11. 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. Několik zamyšlení závěrem
10. Příklad využití bioinformatických nástrojů

5. Biologické sítě

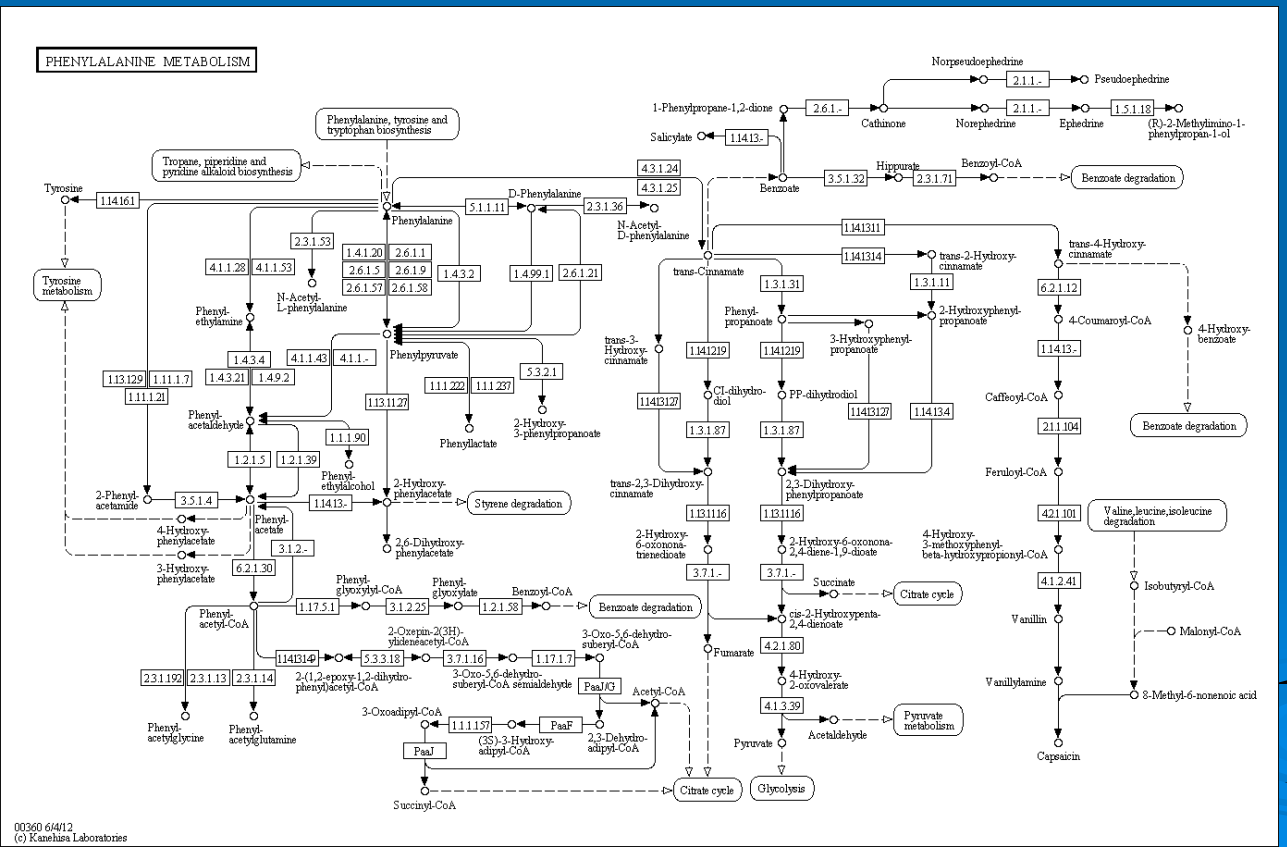
Biologické sítě

- **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 = sada molekul** (např. proteinů, geny, metabolity = *nodes*) **propojených pomocí definovaných, funkčních vztahů** (např. protein-protein interakce = *edges*)

Biologické sítě – příklady

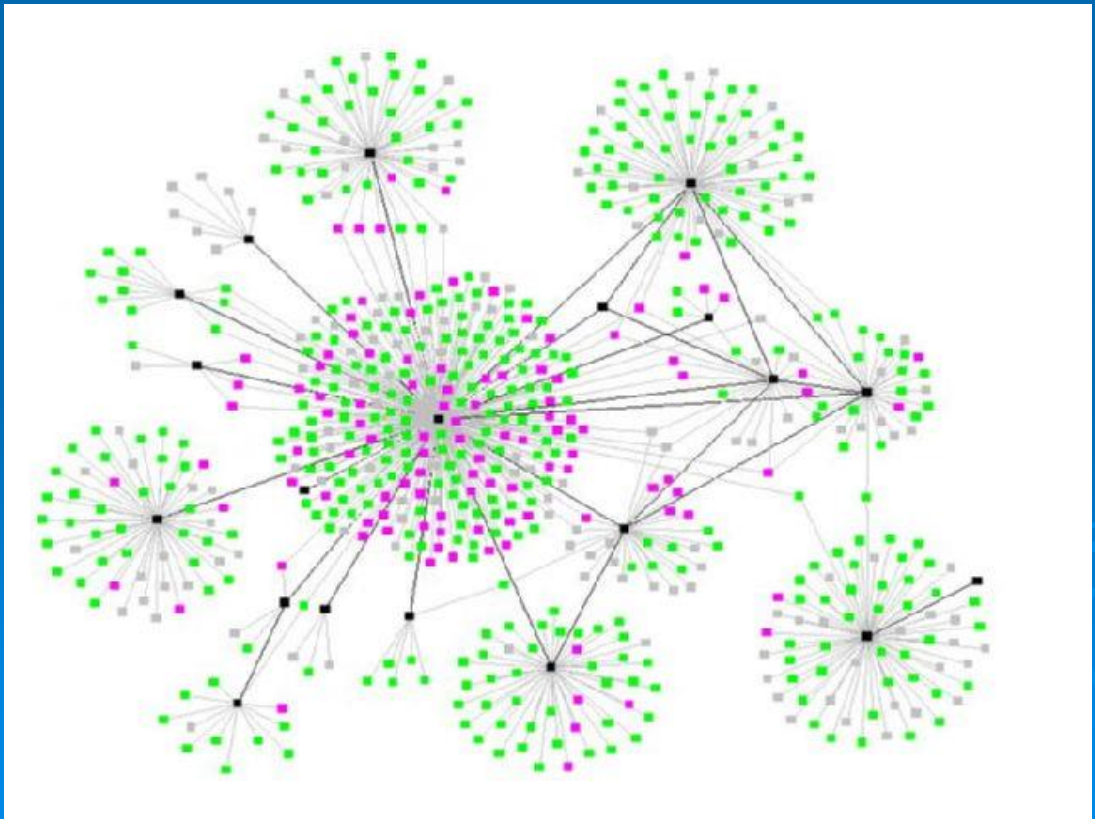
- **metabolické dráhy (*metabolic pathways*)**
  - spojují proteiny (*nodes*) skrze produkty a reaktanty (*edges*)
    - **produkt jednoho = substrát druhého**
  - **např. KEGG; WikiPathways**



část metabolické sítě – metabolismus Phe (KEGG)

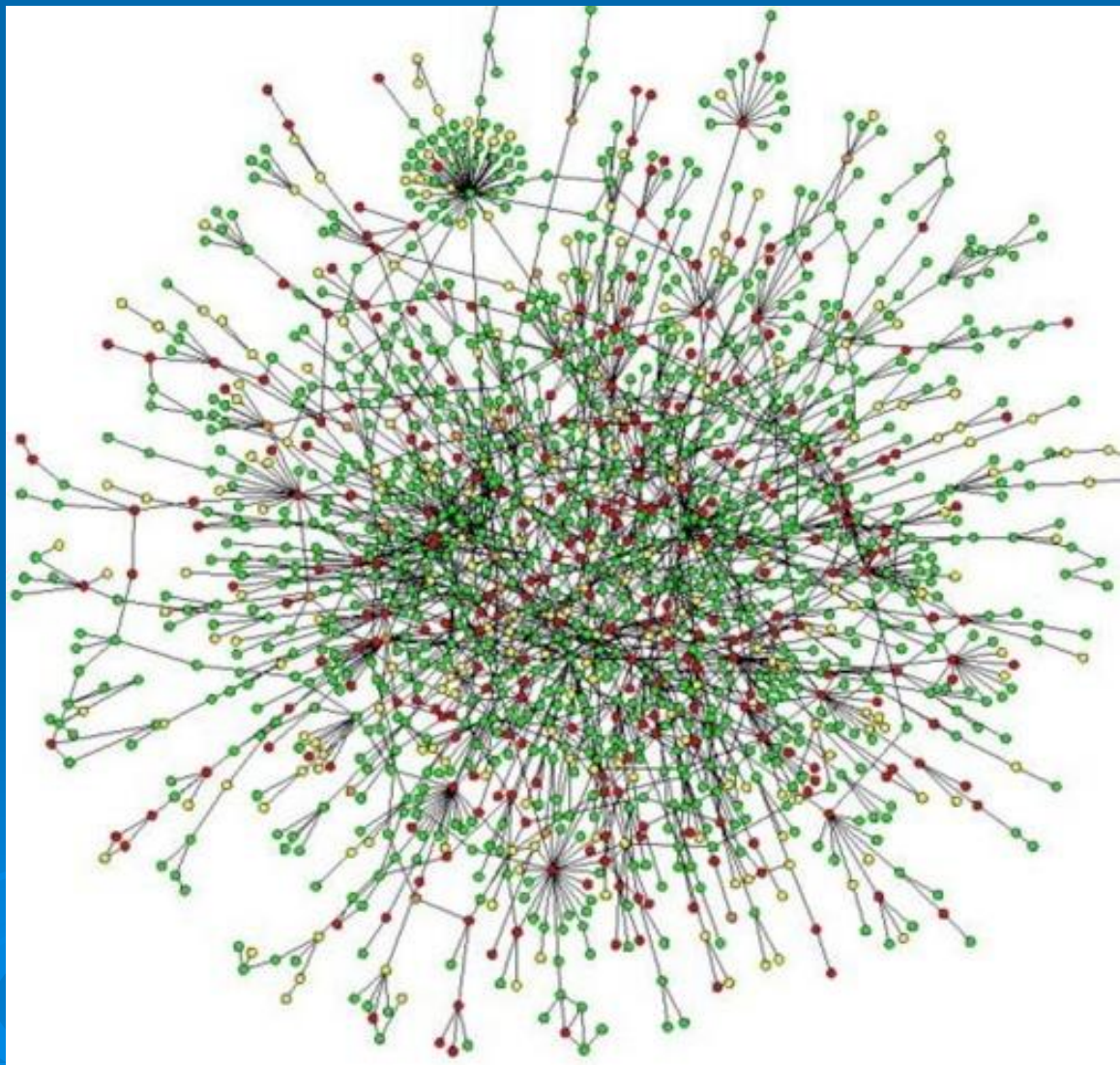
Biologické sítě – příklady (2)

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



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
    - ...



Biologické sítě – příklady (4)

- **proteiny anotované do stejné kategorie genové ontologie (GO) – viz. dále**
  - **proteiny přítomné ve stejné GO kategorii mají společné např.**
    - **a) lokalizaci v buňce (**např. jaderná membrána**)**
    - **b) molekulární funkci (**např. vazba iontů Ca**)**
    - **c) biologický proces, ve kterém participují (**např. metabolismus Ca)

(dle příslušného „GO termínu“, vysvětleno dále...)

6. Biologické sítě

Biologické ontologie, KEGG

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)...**
  - průběh buněčného dělení (*Cell Cycle Ontology*)
  - vývoj rostliny *A. thaliana* (*Arabidopsis development*)
  - **stále živý proces úprav/oprav/doplnění ontologií; není statické**

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

Genová ontologie (GO)

- **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/ + AmiGO prohlížeč (online, offline)**

Genová ontologie (GO) (2)

- kde se berou data pro GO?
  - každá anotace obsahuje informaci o svém původu – *evidence code*

  - A) manuálně přiřazené správcem (*curator*)
    - *experimental evidence codes* ⇒ z reálného experimentu
    - *computational analysis evidence codes* ⇒ z *in silico* analýzy
    - *author statement evidence codes* ⇒ tvrzení autora + citace
    - *curatorial statement codes* ⇒ tvrzení správce, nepatří výše...

    (všechny kategorie se dále dělí...)

  - B) automaticky přiřazené (bez zásahu správce)
    - *automatically-assigned evidence code*
      - *Inferred from Electronic Annotation* (IEA)

Společné znaky anotovaných proteinů v ontologiích

- srovnání proteinů na základě jejich anotace v rámci dané ontologie, např. v GO – sémantická podobnost
  - jak jsou si proteiny blízké z pohledu jejich zařazení do GO termínů
  - většinou používána

- srovnání proteinů na základě jejich funkce – funkční podobnost
  - např. proteiny se stejnou funkcí lokalizované v jiné části buňky
  - z pohledu GO termínů mohou být relativně vzdálené

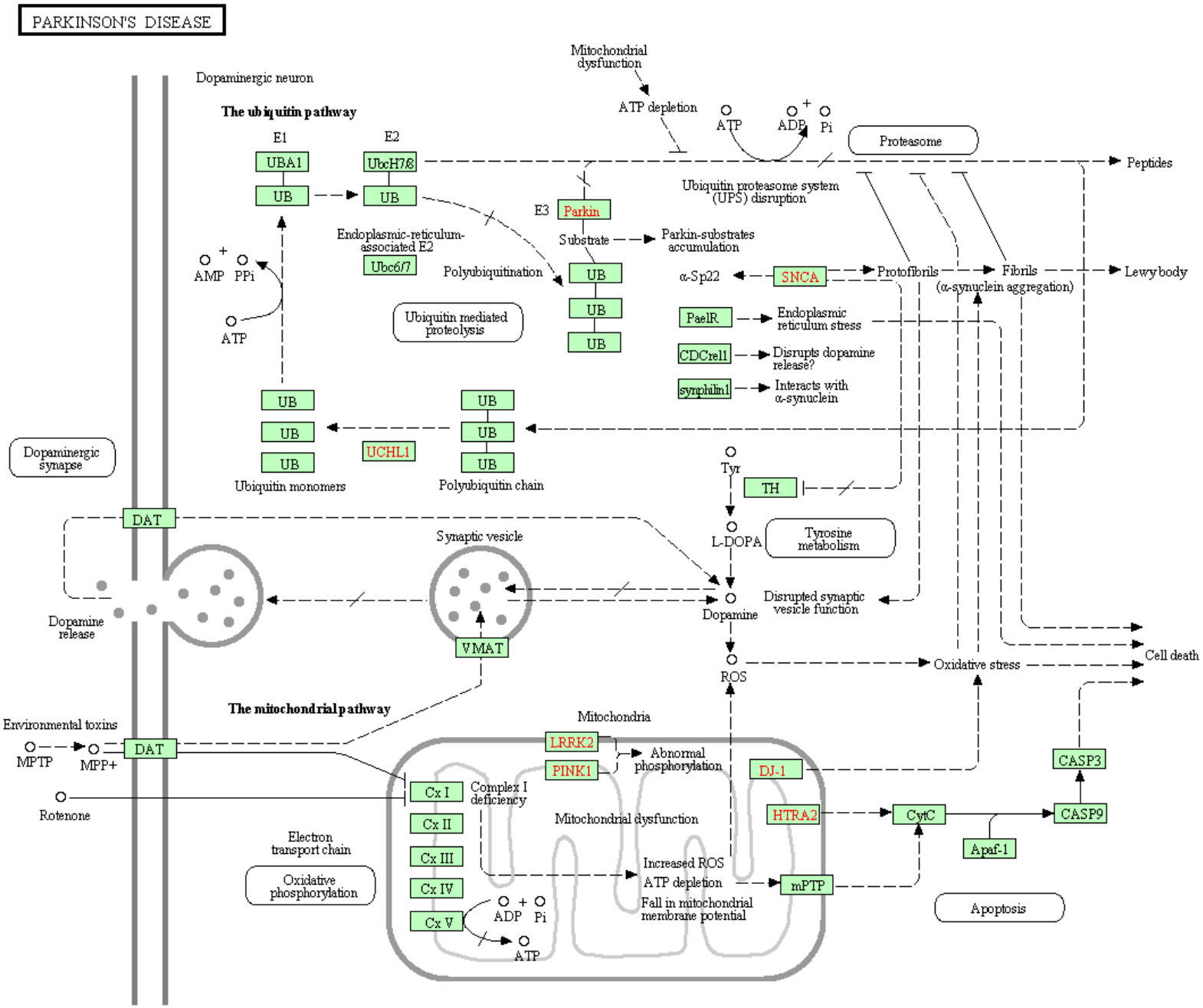
- stejně jdou srovnávat společné znaky a blízkost i např. GO termínů

KEGG

- KEGG = *Kyoto Encyclopedia of Genes and Genomes*
- http://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 – KEGG Pathway

- obdobně i např. http://www.rectome.org



7. Biologické sítě

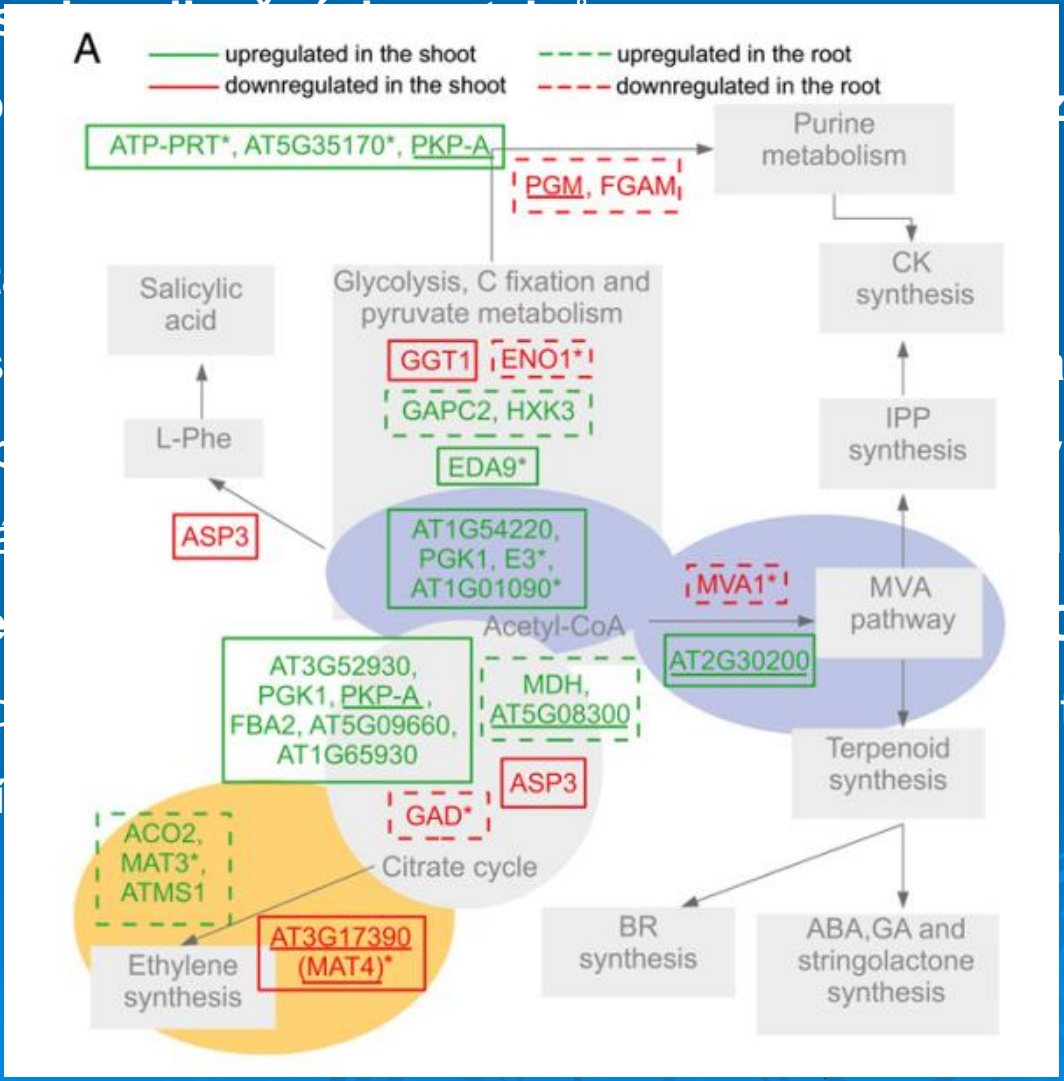
Příklady použití

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?**
  - objevili jsme i jiné, dosud nepotvrzené, ale související metabolické dráhy?
- **jsou známy proteinové komplexy mezi nalezenými proteiny?**
  - dokáží nám tyto pomoci při interpretaci vlivu látky na rostlinu?
- **je mezi proteiny zastoupeno více proteinů z konkrétního GO termínu?**
  - na základě daných GO termínů je možno odvodit souvislosti s funkcí či lokalizací probíhajících (i sekundárních) dějů

Vliv nízkomolekulární látky na rostlinu

- **identifikace s**
- **jsou tyto pro** **ze? (KEGG)**
  - fungoval
- **jaké jiné met**
  - objevili js **bolické dráhy?**
- **jsou známy p** **?**
  - dokáží ná **u?**
- **je mezi prote** **D termínu?**
  - na základ **ti s funkcí či** lokalizací



Interakční partneři zvoleného proteinu

- **vidíme již známé interakční partnery?**
  - pozitivní kontrola průběhu experimentu
- **nově pozorované interakce**
  - **předpoklad možných interakčních partnerů**
    - vhodně provedená negativní kontrola!
    - několik biologických replikátů pro vzorek i kontrolu!
  - **studium biologických vlastností možných interakčních partnerů (GO termíny, metabolické dráhy, ...)**
    - **zapadají tyto do již známých informací o funkci, lokalizaci aj. zvoleného proteinu?**
    - **je možné predikovat nepotvrzenou funkci proteinu?**
    - **jsou patrné souvislosti s lokalizací našeho proteinu?**

Studium proteinu, se vztahem k onkol. 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

„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/vznikem nemoci v konkrétním místě organizmu?

- **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
  - důvodem nedostatečná citlivost/specifita současných přístupů pro studium

- **mnoho dat v databázích z predikčních studií**
  - i přes kontrolu nemusí zcela odpovídat zdrojovým datům a skutečnosti
  - někdy lze vyloučit z analýzy (např. automaticky anotované GO...)

- **stále víme velmi 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**
- **ověřením 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!**
- **není důležité jak to vypadá, ale co a jak se z toho dá vyčíst...**

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**
  - **„mapování“ 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 (~98 %) 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 sekv. homologií**
  - **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**

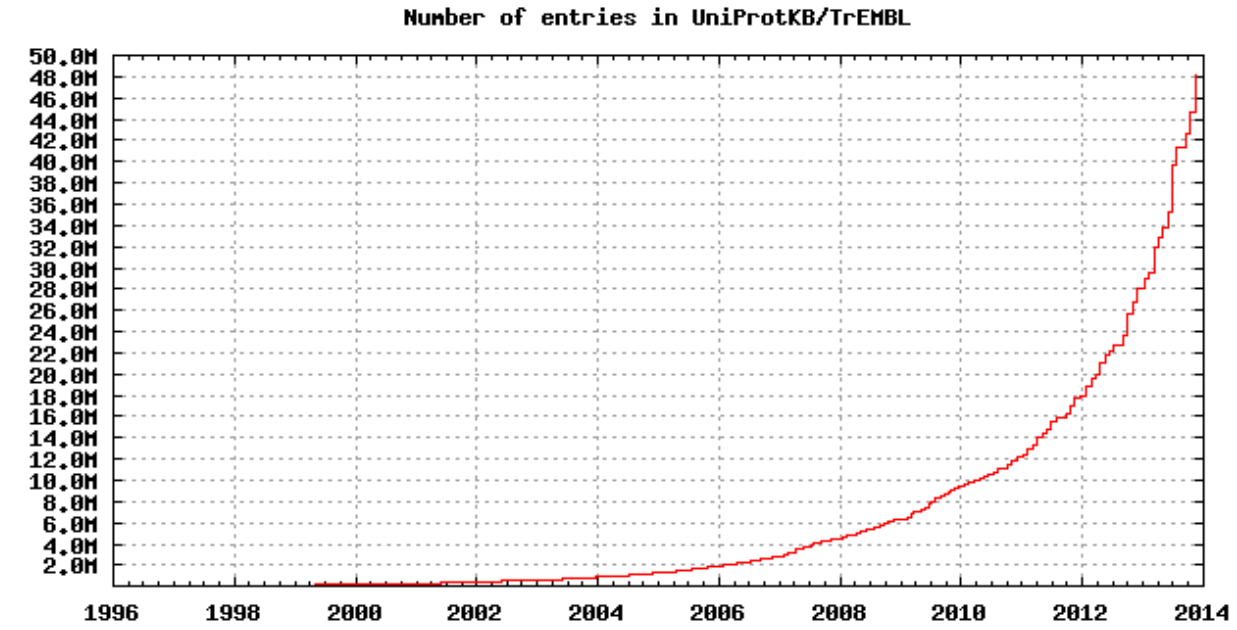
Universal Protein Resource (UniProt) (3)

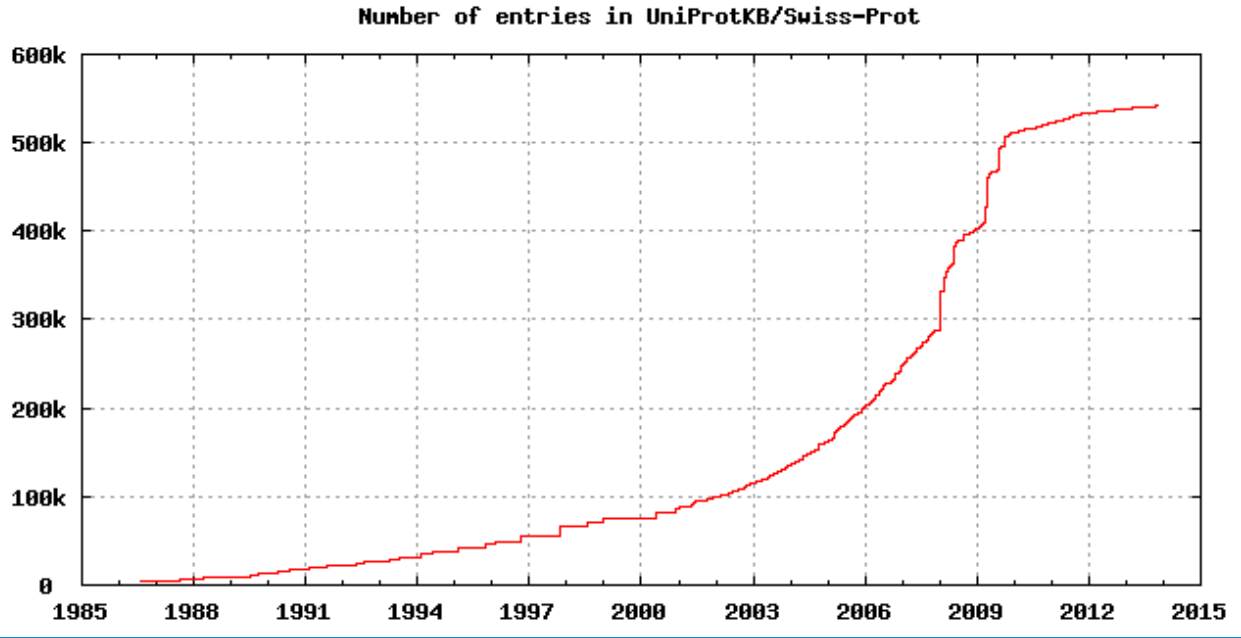
- UniProtKB/SwissProt – manuální zpracování (*curation*) správcem
  - kontrola sekvence – není-li v původní sekvenci chyba
  - sekvenční analýza – manuálně kontrolované predikce atd.
  - studium literárních zdrojů – dodány biologicky relevantní informace k proteinu na základě dostupných publikací; název 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 které 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 opravě správcem (*curation*)**
  - ***Complete Proteome Set* – pro kompletně sekv. organizmy (T+S)**
  - ***Reference Proteome Set* – vybrané modelové organizmy (T+S)**

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.

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.**

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**

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; ...**

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

Reactome

- http://www.reactome.org
- obdoba KEGG, převážně pro lidské dráhy

Pax-DB

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

9. 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, Expasy**)**
  - 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)

- **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
- **bioinformatické aplikace/databáze není možné nevyvíjet/neaktualizovat**
  - při vytváření nástroje/databáze nutno počítat s udržitelností jeho vývoje...
- **školící programy/workshopy/stáže v bioinformatických centrech**
  - EBI, SIB aj.
- **význam spoluprací – jeden tým často nedokáže pojmout celé spektrum použitých nástrojů, přístupů včetně interpretace výstupů**

10. Příklad využití bioinformatických nástrojů

Zavedení problému

- **studium proteinových komplexů vybraného proteinu**

- **imunoprecipitace proteinových komplexů (*IP experiment*)**
  - protilátka proti proteinu (*bait*), u kterého chceme zjistit jeho partnery
  - dva hlavní přístupy
    - protilátka imobilizovaná např. na kuličkách (magnetické, v kolonkách)
    - protilátka se přidá do volného roztoku a následně „lovíme“ protilátku
  - vhodné opakovat experiment s různými protilátkami (epitopy, stérické zábrany, potvrzení předchozích experimentů)
  - nativní prostředí při experimentech – podmínky pro interakce jako *in vivo*
  - výstupem *pull-down* roztoky – proteiny vázající se na *bait* a nesp. proteiny
  - paralelně experimenty bez *bait* – negativní kontrola – *bead proteome*
  - minimálně 3 biologické replikáty, lépe 5 od vzorku i negativní kontroly

LC-MS/MS analýza *pull-down* vzorků

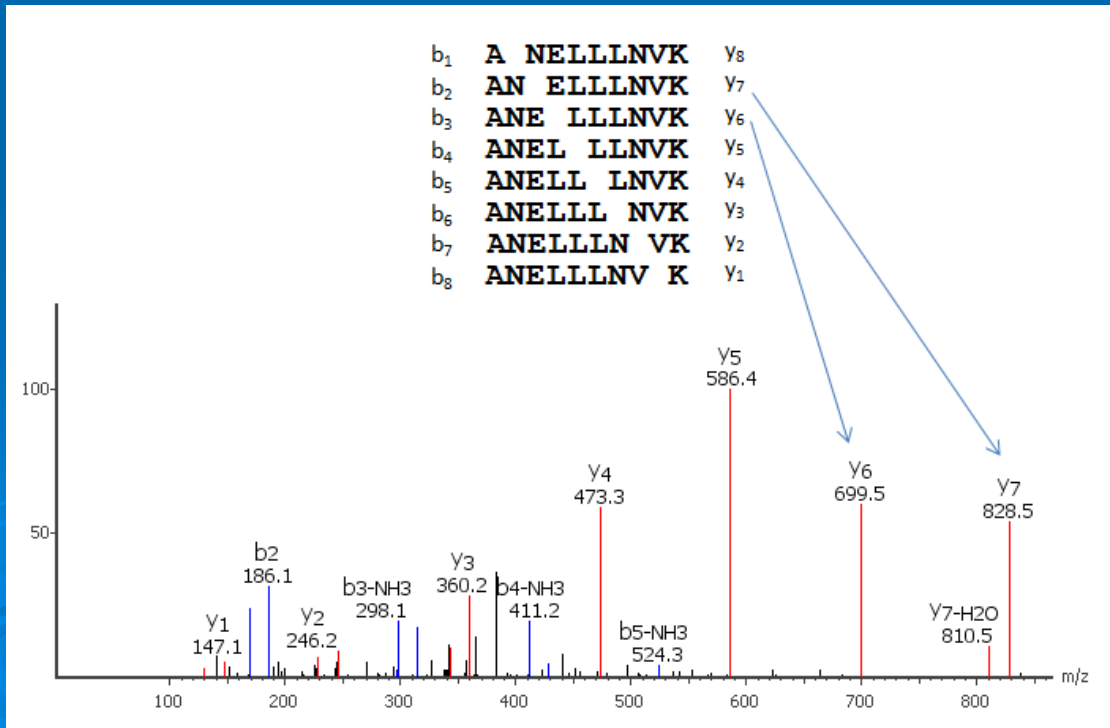
- **digesce proteinů ⇒ peptidy (**např. trypsinem; peptidy končí R nebo K**)**
- **LC-MS/MS analýza směsi peptidů**
  - **peptidy vstupují do MS v pořadí rostoucí hydrofobicity (**LC separace**)**
- **MS zjistí MW peptidů a získá MS/MS spektra**

  (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“**)**



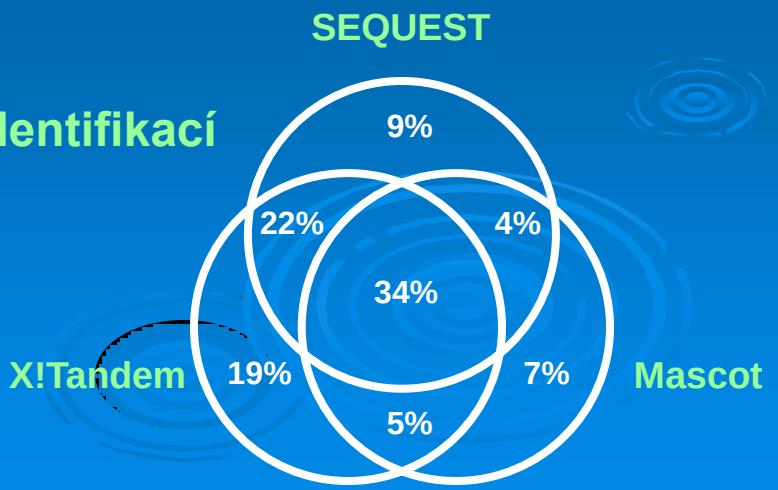
Zpracování LC-MS/MS dat

- **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

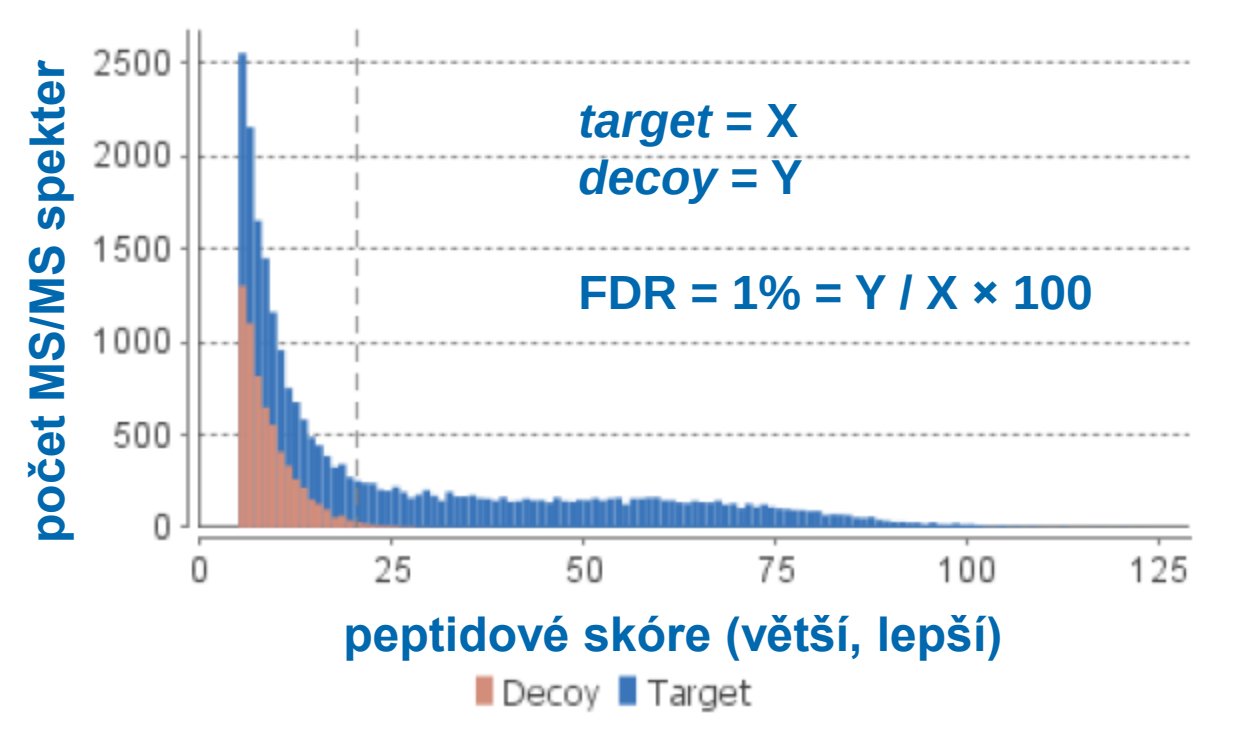
⇒ zvýšení počtu pozitivních identifikací



Zpracování LC-MS/MS dat (2)

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

⇒ jeden z možných přístupů jak určit FDR – peptidová úroveň



Zpracování LC-MS/MS dat (3)

- **z identifikovaných peptidů k proteinům přítomným ve vzorku**
  - problém u ***bottom-up*** přístupu (digesce proteinů, analýza až peptidů)
  - v MS analýze **vidíme jen malou část** z např. tryptických **peptidů** proteinů 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?)

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 v sadě proteinů B

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“)**
    - možno pracovat pouze s **intenzitami A a B peptidů** – „**label-free**“
      - přesnost, správnost?
    - vzorky **A** a **B** byly zpracovány tak, že jsme pomocí MS schopni rozlišit mezi **A** a **B** (např. **SILAC** – „**S**table **I**sotope **L**abeling by **A**mino acids in **C**ell Cultures“ – komplikované u rostlin, nekompletní inkorporace značených AA; **dusík ^{15}N**)

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) DAVID http://david.abcc.ncifcrf.gov/home.jsp**
- **4) PANTHER http://go.pantherdb.org/**
- **5) ANAP http://gmdd.shgmo.org/Computational-Biology/ANAP**
  - **jen pro At**
  - **Source database – čerpá známé informace z databáze interakcí**
  - **Detection method – predikce možných protein-protein interakcí**

    (u predikované interakce uvádí důvod pro predikci**)**

- **6) Cytoscape**
- **....**

Děkuji za pozornost