

10. Bioinformatika a proteiny I

David Potěšil

Core Facility – Proteomics

CEITEC-MU

Masaryk University

Kamenice 5, A2

phone: +420 54949 7304

email: david.potesil@ceitec.muni.cz

Proteomika, Podzim 2013

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.** (Tekaiia, 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)



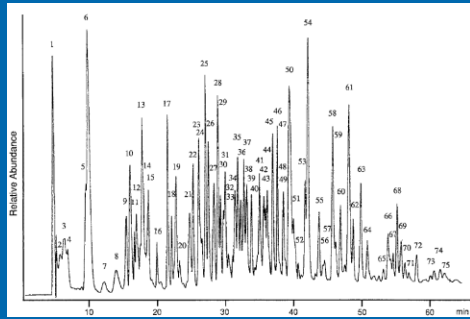
Příbuzné disciplíny

- ***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 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



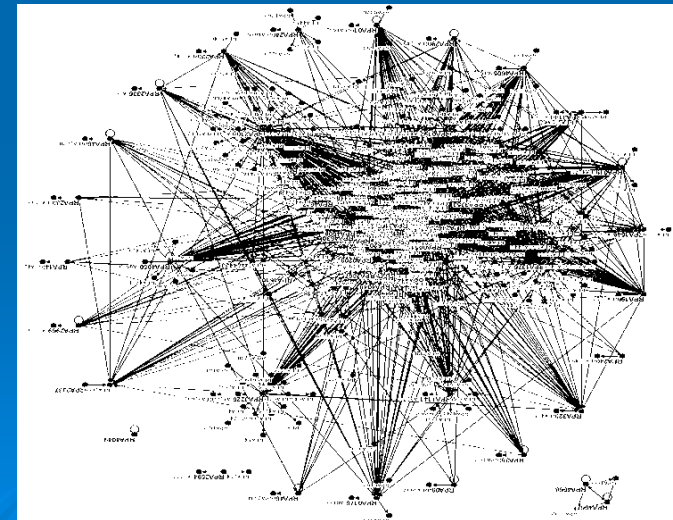
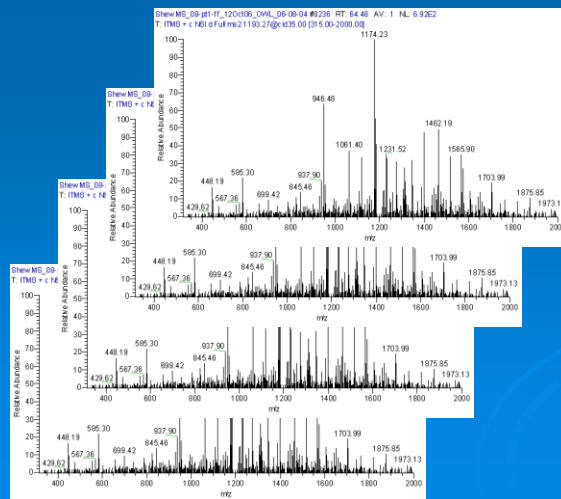
kvant.
informace

peptidy

„black box“

„standardní
nastavení“

MS/MS spektra

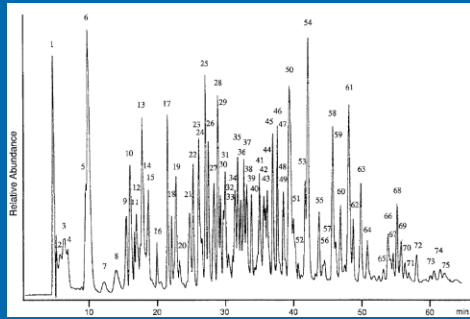


protein-protein interakční síť ?
závěry z analýze této sítě?

Příklad využití bioinformatických nástrojů

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

LC-MS záznam



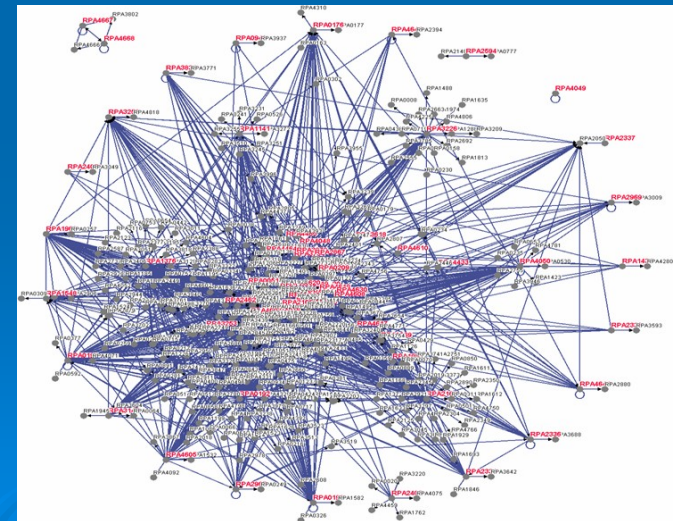
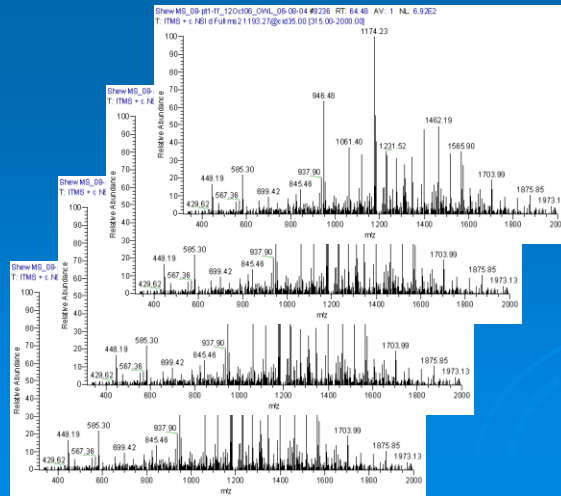
kvant.
informace

„white box“

výstupům
přízpůsobené
nastavení

peptidy

MS/MS spektra



protein-protein interakční síť

analýza sítě: úloha proteinu A z jeho interakcí

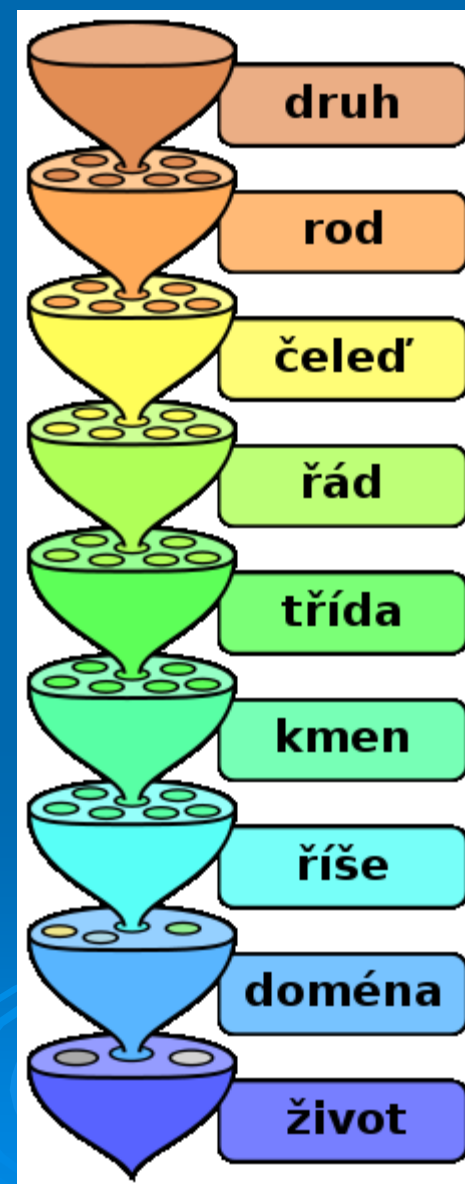
2. Taxonomie a fylogeneze



Taxonomie

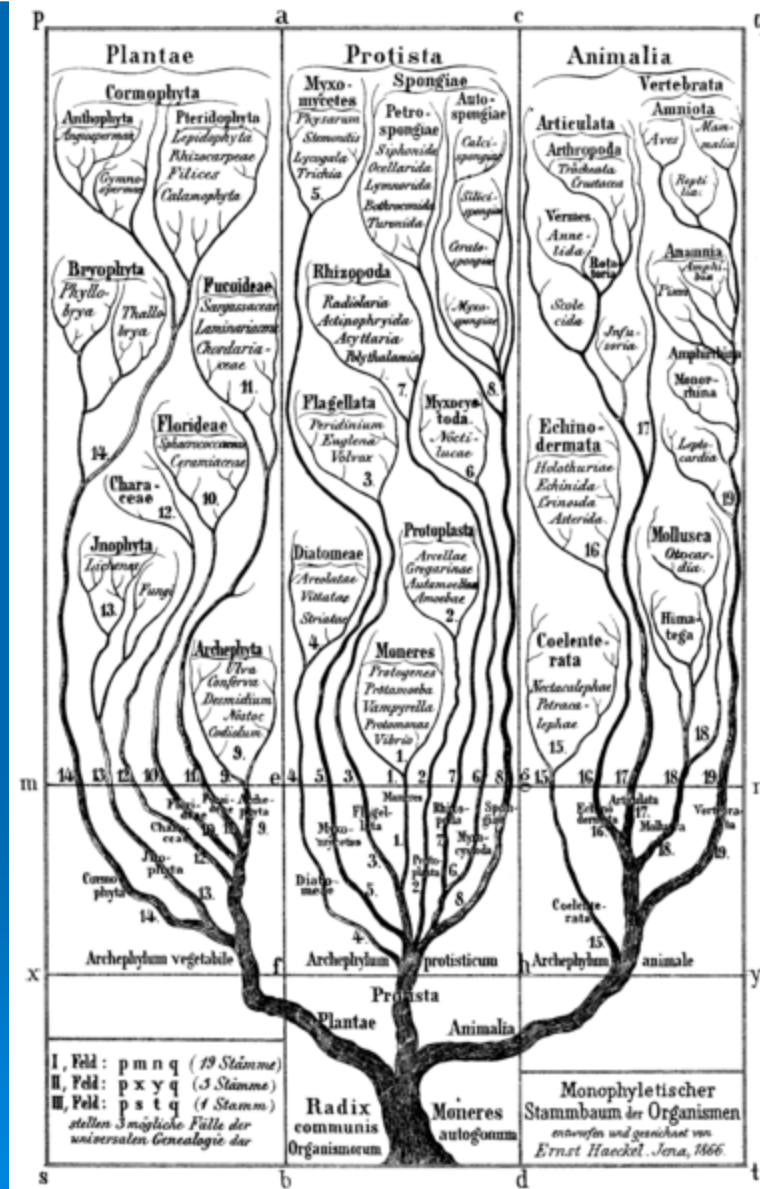
- taxon
 - skupina žijících či již vymřelých organismů se společnými znaky, jimiž se odlišují od jiných taxonů
- taxonomické dělení
 - při objevení nového organismu
 - manuální třídění dle společných a jedinečných znaků
 - snaha o shodu s fylogenezí – evolučním vývojem organismu
 - základní taxonomické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

• <http://www.ncbi.nlm.nih.gov/taxonomy>



Fylogeneze (fylogenetický vývoj)

- evoluční vztah organismů
 - využití morfologických dat a v poslední době hlavně výsledky molekulární sekvenování
- ⇒ evoluční vývoj organismů
- ⇒ **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)
 - různé zobrazení: zakořeněný, nezakořeněný, kruhový
- **iTOL** – **interactive Tree Of Life**
 - automatizované zobrazení fylogenetického stromu – sekv. data (lze nahrát i vlastní seznamy taxonomií aj.)

Fylogenetická podobnost – organizmy s nezveřejněným genomem

- použití dostupných informací pro **evolučně co nejbližší organizmy**
- při studiu pomocí hmotnostní spektrometrie (MS)
 - běžně vychází ze **známých proteinových sekvencí** (znám genom)
 - co když organizmus nemá zveřejněný genom?
 - 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 (taxonomie)
- **příklad** – *Trichinella spiralis* versus *Trichinella pseudospiralis*...

3. Evoluce proteinů, proteinové domény



Jedna z prvních aplikací bioinformatiky

– srovnání primárních sekvencí (sekvenční homologie)

- **BLAST** – **B**asic **L**ocal **A**lignment **S**earch **T**ool (dále podrobněji)
- proč srovnávat primární sekvence?
 - podobnost v primární sekvenci proteinů
 - ⇒ podobnost ve struktuře proteinů
 - ⇒ **podobnost ve funkci proteinů...**

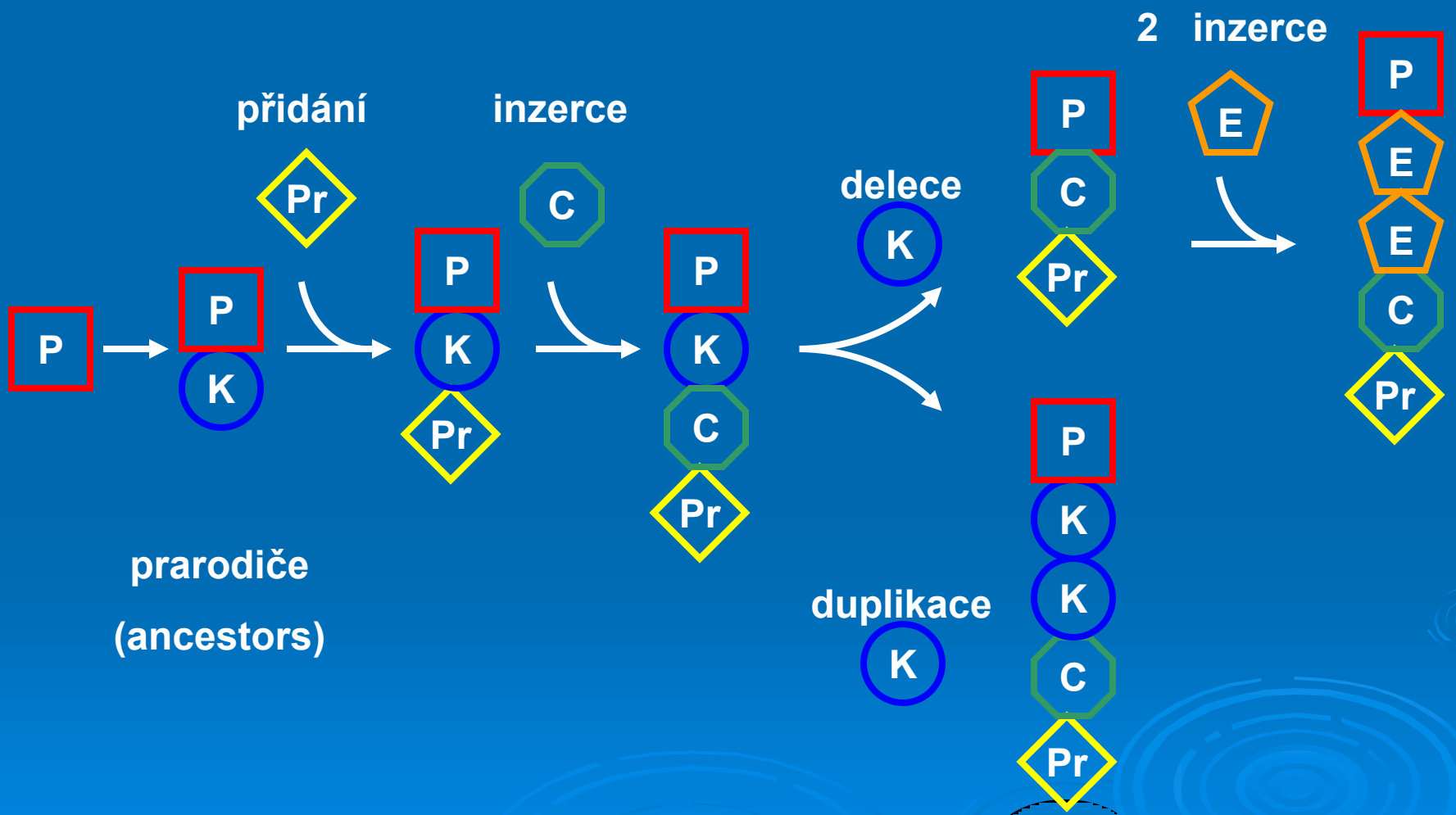
Není tak jednoduché...



Proteinová evoluce a proteinové domény

- proteinová doména = **nezávislá** strukturní, funkční a evoluční jednotka
- 2/3 proteinů jednobuněčných a 80% proteinů mnohobuněčných organismů 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 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 superrodiny a rodiny a podobnost**
 - **strukturní, funkční podobnost** \Rightarrow **doménová superrodina**
 - stejní proteinoví prarodiče, evolučně starší (dlouhodobá mutace sekvence \Rightarrow sekv. podobnost nemusí být zachována)
 - **sekvenční podobnost** \Rightarrow **doménová rodina**
 - evolučně mladší (mutace v krátké době \Rightarrow sekv. podobnost zachována)

Hlavní zdroje pro klasifikaci domén

- klasifikace domén do superrodin a rodin
- **CATH** (***C**lass, **A**rchitecture, **T**opology, **H**omologous Superfamily*)
 - <http://www.cathdb.info/>
- **SCOP** (***S**tructural **C**lassification **O**f **P**roteins*)
 - <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)
 - používají různé cílové proteinové databáze (primární sekvence)
 - UniProtKB (SwissProt a TrEMBL)
 - NCBI RefSeq
 - proteinové databáze pro vybrané kompletně sekvenované organizmy
 - ...
 - 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
 - **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

4. BLAST, srovnávání sekvencí



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
SVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGRVVRAMAIYKQSQHMTEVVRRCPHHE
RCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCCTTIHYNMCNS
SCMGGMNRRPILTIITLEDSSGNLLGRNSFEVVRVCACPRDRRTEENLRKKGEPHHELP
PGSTKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQAGKEPG
GSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD
```

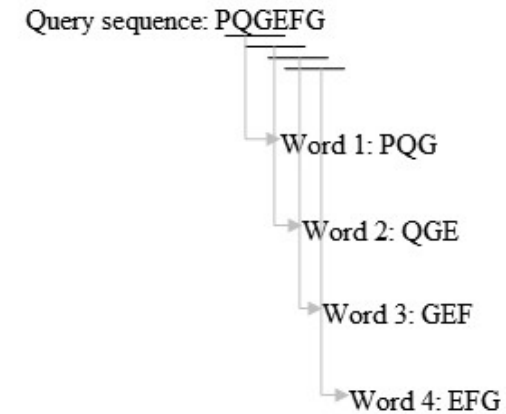
- xml formát
 - komplexní forma s kompletní informací k danému proteinu ze zdrojové databáze
 - specifická pro zdrojovou databázi
 - obsahuje např. kompletní taxonomii zdrojového organismu; známé modifikace; výčet interakčních partnerů, označení v jiných databázích a jiné bioinformaticky (automaticky) 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 matici (**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“
(z případně upravené sekvence; 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* v *databázi*; 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 (2)

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

Délka	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 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 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 -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
Z -2 -1 3 4 -3 0 1 -1 0 -3 -4 0 -3 -3 -2 0 -1 -4 -3 -3 4 1 -1 -4
B -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 1
```



Možnosti dávkové BLAST

- 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

- shodný přístup jako při BLAST programu
- křížové srovnání v případě více srovnávaných sekvencí
- příklad: srovnání vybraných sekvencí **Ig Light Chain gamma**

Příklady webových BLAST rozhraní

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

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
- *identities* – počet identických aminokyselin (AK) z hledaného proteinu
- *positives* – počet AK s podobnými fyzikálně chemickými vlastnostmi

Děkuji za pozornost

