8. Bioinformatika a proteiny I

David Potěšil

Core Facility – Proteomics

CEITEC-MU

Masaryk University

Kamenice 5, A26

telefon: +420 54949 8426

email: david.potesil@ceitec.muni.cz

Obsah přednášky

1. Co je to bioinformatika?
2. Taxonomie a fylogeneze
3. Evoluce proteinů, proteinové domény
4. BLAST, srovnávání sekvencí

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 a aplikace 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)

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

Příbuzné disciplíny

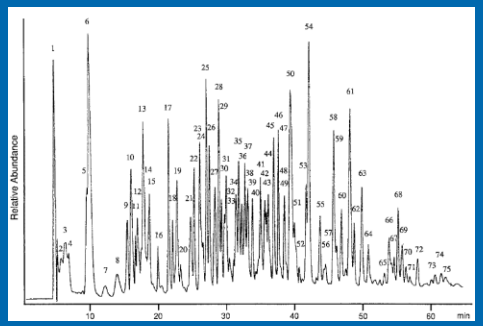
- ***data mining***
  - analýza dat z různých perspektiv a „dolování“ shrnujících (zobecněných) informací
- **matematická a teoretická biologie**
  - matematická prezentace, zpracování a modelování biol. procesů
- **lékařská informatika**
  - tvorba databází medicínských informací a jejich další využití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

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

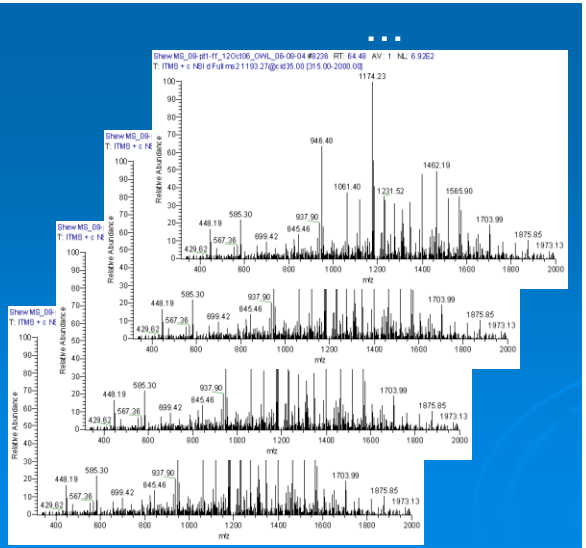
Příklad využití bioinformatických nástrojů

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

LC-MS záznam



MS/MS spektra

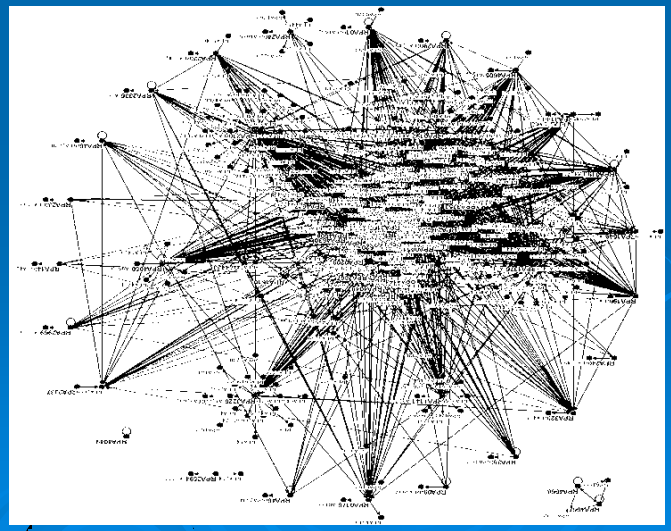


kvant. informace

peptidy (kvalitativní informace)

bioinformat. nástroje („*black box*“)

„standardní“ nastavení



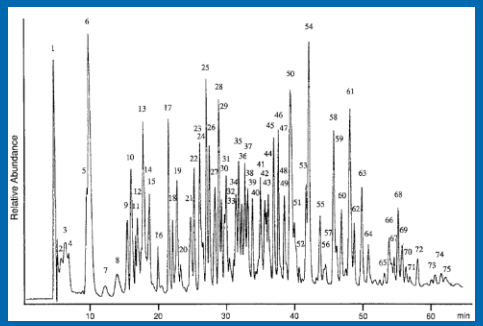
protein-protein interakční síť **?**

závěry z analýze této sítě?

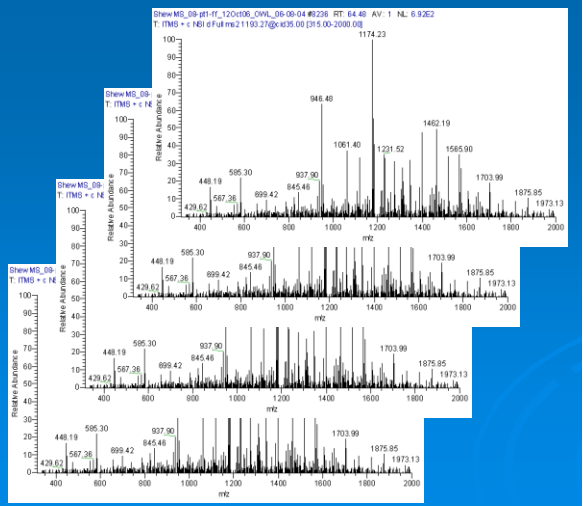
Příklad využití bioinformatických nástrojů

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

LC-MS záznam



MS/MS spektra

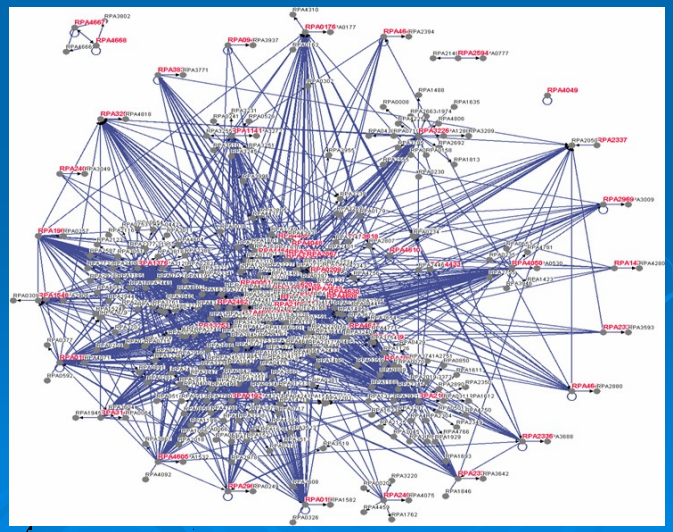


kvant. informace

peptidy (kvalitativní informace)

bioinformat. nástroje („*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 (sekvenční homologie)

- **BLAST – Basic Local Alignment Search Tool** (dále podrobněji)

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

    ⇒ podobnost ve struktuře proteinů

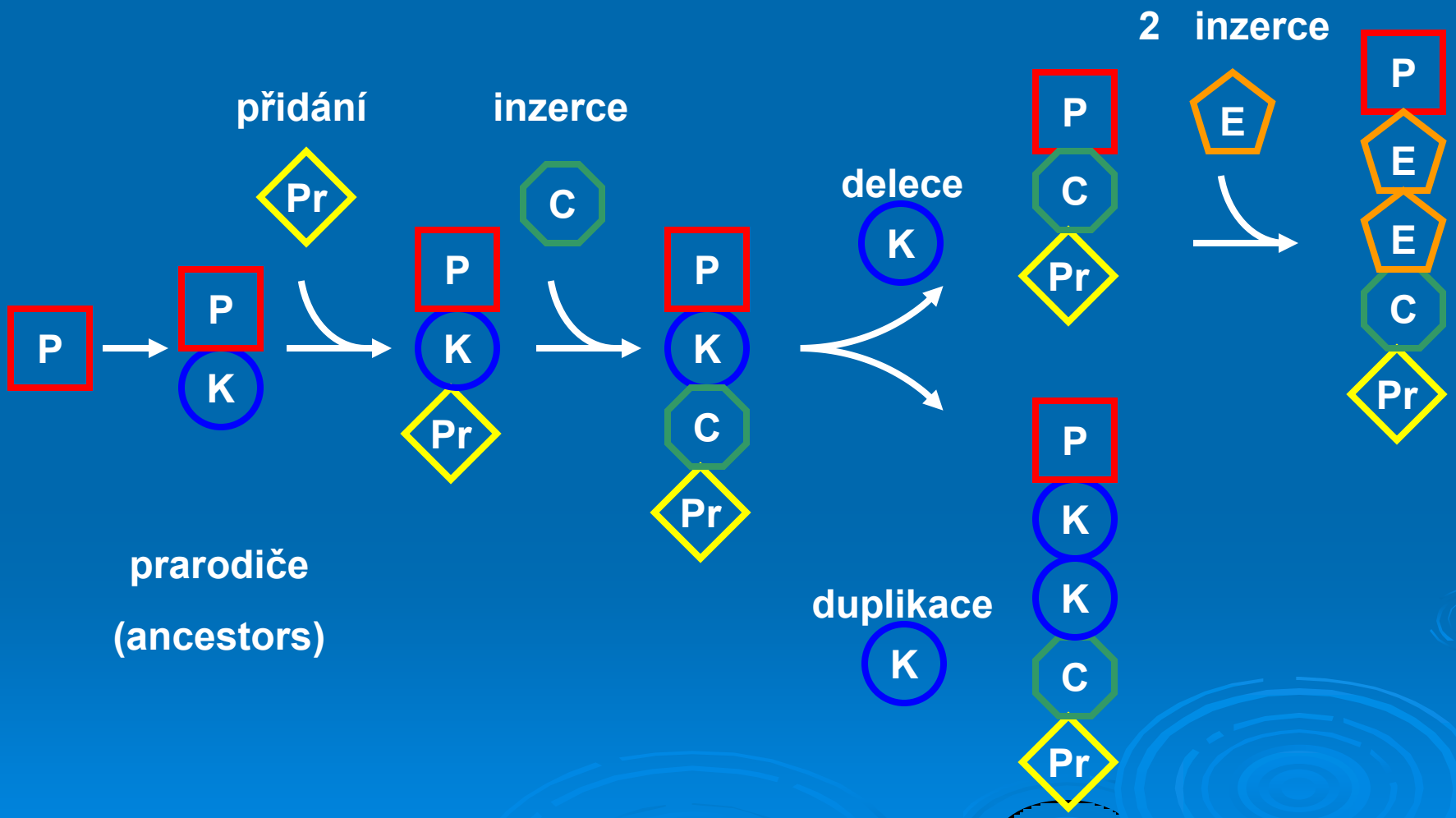
    ⇒ **podobnost ve funkci proteinů...**

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, změna 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 kombinovaných 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)
  - **strukturní podobnost** (sekundární a terciární struktura proteinu/domény)
  - **funkční podobnost** (nezávislá na sekvenční a strukturní podobnosti)

- **doménové rodiny a superrodiny a podobnost**
  - **sekvenční** podobnost ⇒ **doménová rodina**
    - evolučně mladší (mutace v krátké době [redacted]ekv. podobnost zachována)
  - **strukturní, funkční** podobnost ⇒ **doménová superrodina**
    - stejní proteinoví prarodiče, evolučně starší (dlouhodobá mutace sekvence [redacted]ekv. podobnost nemusí být zachována)

Hlavní zdroje pro klasifikaci domén

- klasifikace domén do superrodin a rodin

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

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

- 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)
  - **A. používají různé proteinové databáze** (primární sekvence) **pro klasifikaci**
    - UniProtKB (SwissProt a TrEMBL)
    - NCBI RefSeq
    - proteinové databáze pro vybrané kompletně sekvenované organizmy
    - ...
  - **B. používají různé části proteinu pro predikci rodin/superrodin**
- **integrální zdroje**
  - sbírají informace z více zdrojů a prezentují na jediném místě
  - **InterPro (http://www.ebi.ac.uk/interpro/) – příklad P12345, P04637**
  - **CDD** (*Conserved Domain Database*)

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

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

Co získám znalostí proteinové rodiny/superrodiny?

- **předpokládaná funkce proteinu**
  - pokud není protein sám o sobě již detailně prostudován...
  - navazující GO (*gene ontology*) termíny – viz. příští přednáška

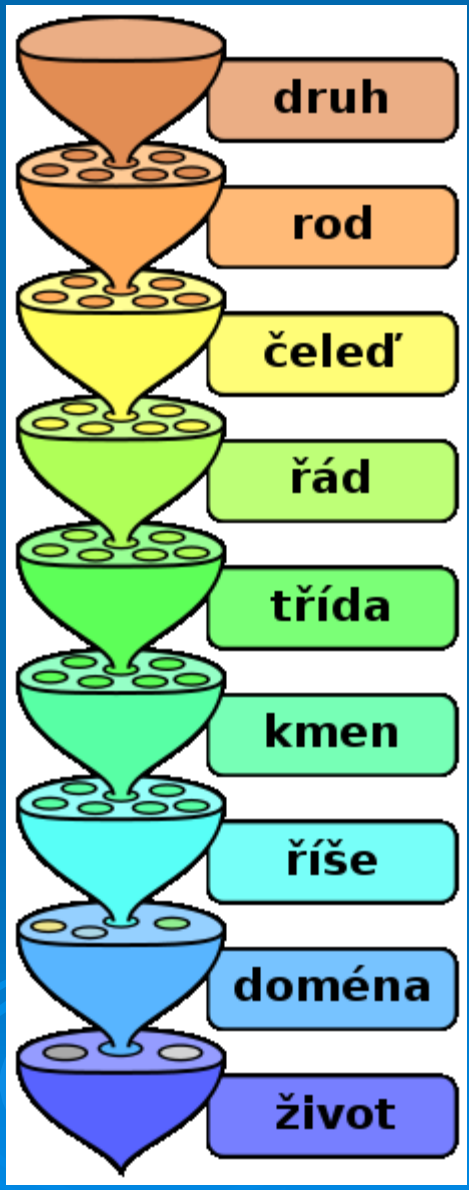
- **klasifikace v systému proteinových rodin/superrodin**
  - návaznosti na jiné rodiny, metabolické dráhy atd.

- **důležité např. při studiu seznamu proteinů/genů se změněnou hladinou/expresí**
  - ***datamining***
    - proteiny většinou nepůsobí samostatně, paralelní dráhy, atd.
    - případně lze pozorovat změny u proteinů následujících/předcházejících v kaskádě změn v reakci na konkrétní stimul

3. Taxonomie a fylogeneze

Taxonomie

- taxon
  - skupina žijících či již vymřelých organizmů se společnými znaky, jimiž se odlišují od jiných taxonů (organizmů v těchto taxonech)
- taxonomické dělení
  - při objevení nového organizmu
  - manuální třídění dle společných a jedinečných znaků
  - snaha o shodu s fylogenezí – evolučním vývojem organizmu
  - základní taxonomické kategorie – viz. obr.



wikipedie

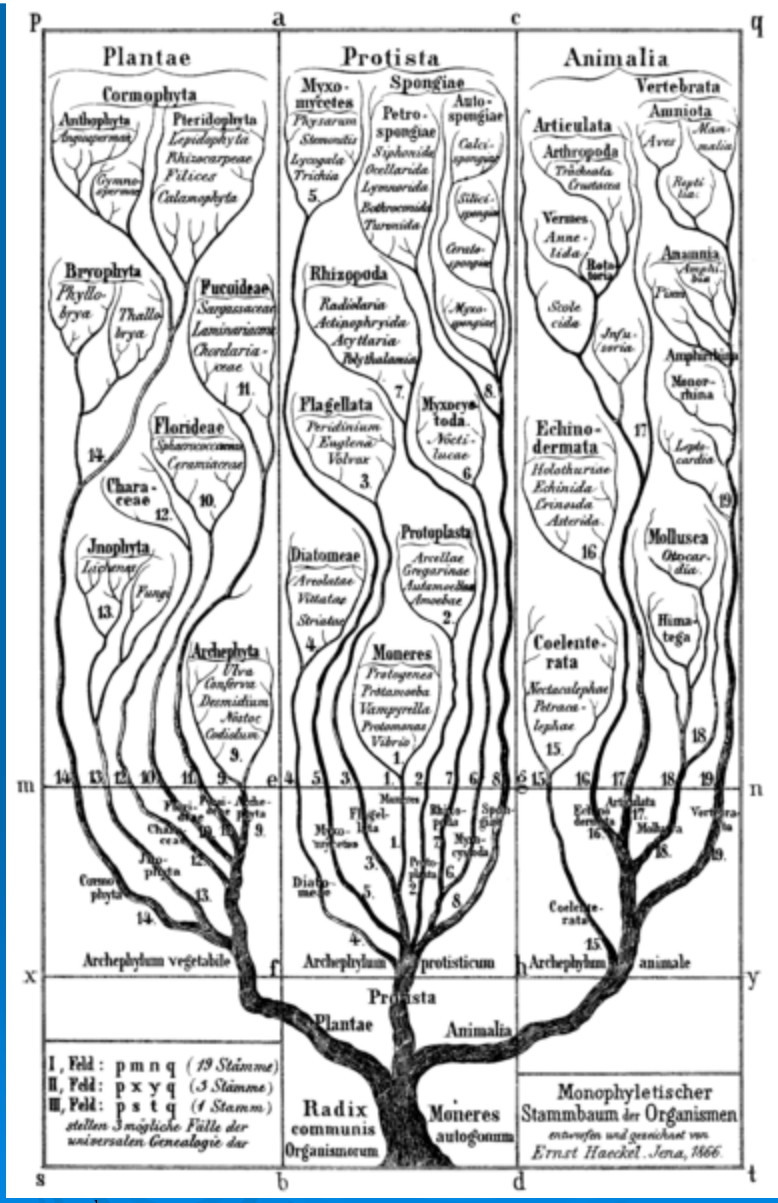
- http://www.ncbi.nlm.nih.gov/taxonomy

Fylogeneze (fylogenetický vývoj)

- evoluční vztah organizmů
- využití morfologických dat a v poslední době hlavně výsledky molekulárního sekvenování

[redacted]voluční vývoj organizmů

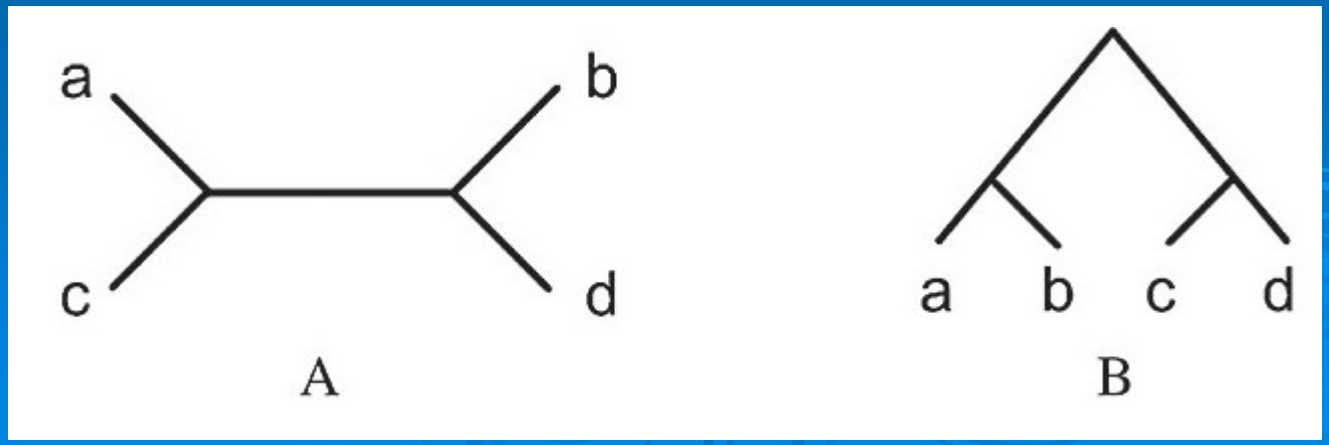
⇒ fylogenetický strom



fylogenetický strom - Haeckel (1866)

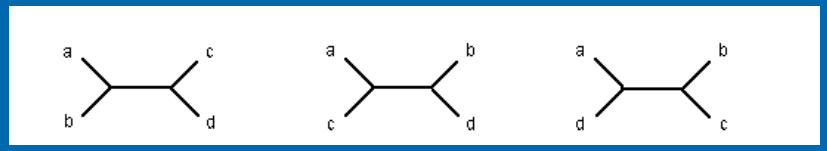
Fylogenetické stromy

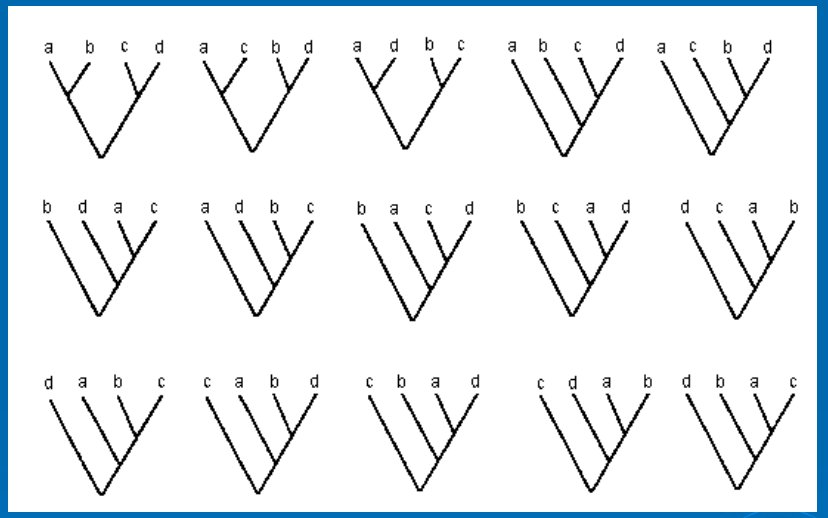
- grafické znázornění příbuzenských vztahů mezi různými taxonomickými jednotkami / jednotlivými druhy / geny
- tvorba fylogenetických stromů
  - definování „podobnosti“ mezi např. taxonomickými jednotkami
    - morfologické vlastnosti – vzdálenost dána důležitostí morf. znaků
    - sekvenční podobnost na úrovni genomů (i proteinů)
  - podobnější tax. jednotky jsou si ve fylogenetickém stromu blíže
  - různé zobrazení fyl. stromů: nezakořeněný (A), zakořeněný (B), aj.



Fylogenetické stromy (2) – příklad komplexnosti

- tvorba fylogenetických stromů
  - možnosti pro případ 4 organizmů celkem 3 (nezakořeněný), resp. 15 (zakořeněný)





  - možnosti pro případ 10 organizmů – celkem ~2 resp. ~34 M...

Pouze jeden je správný...

⇒ využití morfologických, sekvenčních či jiných informací

Fylogenetické stromy (3) – vybrané nástroje

- iTOL – interactive Tree Of Life (http://itol.embl.de/index.shtml)
  - automatizované zobrazení fylogenetického stromu – sekvenční data
  - pro organizmy se známým genomem, případně vlastní data
  - struktura nemusí nutně odpovídat evoluci – nepřesná data, gen. anomálie (např. horizontální přenos genů)

- phyloT (http://phylot.biobyte.de/)
  - pracuje s taxonomickým zařazením dle NCBI (http://www.ncbi.nlm.nih.gov/taxonomy/)
    - manuálně editované řazení organizmů, které jsou přítomné ve veřejných sekvenčních databázích (~10% z celkového počtu známých organizmů...)
  - export výsledků; zobrazení v iTOL s obdobnými možnostmi zobrazení

Fylogenetická podobnost – organizmy s nezveřejněným genomem

- použití dostupných informací pro evolučně co nejbližší organizmy
- příklad 1 – identifikace proteinů pomocí hmotnostní spektrometrie (MS)
  - běžně se vychází ze známých prot. sekvencí (znám genom)
  - co když organizmus nemá zveřejněný genom?
    - použití proteinové databáze pro evolučně blízký organizmus
      - A) databázové hledání přímo proti této databázi
      - B) *de novo* sekvenace peptidů a BLAST *de novo* peptidů
  - například *Trichinella spiralis* versus *Trichinella pseudospiralis*
    - dříve ***T.pseudospiralis*** bez veřejně dostupného genomu/proteomu
    - využíván proteom pro ***T. spiralis***
    - dnes už sekvenční informace pro oba organizmy...
      - jen další ukázka dynamiky celého odvětví!!!

Fylogenetická podobnost – organizmy s nezveřejněným genomem

- příklad 2 – nedostačující anotace proteinů pro organizmus zájmu
  - genom/proteom znám, ale není známa funkce/popis daných proteinů
  - použití modelového organismu s lepší anotací proteinů a evolučně blízkého
    - BLAST jednotlivých proteinů vůči databázi model. organizmu
  - například *N.tobaccum* versus *A.thaliana*
    - zlepšení anotace proteinů tabáku tím, že provedeme BLAST jednotlivých tabákových proteinů proti *A.thaliana* databázi
    - nejlepší hit z *A.thaliana* se vezme pro anotaci tabákového proteinu
    - je možné použít i více modelových organizmů...

4. BLAST, srovnávání sekvencí

Základní formáty proteinových sekvencí/databází

- **FASTA formát – hlavička specifická pro zdrojovou databázi, relativně málo informací; postačuje pro získání a další zpracování proteinové sekvence**

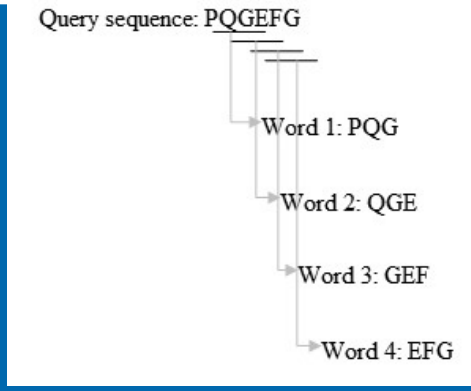
```
>sp|P04637|P53_HUMAN Cellular tumor antigen p53 OS=Homo sapiens GN=TP53 PE=1 SV=4
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGP
DEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAK
SVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHE
RCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNYMCNS
SCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGRDRRTEEENLRKKGEPHHELP
PGSTKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQAGKEPG
GSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD
```

- **xml formát**
  - **komplexní forma s kompletní informací k danému proteinu ze zdrojové databáze**
  - **konkrétní forma specifická pro zdrojovou databázi – xml schéma**
  - **obsahuje např. kompletní taxonomii zdrojového organizmu; známé modifikace; výčet interakčních partnerů, označení v jiných databázích a jiné bioinformaticky (automaticky) zpracovatelné informace**

BLAST – Basic Local Alignment Search Tool

- **srovnání proteinových či nukleotidových sekvencí** (většinou FASTA formát)
- **různé algoritmy dle vstupu** (protein či nukleotid) **a typu srovnání**
- **nejběžnější algoritmy** (pro proteiny)
  - **blastp – protein-proteinová databáze**
  - **blastx – nukleotid** (překlad na proteinovou sekvenci)**-proteinová databáze**
- **vybrané speciální algoritmy – k hledání vzdáleně příbuzných proteinů**
  - **PSI-BLAST – Position Specific Iteration BLAST**
    - **po blastp ze zvoleného počtu sekvencí vytvoří novou pozičně-specifickou skórovací matrici (PSSM), kterou použije v dalším hledání; tento postup je možno několikrát opakovat**
  - **DELTA-BLAST – obdoba PSI-BLAST; využívá předpřipravené PSSM dle konzervativních domén v NCBI databázi rychlejší a citlivější**

Základní kroky BLAST algoritmů



1. generování k-písmenných úseků – „slov“ (parametr *word size*)
   - proteiny – běžně K = 3; nukleotidy – běžně K = 11
2. prohledání každého „slova“ vůči cílové databázi a ponechání těch slov, kde se našla shoda překračující stanovené limitní skóre *high scoring words*
3. hledání ***high scoring words*** z databáze; hledána úplná shoda – *exact match*
4. rozšíření *exact match* na obě strany původního k-písmenného slova a hledání *high-scoring segment pairs* (HSPs) pro každý *exact match* – rozšiřování do doby, dokud neklesá skóre pro původní *exact match*
5. zhodnocení statistické významnosti jednotlivých HSPs
6. spojení HSPs do delších úseků
7. výpočet *expectation value* (E)

Substituční skórovací matice pro výpočet skóre

- matice „substitučních“ skóre pro jednotlivé AK, které se uvažují při srovnávání prvotních k-písmenných slov i při jejich rozšiřování a hledání HSPs
- nejběžnější matice BLOSUM62
  - skóre se odvíjí od četnosti aminokyselinové záměny v reálných proteinech jejichž identita je větší než 62%
  - + hodnoty – častější substituce
  - - hodnoty – méně časté subst.
  - po diagonále největší skóre (62% ident. [redacted] ejčastější jev)
  - zaokrouhleno na celá čísla

```
#  Matrix made by matblas from blosum62.iij
#  * column uses minimum score
#  BLOSUM Clustered Scoring Matrix in 1/2 Bit Units
#  Blocks Database = /data/blocks_5.0/blocks.dat
#  Cluster Percentage: >= 62
#  Entropy =   0.6979, Expected =  -0.5209
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  Z  X  *
A  4 -1 -2 -2  0 -1 -1  0 -2 -1 -1 -1 -1 -2 -1  1  0 -3 -2  0 -2 -1  0 -4
R -1  5  0 -2 -3  1  0 -2  0 -3 -2  2 -1 -3 -2 -1 -1 -3 -2 -3 -1  0 -1 -4
N -2  0  6  1 -3  0  0  0  1 -3 -3  0 -2 -3 -2  1  0 -4 -2 -3  3  0 -1 -4
D -2 -2  1  6 -3  0  2 -1 -1 -3 -4 -1 -3 -3 -1  0 -1 -4 -3 -3  4  1 -1 -4
C  0 -3 -3 -3  9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1 -3 -3 -2 -4
Q -1  1  0  0 -3  5  2 -2  0 -3 -2  1  0 -3 -1  0 -1 -2 -1 -2  0  3 -1 -4
E -1  0  0  2 -4  2  5 -2  0 -3 -3  1 -2 -3 -1  0 -1 -3 -2 -2  1  4 -1 -4
G  0 -2  0 -1 -3 -2 -2  6 -2 -4 -4 -2 -3 -3 -2  0 -2 -2 -3 -3 -1 -2 -1 -4
H -2  0  1 -1 -3  0  0 -2  8 -3 -3 -1 -2 -1 -2 -1 -2 -2  2 -3  0  0 -1 -4
I -1 -3 -3 -3 -1 -3 -3 -4 -3  4  2 -3  1  0 -3 -2 -1 -3 -1  3 -3 -3 -1 -4
L -1 -2 -3 -4 -1 -2 -3 -4 -3  2  4 -2  2  0 -3 -2 -1 -2 -1  1 -4 -3 -1 -4
K -1  2  0 -1 -3  1  1 -2 -1 -3 -2  5 -1 -3 -1  0 -1 -3 -2 -2  0  1 -1 -4
M -1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  5  0 -2 -1 -1 -1 -1  1 -3 -1 -1 -4
F -2 -3 -3 -3 -2 -3 -3 -3 -1  0  0 -3  0  6 -4 -2 -2  1  3 -1 -3 -3 -1 -4
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4  7 -1 -1 -4 -3 -2 -2 -1 -2 -4
S  1 -1  1  0 -1  0  0  0 -1 -2 -2  0 -1 -2 -1  4  1 -3 -2 -2  0  0  0 -4
T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5 -2 -2  0 -1 -1  0 -4
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1  1 -4 -3 -2 11  2 -3 -4 -3 -2 -4
Y -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  3 -3 -2 -2  2  7 -1 -3 -2 -1 -4
V  0 -3 -3 -3 -1 -2 -2 -3 -3  3  1 -2  1 -1 -2 -2  0 -3 -1  4 -3 -2 -1 -4
B -2 -1  3  4 -3  0  1 -1  0 -3 -4  0 -3 -3 -2  0 -1 -4 -3 -3  4  1 -1 -4
Z -1  0  0  1 -3  3  4 -2  0 -3 -3  1 -1 -3 -1  0 -1 -3 -2 -2  1  4 -1 -4
X  0 -1 -1 -1 -2 -1 -1 -1 -1 -1 -1 -1 -1 -1 -2  0  0 -2 -1 -1 -1 -1 -1 -4
* -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4  1
```

Substituční skórovací matice pro výpočet skóre (2)

- **typ matice by měl být uzpůsoben délce hledané sekvence**
- ***word size* se doporučuje snížit u proteinů na 2 v případě krátkých sekvencí** (peptidy či menší proteiny)

Délka (počet AK)	**Substituční matice**
<35	PAM-30
35-50	PAM-70
50-85	BLOSUM-80
>85	BLOSUM-62

Substituční skórovací matice – skóre jednoho HSP

Query sequence: R P P Q G L F

Database sequence: D P **P E G** V V

Exact match is scanned.

Score: -2 **7 7 2 6 1** -1

HSP

Optimal accumulated score = 7+7+2+6+1 = 23

```
#  Matrix made by matblas from blosum62.iij
#  * column uses minimum score
#  BLOSUM Clustered Scoring Matrix in 1/2 Bit Units
#  Blocks Database = /data/blocks_5.0/blocks.dat
#  Cluster Percentage: >= 62
#  Entropy =   0.6979, Expected =  -0.5209
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  Z  X  *
A  4 -1 -2 -2  0 -1 -1  0 -2 -1 -1 -1 -1 -2 -1  1  0 -3 -2  0 -2 -1  0 -4
R -1  5  0 -2 -3  1  0 -2  0 -3 -2  2 -1 -3 -2 -1 -1 -3 -2 -3 -1  0 -1 -4
N -2  0  6  1 -3  0  0  0  1 -3 -3  0 -2 -3 -2  1  0 -4 -2 -3  3  0 -1 -4
D -2 -2  1  6 -3  0  2 -1 -1 -3 -4 -1 -3 -3 -1  0 -1 -4 -3 -3  4  1 -1 -4
C  0 -3 -3 -3  9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1 -3 -3 -2 -4
Q -1  1  0  0 -3  5  2 -2  0 -3 -2  1  0 -3 -1  0 -1 -2 -1 -2  0  3 -1 -4
E -1  0  0  2 -4  2  5 -2  0 -3 -3  1 -2 -3 -1  0 -1 -3 -2 -2  1  4 -1 -4
G  0 -2  0 -1 -3 -2 -2  6 -2 -4 -4 -2 -3 -3 -2  0 -2 -2 -3 -3 -1 -2 -1 -4
H -2  0  1 -1 -3  0  0 -2  8 -3 -3 -1 -2 -1 -2 -1 -2 -2  2 -3  0  0 -1 -4
I -1 -3 -3 -3 -1 -3 -3 -4 -3  4  2 -3  1  0 -3 -2 -1 -3 -1  3 -3 -3 -1 -4
L -1 -2 -3 -4 -1 -2 -3 -4 -3  2  4 -2  2  0 -3 -2 -1 -2 -1  1 -4 -3 -1 -4
K -1  2  0 -1 -3  1  1 -2 -1 -3 -2  5 -1 -3 -1  0 -1 -3 -2 -2  0  1 -1 -4
M -1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  5  0 -2 -1 -1 -1 -1  1 -3 -1 -1 -4
F -2 -3 -3 -3 -2 -3 -3 -3 -1  0  0 -3  0  6 -4 -2 -2  1  3 -1 -3 -3 -1 -4
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4  7 -1 -1 -4 -3 -2 -2 -1 -2 -4
S  1 -1  1  0 -1  0  0  0 -1 -2 -2  0 -1 -2 -1  4  1 -3 -2 -2  0  0  0 -4
T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5 -2 -2  0 -1 -1  0 -4
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1  1 -4 -3 -2 11  2 -3 -4 -3 -2 -4
Y -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  3 -3 -2 -2  2  7 -1 -3 -2 -1 -4
V  0 -3 -3 -3 -1 -2 -2 -3 -3  3  1 -2  1 -1 -2 -2  0 -3 -1  4 -3 -2 -1 -4
B -2 -1  3  4 -3  0  1 -1  0 -3 -4  0 -3 -3 -2  0 -1 -4 -3 -3  4  1 -1 -4
Z -1  0  0  1 -3  3  4 -2  0 -3 -3  1 -1 -3 -1  0 -1 -3 -2 -2  1  4 -1 -4
X  0 -1 -1 -1 -2 -1 -1 -1 -1 -1 -1 -1 -1 -1 -2  0  0 -2 -1 -1 -1 -1 -1 -4
* -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4  1
```

Příklady webových BLAST rozhraní

- **Pubmed (http://blast.ncbi.nlm.nih.gov/Blast.cgi)**
- **UniProt (http://www.uniprot.org/blast/)**

Offline možnosti

- **BioEdit (http://www.mbio.ncsu.edu/bioedit/bioedit.html)**
  - nejen pro BLAST...
  - možnost použití vlastních databází atd.
- **blast+ (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/)**
  - sada nástrojů pro práci v příkazové řádce
  - příklad příkazu:
    - blastp -db „**databáze**“ -out „**kam zapsat výstup**“ -word_size 3 -gapopen 11 -gapextend 1 -threshold 11 -outfmt "**6 std positive ppos**" -num_threads 4 -comp_based_stats 2
  - http://www.ncbi.nlm.nih.gov/books/NBK279675/ - seznam možností

Zhodnocení výstupu BLAST

- *expectation value* (E) – hlavní parametr
  - **počet sekvencí z databáze, které se přiřadí hledané sekvenci se stejným skóre pouze dílem náhody – relevantní E pod ~0,05-0,001**
    - záleží na konkrétní aplikaci a následné validaci výstupů...
    - hodnotí se i délka sekvence – krátkých sekvencí obecně vyšší E
- *identities* **– počet identických aminokyselin (AK) z hledaného proteinu**
- *positives* **– počet AK s podobnými fyzikálně chemickými vlastnostmi**

Možnosti dávkové BLAST (Pubmed)

- https://blast.ncbi.nlm.nih.gov/Blast.cgi
- několik desítek až stovek proteinů
- možnost procházet individuální výsledky
- možnost stažení shrnutých výsledků + zpracování v externím programu
- příklad – proteiny *Nicotiana tabacum*

Srovnání sekvencí dvou či více proteinů (UniProt)

- http://www.uniprot.org/align/
- obdobný přístup jako při BLAST
- křížové srovnání v případě více srovnávaných sekvencí
- příklad: srovnání vybraných sekvencí Ig Light Chain gammna

Děkuji za pozornost