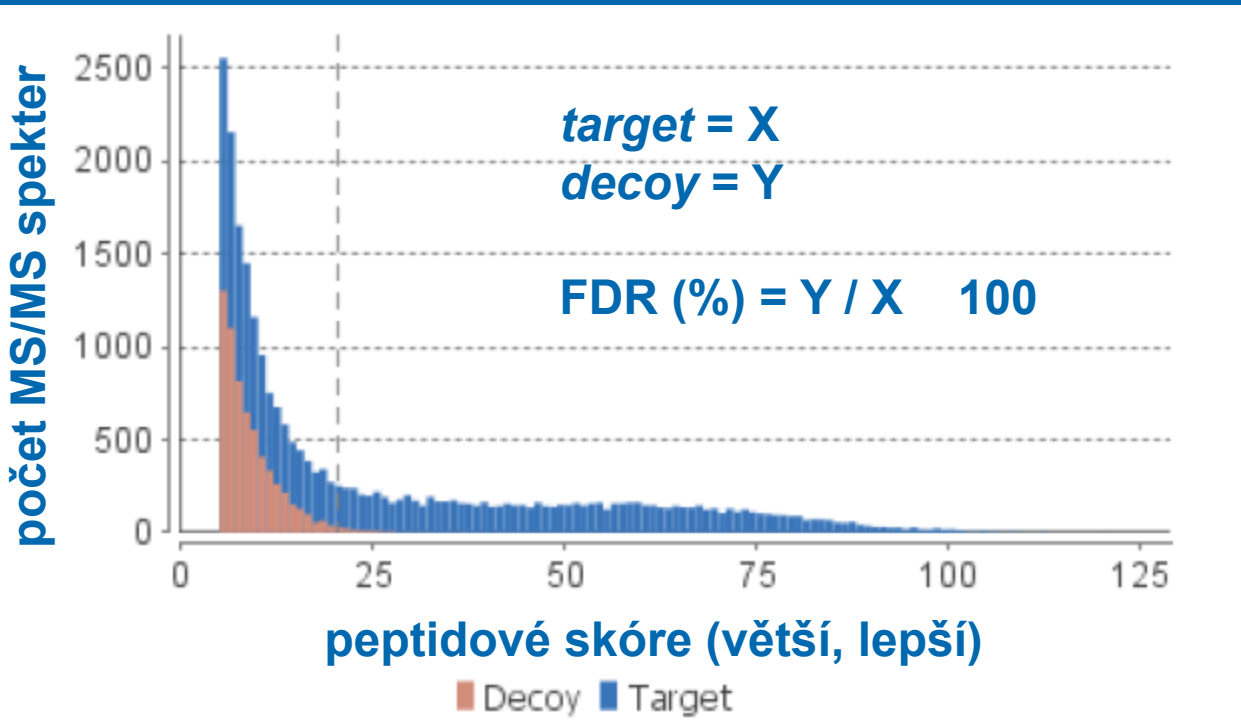


Zpracování LC-MS/MS dat (2)

- falešná pozitivita a negativita ve výsledcích databázového hledání
- *decoy* proteinová databáze a FDR (*false discovery rate*)
  - *decoy* databáze – např. obrácené sekvence, náhodné sekvence proteinů
  - identifikace peptidů v cílové (např. TAIR) i *decoy* proteinové databázi

⇒ **jeden z možných přístupů jak určit FDR – peptidová úroveň**

		The MS/MS spectrum comes from a peptide sequence in the database	
		True	False
Search reports a match to the correct sequence	True	True positive	False positive
	False	False negative	True negative



Zpracování LC-MS/MS dat (3)

- **z identifikovaných peptidů k proteinům přítomným ve vzorku (*protein inference problem*)**
  - problém u ***bottom-up*** přístupu (digesce proteinů, analýza až peptidů)
  - v MS analýze **vidíme jen část** z např. tryptických **peptidů** proteinů (max. kolem 60-70% sekvenčního pokrytí proteinu, min. 1 peptid na protein) a navíc nevíme ze kterých proteinů pozorované peptidy původně pochází...

    ⇒ **problém s určením seznamu proteinů přítomných ve vzorku**

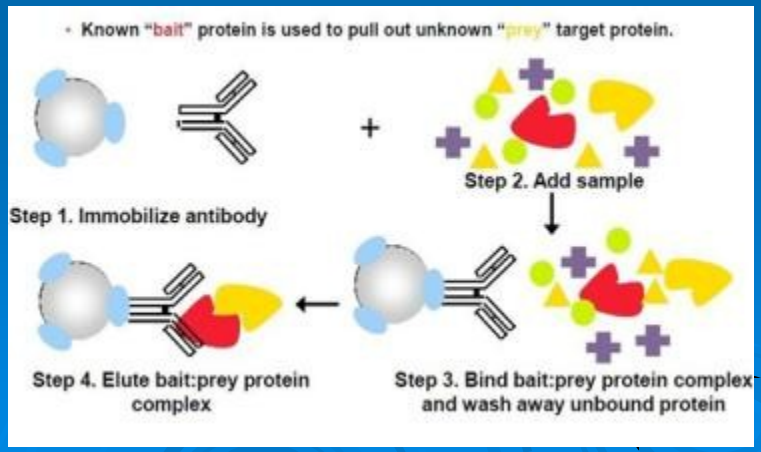
    (sadě peptidů může odpovídat více proteinů – isoformy, sekv. homology; proteiny identifikované jen na jeden peptid?)
  - **peptid může teoreticky pocházet z jednoho i více proteinů**

Pohled na seznamy identifikovaných proteinů

- dva seznamy identifikovaných proteinů v našem IP experimentu
  - vzorek po IP experimentu s naším proteinem – sada proteinů A
  - slepý vzorek; „bead proteome“ – sada proteinů B

- co nás zajímá v našem IP experimentu nejvíce?

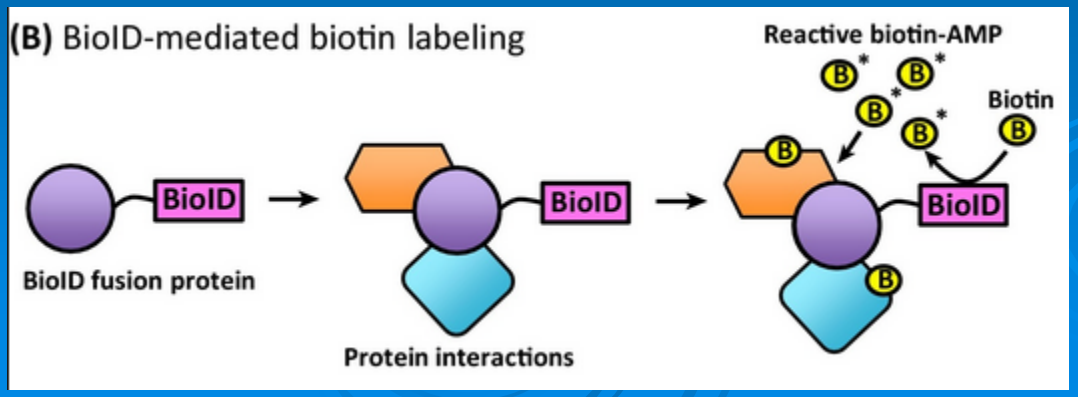
- sada proteinů A, které zároveň nejsou či jsou významně méně zastoupeny v sadě proteinů B



https://www.creative-proteomics.com

Pohled na seznamy identifikovaných proteinů (2)

- **proteiny „navíc“ v A**
  - **1) kvalitativní změny (A: „ano“, B: „ne“)**
    - citlivost použitého přístupu...
    - proteiny identifikované relativně slabě v A mohou být v B také přítomny!
  - **2) kvantitativní změny (A: „více“, B: „méně“)**
    - problém s nespecificky se vázanými proteiny – nevážou se sami, nesou si sebou i své interakční partnery!!!
      - BioID alternativa k IP experimentu tento problém nespecifických interakcí elegantně řeší



Co se seznamem proteinů „navíc“? – vybrané možnosti

1. **manuální prohledání dostupných informací v literatuře**

2. **www.UniProt.org** (ID mapping; informace, další databáze; GO, *pathways*)
3. **STRING https://string-db.org**
4. **DAVID http://david.abcc.ncifcrf.gov/home.jsp**
5. **PANTHER http://go.pantherdb.org/**
6. **R/Python/Jupyter**
7. **Cytoscape**

- **...**

Děkuji za pozornost

Minitest

1) (Vyberte dle Vás nejsprávnější tvrzení) Co je to bioinformatika?

A. disciplína plná *blackbox* nástrojů
B. studium a aplikace metod pro uchování, zpětné vyvolání a analýzu biologických dat
C. sada nástrojů, které pravděpodobně nebudu nikdy potřebovat, a tudíž je tato odpověď správnější než ta správná...
D. žádná z uvedených možností

2) (Vyberte správné tvrzení) Evoluce proteinů

A. se děje replikací, duplikací a delecí jednotlivých aminokyselin
B. se zastavila před přibližně 100 miliony let
C. se u vícedoménových proteinů děje také replikací, duplikací či delecí jednotlivých domén
D. se ani nepřímo nevyužívá při studiu sekvenčně homologních proteinů k predikci očekávané funkce nově objevených proteinů

3) (Vyberte NEsprávné tvrzení) Při analýze biologických sítí jako KEGG, Reactome, Gene Ontology je třeba

A. uvažovat relativně malou míru falešné pozitivity a negativity z důvodu naší téměř kompletní znalosti probíhajících procesů v živých systémech
B. uvažovat, že většina informací není manuálně ověřená (*Gene Ontology*)
C. uvažovat, že většina informací je manuálně ověřená (*KEGG*, *Reactome*)
D. přistupovat k výsledkům kriticky a optimálně ověřovat zdrojová data pro klíčové závěry z analýzy sítí

4) (Vyberte dle Vás nejhorší řešení uvedeného zadání) Jsem nucen(a) zjistit primární sekvenci pro řádově stovky proteinů pro které mám UniProtKB identifikátor:

A. najdu si každý protein v UniProtKB databázi, doklikám se k jeho sekvenci, kterou si uložím a následně všechny sekvence zkombinuji
B. napíši si skript/program pro stažení primární sekvence na základě proteinového identifikátoru – bude mi to možná trvat déle, ale potřeboval jsem to již několikrát...
C. poprosím kolegu, zda-li nemá/nezná nástroj, kterým bych tento úkol zvládl sám případně s jeho pomocí automaticky
D. najedu si na stránky UniProtKB a pomocí nástroje *Retrieve/ID Mapping* si potřebnou informaci získám

5) Vyjmenuj alespoň 3 bioinformatické nástroje, které jsi během svého studia použil nebo se chystáš použít a pro jaké účely (případně alespoň ty, o kterých jsi slyšel). Pokud se jedná o nástroje, o kterých by měli podle Tebe vědět i ostatní, podtrhni je, prosím (budou zváženy pro zmínění dalším ročníkům...)

6) Zazněla (a pokud ano, zmiň pro Tebe ty nejpodstatnější) v rámci některé z přednášek „Bioinformatika a proteiny“ obecná informace, která Ti rozšířila obzory, případně Ti pomohla k samostatnějšímu a kritičtějšímu (konstruktivně) uvažování nad věcmi, které během studia děláš?