11. Bioinformatika a proteiny II

David Potěšil

Core Facility – Proteomics

CEITEC-MU

Masaryk University

Kamenice 5, A2

phone: +420 54949 7304

email: david.potesil@ceitec.muni.cz

Obsah přednášky

1. Co je to bioinformatika?
2. Evoluce proteinů, proteinové domény
3. Biologické sítě
4. Příklad využití bioinformatických nástrojů

1. Co je to bioinformatika?

Co představuje „bioinformatika“?

- **vícero názorů...**[1]
  - **Bioinformatics is conceptualizing biology in terms of macromolecules** (in the sense of physical-chemistry) **and, then, applying “informatics” techniques** (derived from disciplines such as applied math, computer science, and statistics) **to understand and organize the information** associated with these molecules, on a large scale. (Luscombe, 2001, p. 346)
  - **The mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information**. (Tekaia, n.d.)
  - **Bioinformatics is the field of science in which biology, computer science, and information technology merge to form a single discipline.** The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned. (National Center for Biotechnology Information, n.d.)
  - **Computational biology is not a “field”, but an “approach” involving the use of computers to study biological processes** and hence it is an area as diverse as biology itself. (Schulte, n.d.)
  - **Biomedical informatics is the science underlying the acquisition, maintenance, retrieval and application of biomedical knowledge and information to improve patient care, medical education and health sciences research**. (Friedman, n.d.)

1. Fenstermacher, D. Introduction to bioinformatics. *Journal of the American Society for Information Science and Technology* **56, 440–446 (2005).**

Co představuje „bioinformatika“? (2)

- „The enormous amount of data gathered by biologists – and the need to interpret it – requires tools that are in the realm of computer science. Thus, bioinformatics.“ [2]

- **studium metod pro uchování, zpětné vyvolání a analýzu biologických dat**
  - **sekvence nukleových kyselin (NK) a proteinů**
  - **proteinové struktury**
  - **funkce proteinů**
  - **metabolické a regulační dráhy („pathways“)**
  - **molekulární interakce** (např. protein-protein, protein-NK, NK-NK)

1. Cohen, J. Bioinformatics - an introduction for computer scientists. *ACM Comput. Surv.* **36, 122–158 (2004).**

Příbuzné disciplíny

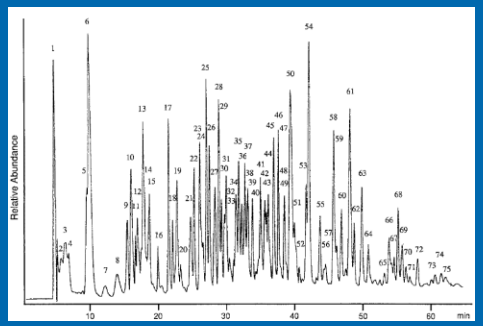
- **„data mining“**
  - analýza dat z různých perspektiv a „dolování“ shrnujících (zobecněných) informací
- **matematická a teoretická biologie**
  - matematická reprezentace, zpracování a modelování biol. procesů
- **lékařská informatika**
  - tvorba databáze medicínských informací a jejich další využití
- **biostatistika**
  - aplikace a vývoj statistických metod pro řešení biologických a klinických problémů

- **častý překryv s těmito i s dalšími obory (**záleží na konkrétní aplikaci**)**

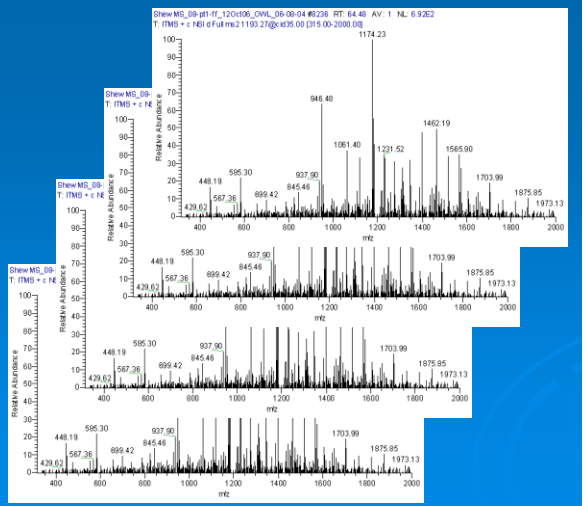
Příklad využití bioinformatických nástrojů

- **protein-protein interakce založené na datech z hmotnostní spektrometrie spojené s kapalinovou chromatografií** (LC-MS analýza peptidů)

LC-MS záznam

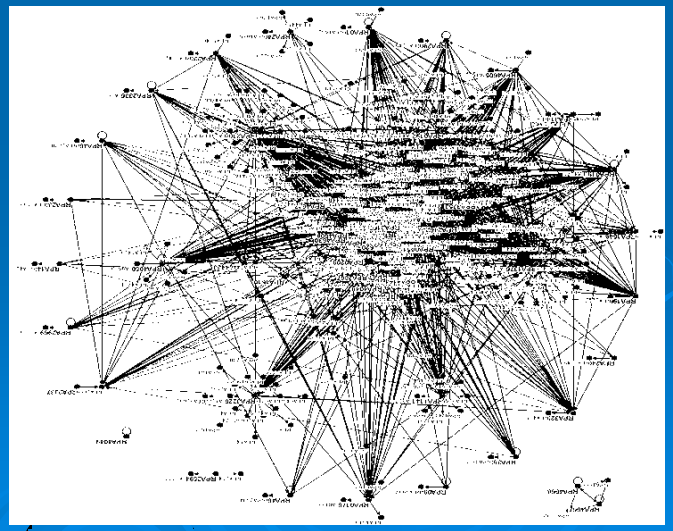


MS/MS spektra



„black box“

„standardní nastavení“

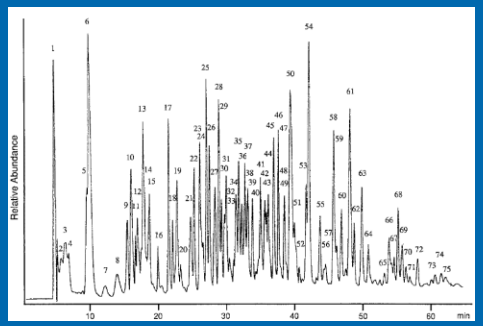


protein-protein interakční síť ?
závěry z analýze této sítě?

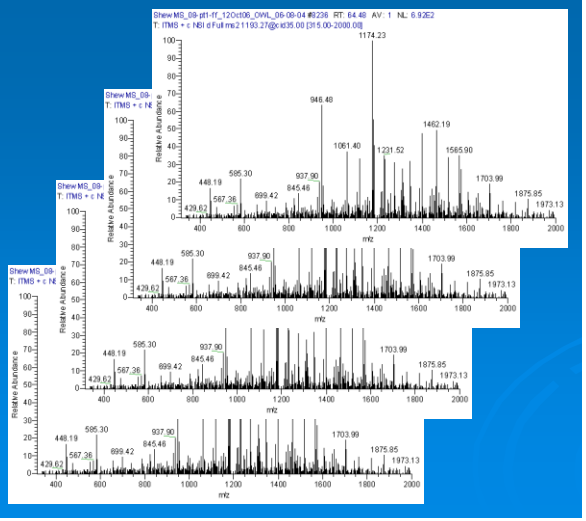
Příklad využití bioinformatických nástrojů

- **protein-protein interakce založené na datech z hmotnostní spektrometrie spojené s kapalinovou chromatografií** (LC-MS analýza peptidů)

LC-MS záznam

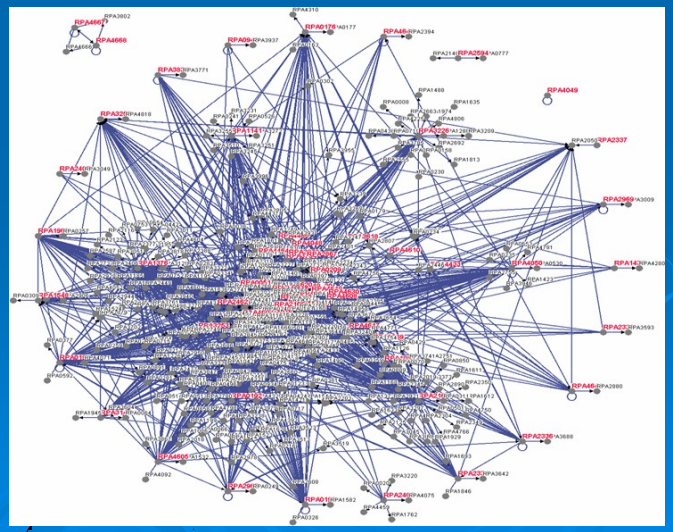


MS/MS spektra



„white box“

výstupům přizpůsobené nastavení



protein-protein interakční síť

analýza sítě: úloha proteinu A z jeho interakcí

2. Evoluce proteinů, proteinové domény

Jedna z prvních aplikací bioinformatiky

– srovnání primárních sekvencí (homologie)

- **BLAST – Basic Local Alignment Searh Tool**

- **proč srovnávat primární sekvence?**
  - podobnost v primární sekvenci proteinů

    odobnost ve struktuře proteinů

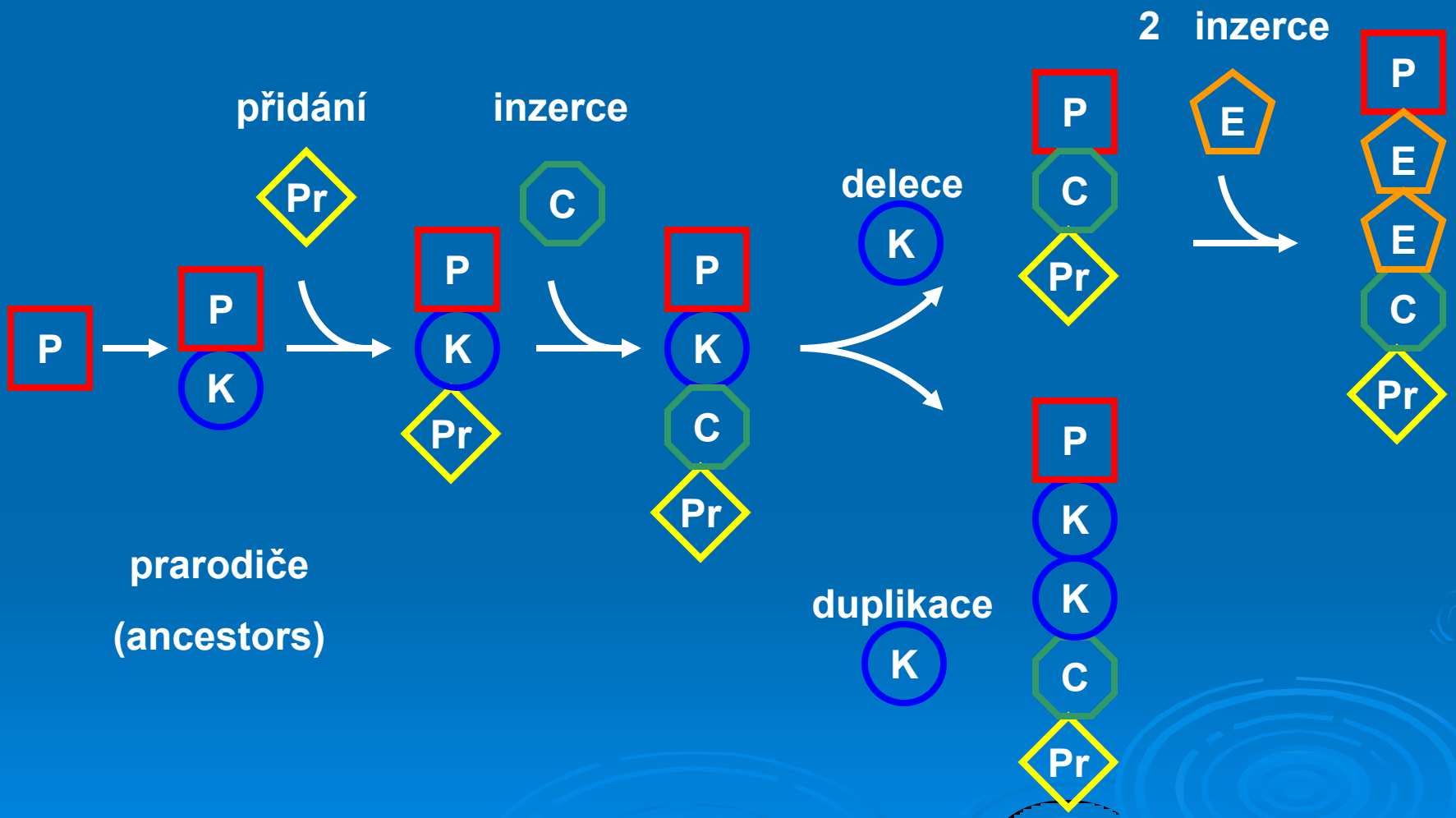
    ⇒ **podobnost ve funkci proteinů...**

Není tak jednoduché...

Proteinová evoluce a proteinové domény

- **proteinová doména = nezávislá strukturní, funkční a evoluční jednotka**
- **2/3 proteinů jednobuněčných a 80% proteinů mnohobuněčných organizmů je složených z více domén**
- **vznik „nových“ proteinů (proteinová, molekulární evoluce)**
  - **kombinace, duplikace, divergence stávajících domén (na úrovni genů)**
  - **kombinace/duplikace/změna domén často odlišná funkce proteinu**
    - změna struktury, spolupráce se sousedními doménami...
      - jednodoménové proteiny, stejná doména: ~67% šance na podobnou funkci
      - dvoudoménový protein, 1 stejná doména: ~35% šance na podobnou funkci
  - **v průběhu evoluce dále nastávaly mutace v duplikovaných či zkombinovaných doménách často se zachováním strukturní podobnosti sekvenčně odlišné, strukturně podobné**

Proteinová evoluce a proteinové domény – příklad



proteinová evoluce v čase a událostech

Doménové superrodiny a rodiny („superfamilies“, „families“)

- **proteinové domény je možné klastrovat na základě podobnosti**
- **podobnost možná na více úrovních**
  - **sekvenční podobnost (**primární struktura proteinu/domény**)**
  - **strukturální podobnost (**sekundární a terciární struktura proteinu/domény**)**
  - **funkční podobnost (**nezávislá na sekvenční a strukturní podobnosti**)**

- **doménové superrodiny a rodiny a podobnost**
  - **strukturní, funkční** podobnost ⇒ **doménová superrodina**
    - stejní proteinoví prarodiče, evolučně starší (mutace sekvence)
  - **sekvenční** podobnost ⇒ **doménová rodina**
    - evolučně mladší

Hlavní zdroje pro klasifikaci domén

- klasifikace domén do superrodin a rodin

- SCOP („Structural Classification Of Proteins“)
  - http://scop.mrc-lmb.cam.ac.uk/scop/

- CATH („Class, Architecture, Topology, Homologous Superfamily“)
  - http://www.cathdb.info/

- čerpají známé proteinové sekvence z Protein Data Bank (PDB)
- zpracovávanou jednotkou je proteinová „doména“

Proteinové rodiny a superrodiny

- **obdobně jako u proteinových domén**
  - **častější klastrování na základě „sekvenční podobnosti“** (převážně „multiple sequence alignment“ algoritmy) **sequence signatures“**
  - **využití primárních sekvencí proteinů ve zvolené databázi**
- **při klastrování je možno zvažovat různé části proteinu**
  - funkční místa proteinu
  - funkční konzervativní motivy
  - funkční domény
  - strukturní domény

- **proteinová rodina = „sekvenčně podobné“ proteiny**
- **proteinová superrodina = evolučně spjaté proteinové rodiny** (není nutná sekvenční podobnost) **– souhrn proteinů v evolučně spjatých prot. rodinách**

Proteinové rodiny a superrodiny – online zdroje

- **různé databáze proteinových rodin a superrodin** (viz. dále)
  - **používají různé cílové proteinové databáze (primární sekvence)**
    - UniProtKB (SwissProt a TrEMBL)
    - NCBI RefSeq
    - proteinové databáze pro vybrané kompletně osekvenované orgaizmy
    - ...
  - **používají různé části proteinu pro predikci rodin/superrodin**
- **integrální zdroje**
  - sbírají informace z více zdrojů a prezentují na jediném místě
  - **InterPro**
  - **CDD**

Bioinformatic tool/URL	Database source	Clustering method	Cluster information based on	Protein families or signatures
Signature databases				
ProtClustDB Dec 2 2010/ http://www.ncbi.nlm.nih.gov/proteinclusters	NCBI RefSeq	Clique based	Functional domains	627757, 10885 (curated)
Pfam 25.0/ http://pfam.sanger.ac.uk/	UniProtKB	HMMs	Functional domains	12273 (Pfam-A)
PROSITE 20.68/ http://expasy.org/prosite/	UniProtKB	Patterns, profiles	Functional sites	1598
PRINTS 41.1/ http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/index.php	UniProtK	Fingerprints	Functional conserved motifs	2050
ProDom 2006.1/CG267/ http://prodom.prabi.fr/prodom/current/html/home.php	UniProtKB/267 completed genomes (one from plants)	MKDOM2	Functional domains	574656/301126
SMART 6.1/ http://smart.embl-heidelberg.de/	UniProtKB/760 completed genomes (one from plants)	HMMs	Functional domains	895
TIGRFAMs 10.0/ http://www.jcvi.org/cms/research/projects/tigrfams/overview/	UniProtKB	HMMs	Functional domains	4025
PIRSF 2.74/ http://pir.georgetown.edu/pirwww/dbinfo/pirsf.shtml	UniProtKB	HMMs	Functional domains	3248 (curated)
SUPERFAMILY 1.75/ http://supfam.cs.bris.ac.uk/SUPERFAMILY/	1452 completed genomes (27 from plants)/UniProtKB/PDB	HMMs	SCOP domains	2019
GENE3D 10.0.0/ http://gene3d.biochem.ucl.ac.uk/Gene3D/	1867 completed genomes	HMMs	CATH domains	2549
PANTHER 7.0/ http://www.pantherdb.org/	48 completed genomes (three from plants)	HMMs	Functional domains	6594
Integrative signature databases				
InterPro 31.0/ http://www.ebi.ac.uk/interpro/	UniProtKB	Signature integration	Gene3D, HAMAP, PANTHER, Pfam, PIRSF, PRINTS, ProDom, PROSITE, SMART, SUPERFAMILY, TIGRFAMs signatures	21185
CDD 2.26/ http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml	NCBI Database	PSSMs	NCBI-curated domains, Pfam, SMART, COGs, ProtClustDB signatures	41593

A4

1. Chothia, C. & Gough, J. Genomic and structural aspects of protein evolution. *Biochemical Journal* 419, 15 (2009).

Fylogenetická podobnost

- fylogenetika
  - studuje evoluční vztah organizmů
  - používá molekulární sekvenování a morfologická data

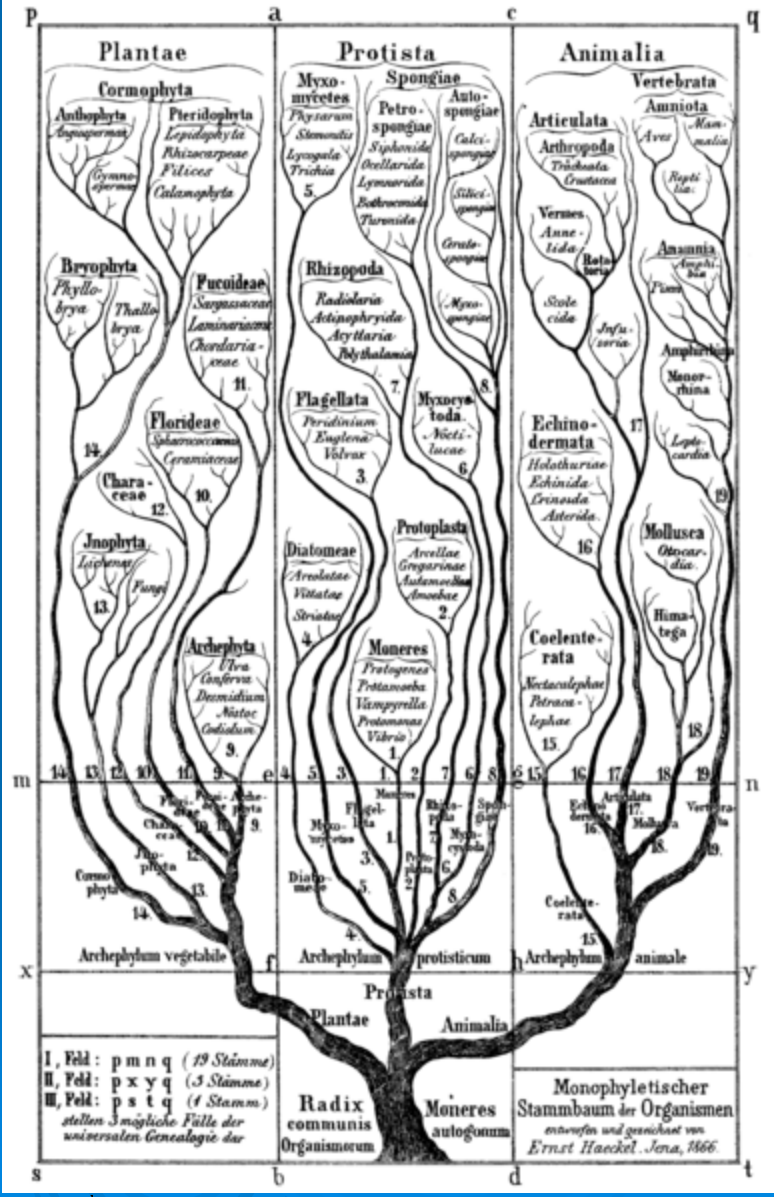
    [redacted]voluční vývoj organizmů

    ⇒ fylogenetický strom

- změny v dědičné informaci v průběhu evoluce

    [redacted]rganizmy evolučně blíže

    ⇒ proteiny sekvenčně blíže



fylogenetický strom - Haeckel (1866)

Fylogenetická podobnost – organizmy s nezveřejněným genomem

- **použití dostupných informací pro evolučně co nejbližší organizmy**
- **při studiu pomocí hmotnostní spektrometrie (MS)**
  - MS běžně vychází ze známých proteinových sekvencí pro organizmy se známým genomem...
  - použití **EST databáze** („expression sequence tags“)
    - většinou nekompletní úseky cDNA/mRNA (exprimovaných genů)
    - sekvenční data relativně nízké kvality („single pass“ sekvenace)
    - lze přepsat do aminokyselinové sekvence
  - ***de novo* sekvenace/identifikace peptidů** (z MS/MS spektra se přímo určí možný peptid)
    - **Blast *de novo* peptidů** proti zvolené databázi (omezená taxonomie)

3. Biologické sítě

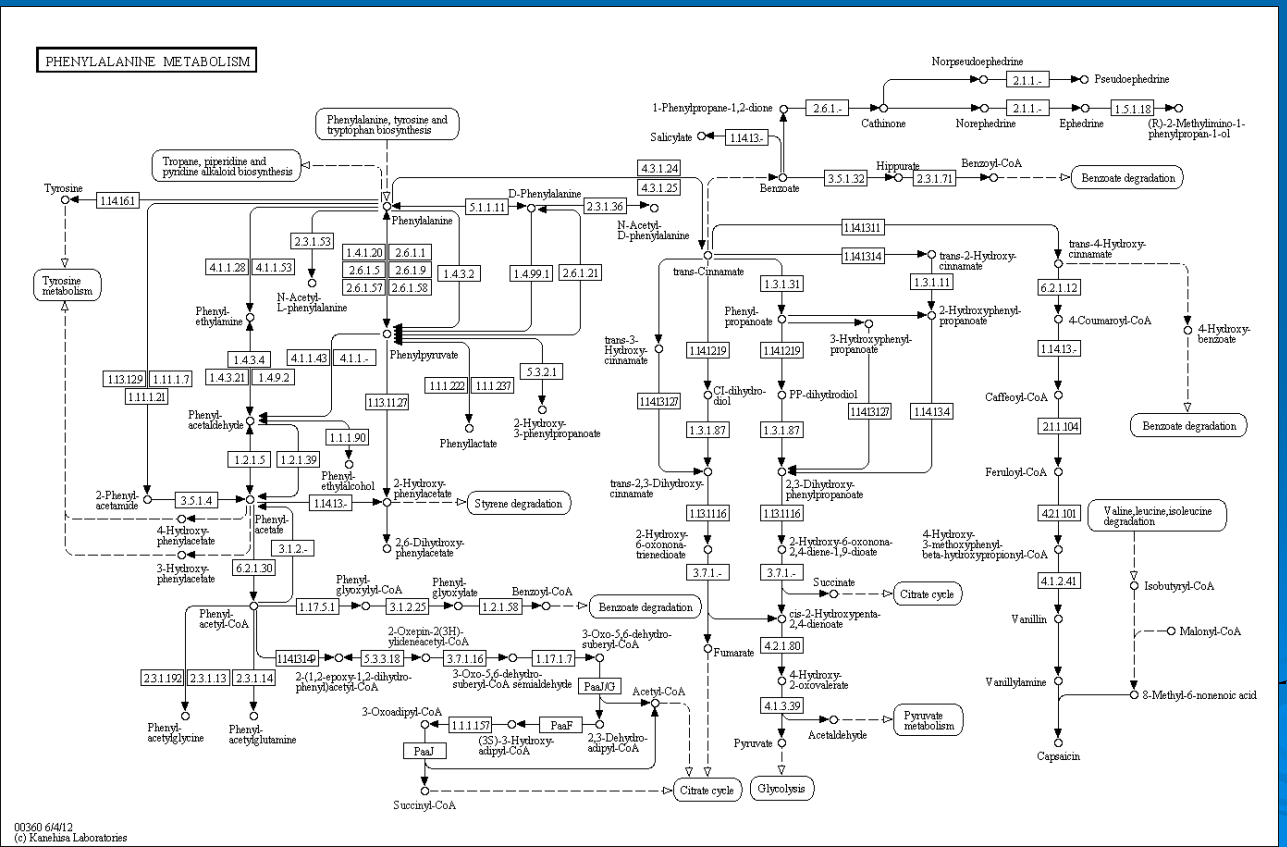
Biologické sítě

- **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 skrze produkty a reaktanty
    - **produkt jednoho = substrát druhého**
  - **např.; např. KEGG; WikiPathways**



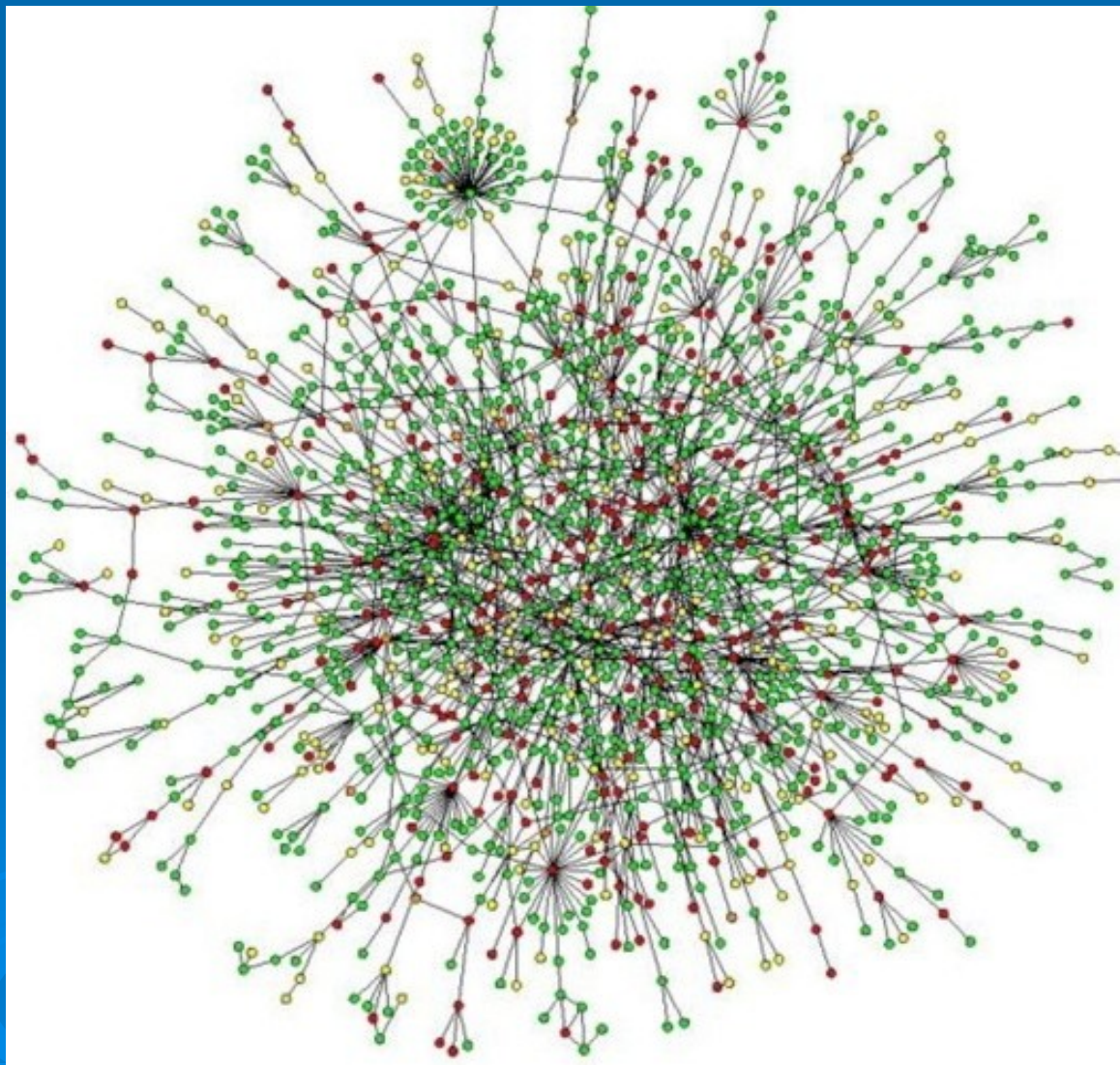
část metabolické sítě – metabolismus Phe (KEGG)

Biologické sítě – příklady (2)

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

Biologické sítě – příklady (3)

- **protein-protein fyzické interakce – ze sítě samotné není přímá informace o významu dané interakce...**
  - **příklady databází**
    - MINT
    - DIP
    - BioGRID
    - STRING
    - ...



Biologické sítě – příklady (4)

- **proteiny anotované do stejné kategorie genové ontologie (GO) – viz. dále**
  - **proteiny přítomné ve stejné GO kategorii mají společné**
    - **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...)

3. 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 (proteinů)
- **společné termíny pro všechny organizmy**
- **http://www.geneontology.org/ + AmiGO prohlížeč (online, offline)**

- **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, ...)**

Genová ontologie (GO) (2)

- kde se berou data pro GO?
  - každá anotace obsahuje informaci o svém původu – „evidence code“
  - A) manuálně přiřazené správcem/kurátorem („curator“)
    - „experimental evidence codes“ [redacted] reálného experimentu
    - „computational analysis evidence codes“ [redacted] *in silico* analýzy
    - „author statement evidence codes“ [redacted]rzení autora + citace
    - „curatorial statement codes“ [redacted]rzení 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

- srovnání biologických entit na základě jejich anotace v rámci dané ontologie – sémantická podobnost
- srovnání biologických entit na základě jejich funkce – funkční podobnost
- např. proteiny se stejnou funkcí lokalizované v jiné části buňky

KEGG

- KEGG = „Kyoto Encyclopedia of Genes and Genomes“
- 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

3. Biologické sítě

Příklady použití

Biologické sítě – použití („network analysis“)

- **sledujeme vliv nízkomolekulární látky na rostlinu a identifikovali jsme sadu rozdílně exprimovaných proteinů...**
  - jsou tyto proteiny zahrnuty v příslušné metabolické dráze? (KEGG)
  - jaké metabolické dráhy byly „významně“ zastoupeny? (test na statistickou významnost „identifikace“ (= pokrytí) dráhy v našem proteinovém listu) – „+“ i „–“ regulace
  - jsou známy proteinové komplexy mezi nalezenými proteiny?
  - je mezi proteiny zastoupeno více proteinů z konkrétního GO termínu?

- **studujeme interakční partnery zvoleného proteinu...**
  - vidíme již známé interakční partnery?
  - možno predikovat potencionální protein-protein interakce; následuje studium společných vlastností – GO termíny, metabolické dráhy?

Biologické sítě – použití („network analysis“) (2)

- **pozorujeme protein, který má vztah k onkologickému onemocnění...**
  - interaguje tento protein s dalšími proteiny?

    (**zvýšená pravděpodobnost**, že se tyto proteiny **aktivně** účastní daného onemocnění **potencionální cíle dalšího studia a nové léčby**)

- **„zdraví versus nemocní“ – rozdílné proteiny**
  - kterých metabolických drah se účastní? vysvětluje to důsledky nemoci?
  - jsou rozdílné proteiny převážně lokalizované v některé z organel?

Biologické sítě – použití („network analysis“) (3)

- **na co si dát pozor?**
  - falešně pozitivní i negativní informace v biologických sítích
    (nedostatečná citlivost současných přístupů pro studium)
  - mnoho dat v databázích z predikčních studií
    (i přes kontrolu nemusí zcela odpovídat)
  - stále víme málo...
  - **volba vhodných otázek, na které nám sítě dokážou dat odpověď**

Biologické sítě – použití („network analysis“) (4)

- jak se postavit k výstupům?
  - manuální validace výstupů
  - ověřením původních zdrojů
  - pochybovat a ptát se
  - experimentální ověření našich závěrů
  - není důležité jak to vypadá, ale co se z toho dá vyčíst...

4. Příklad využití bioinformatických nástrojů

Zavedení problému

- **studium proteinových komplexů vybraného proteinu**

- **imunoprecipitace proteinových komplexů (IP experiment)**
  - protilátka proti proteinu, u kterého chceme zjistit jeho partnery („bait“)
  - dva hlavní přístupy
    - protilátka imobilizovaná např. na kuličkách (magnetické, v kolonkách)
    - protilátka se přidá do volného roztoku a následně „lovíme“ protilátku
  - nutno opakovat experiment s různými protilátkami (epitopy, stérické zábrany, potvrzení předchozích experimentů)
  - nativní prostředí při experimentech
  - paralelně experimenty bez „bait“ – „bead proteome“

LC-MS/MS analýza „pull-down“ vzorků

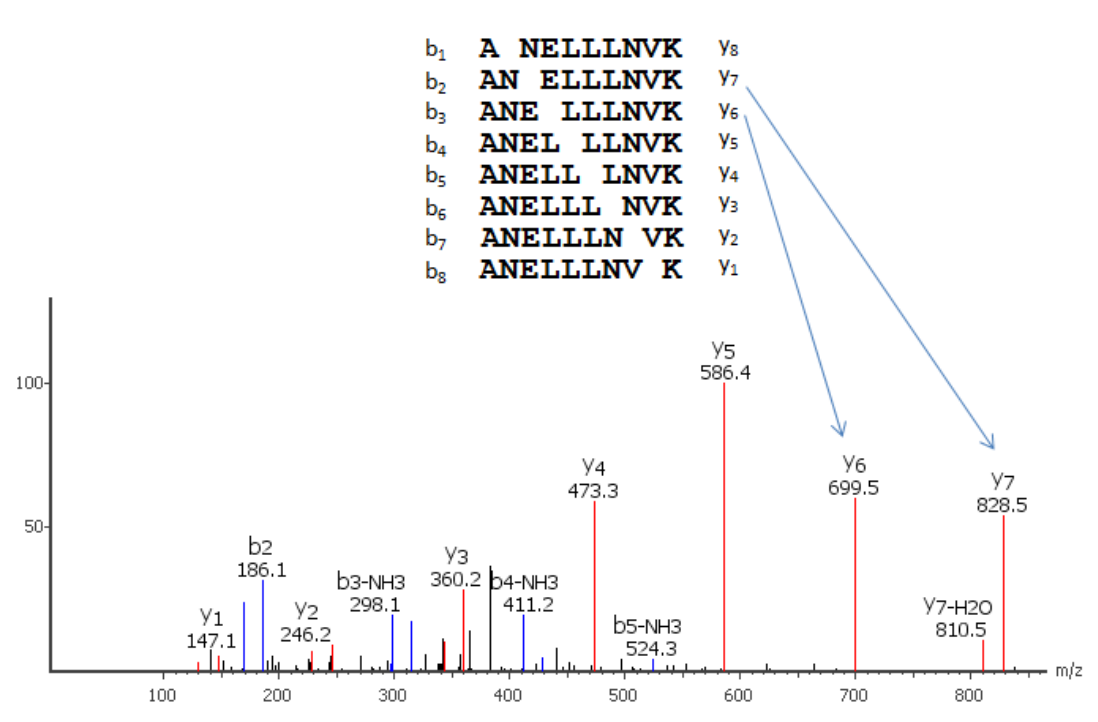
- **digesce proteinů [redacted]eptidy**
- **LC-MS/MS analýza směsi peptidů**
  - **peptidy vstupují do MS v pořadí rostoucí hydrofobicity** (LC separace)
- **MS zjistí MW peptidů a získá MS/MS spektra**

  (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 [redacted]**
   (CID; „collision induced dissociation“)

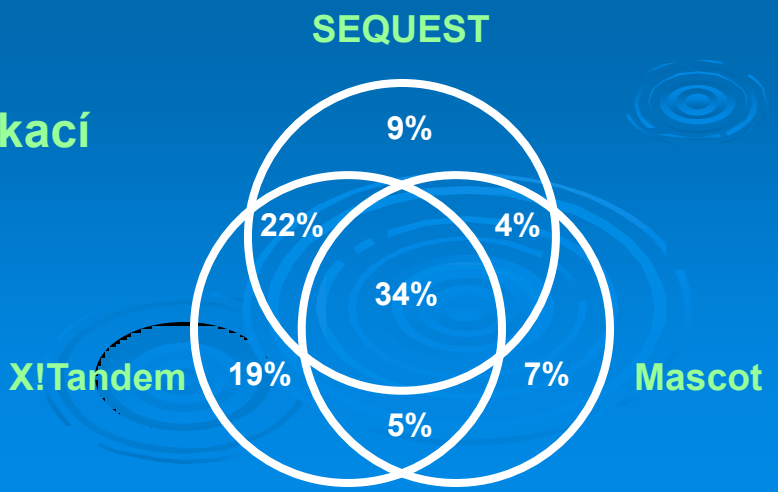


Zpracování LC-MS/MS dat

- LC-MS/MS data z analýz „pull-down“ vzorků po digesci = MS/MS spektra
- řádově 1 000 – 100 000 MS/MS spekter
- identifikace peptidů
  - vycházíme z proteinové databáze, např. UniRef100, *Homo sapiens*
  - *in silico* se vytvoří seznam možných peptidů
  - >20 algoritmů pro automat. přiřazení MS spektra možným peptidům (Sequest, Mascot, XTandem!, OMSSA, Phenyx, Andromeda, ...)
  - jiný algoritmus ný přístup úzná citlivost dlišné výsledky

kombinace algoritmů ⇒

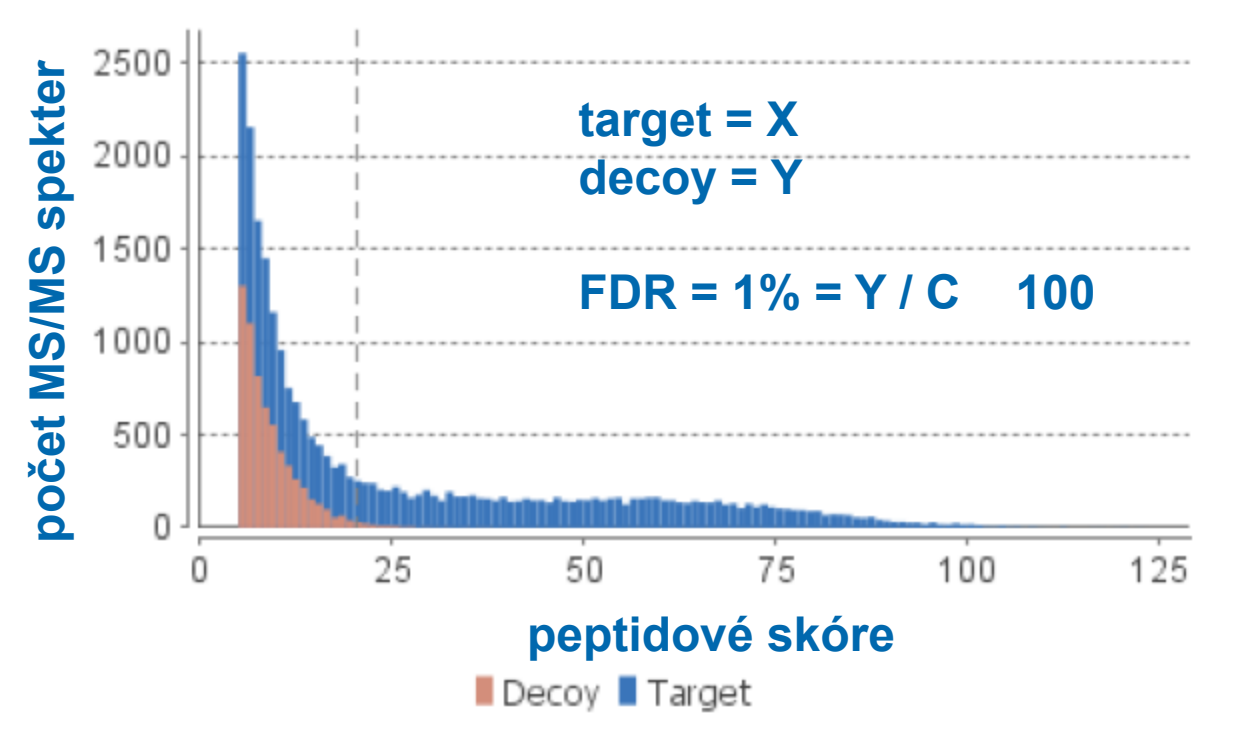
zvýšení počtu identifikací



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 i „decoy“ proteinové databázi

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



Zpracování LC-MS/MS dat (3)

- **z identifikovaných peptidů k proteinům**
  - problém u „**bottom-up**“ přístupu
  - v MS analýze **vidíme jen malou část** z např. tryptických **peptidů** proteinů a navíc nevíme ze kterých proteinů původně pochází...

    ⇒ **problém s určením seznamu proteinu**

    (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 – Proteiny A
  - slepý vzorek; „bead proteome“ – Proteiny B

- co nás zajímá v našem IP experimentu nejvíce?

Pohled na seznamy identifikovaných proteinů (2)

- **proteiny „navíc“ v A**
  - **1) kvalitativní změny (A: „**ano“, **B**: „ne“**) – citlivost použitého přístupu...**
  - **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** (**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“?

- **1) manuální prohledání dostupných informací v literatuře**

- **2) www.UniProt.org (**ID mapping, informace a odkazy na další databáze**)**
- **3) DAVID http://david.abcc.ncifcrf.gov/home.jsp**
- **4) ANAP http://gmdd.shgmo.org/Computational-Biology/ANAP/ANAP_V1.0/**
  - **doplnění k přednášce**
    - **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**)**

- **5) Cytoscape**
- **....**

Děkuji za pozornost