Databáze biologických dat

Bioinformatika

Ivana Burgetová, Tomáš Martínek

burgetova@fit.vutbr.cz, martinto@fit.vutbr.cz

Osnova

- Motivace
- Databáze sekvencí
- Genomové databáze
- Databáze struktur
- Ostatní databáze
- Integrační nástroje
- Shrnutí

Motivace

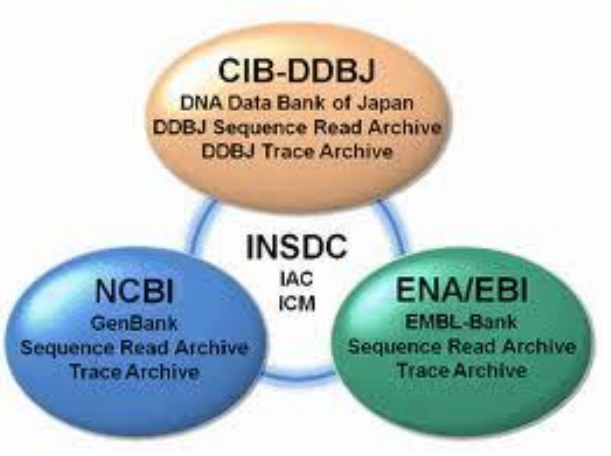
- Rozvoj molekulární biologie a genomického výzkumu ⇒ obrovské množství dat
- Biologická data – jedna z největších a nejbohatších odborných datových množin
- Jednou z klíčových úloh bioinformatiky je:
  - Vytvářet a spravovat databáze biologických dat
  - Poskytnout uživatelsky přívětivé rozhraní pro přístup k datům
  - Vytvářet nástroje, které analyzují data v databázích
  - Vytvářet nástroje pro vizualizaci dat
  - Vytvářet platformy, které integrují existující nástroje pro analýzu s dostupnými databázemi

Osnova

- Motivace
- Databáze sekvencí
- Genomové databáze
- Databáze struktur
- Ostatní databáze
- Integrační nástroje
- Shrnutí

Sekvence nukleových kyselin

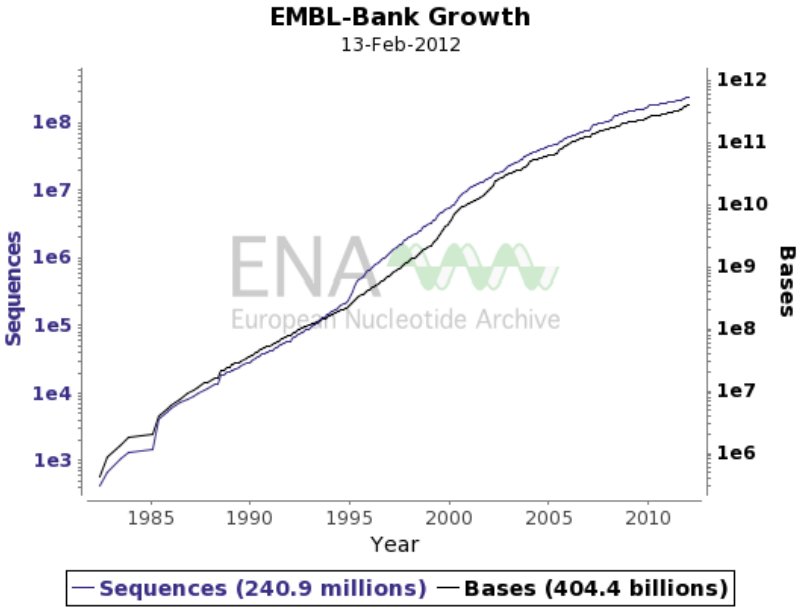
- Hlavní databáze nukleotidů
  - EMBL - Nucleotide Sequence Database
    - spravovaná EBI (European Bioinformatics Institute)
    - hlavní evropská databáze nukleotidových sekvencí
  - DDBJ – DNA Data Bank of Japan
    - spravovaná NIG (National Institute of Genetics)
  - GenBank – Genetic sequence database
    - spravovaná NCBI (National Center for Biotechnology Information)

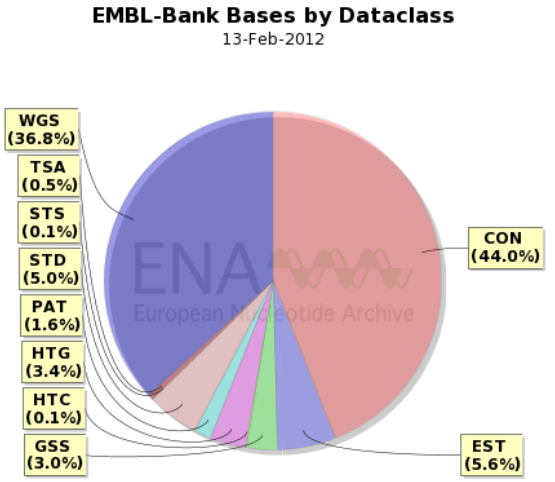


- INSCD - International Nucleotide Sequence Database Collaboration
  - Organizace zastřešující kooperaci a výměnu dat mezi všemi třemi hlavními databázemi
  - Výměna dat mezi jednotlivými DB probíhá každý den

Sekvence nukleových kyselin

- Data v DDBJ/EMBL/GenBank
  - Zjištěny sekvenováním molekuly DNA
  - Nárůst objemu dat – exponenciální nárůst (objem dat se zdvojnásobí každých 6 až 9 měsíců)
  - Aktuálně 240 miliónů záznamů (404 miliard nukleotidů)
  - data dostupná přes FTP
- Struktura dat:
  - WGS – Whole Genom Shotgun (úseky pro sestavení genomu)
  - CON – Entry constructed from segment entry sequences
  - EST – Expressed Sequence Tags (mRNA úseky)
  - HTG - High Thoughput Genome sequencing
  - GSS - Genome Survey Sequence (podobné EST)
  - PAT – Patent
  - STD – nezařazené





GenBank – Formát záznamu

- Formát záznamu sekvence v databázi GenBank
  - Textový formát záznamu, který kromě vlastní sekvence nukleotidů obsahuje i další informace jako:
    - LOCUS – jméno záznamu, délka sekvence, typ molekuly, kód divize, datum posledního zveřejnění záznamu
    - DEFINITION – sumarizace biologického obsahu záznamu (stručný popis sekvence)
    - ACCESSION – unikátní číslo záznamu
    - KEYWORDS – fráze popisující produkt genu a další relevantní informace
    - SOURCE – popis tkáně, ze které byla sekvence zjištěna
    - ORGANISM – zdrojový organismus
    - REFERENCE – odkazy na literaturu, kde byly zveřejněny informace o sekvenci, minimálně 1 odkaz
    - FEATURES - detailní popis vlastností sekvence (biologické informace obsažené v sekvenci)
    - ORIGIN - Vlastní sekvence nukleotidů

GenBank – Příklad záznamu

```
LOCUS       DMU54469 2881 bp DNA linear INV 22-FEB-1998
DEFINITION Drosophila melanogaster eukaryotic initiation factor 4E (eIF4E)
           gene, alternative splice products, complete cds.
ACCESSION  U54469 VERSION U54469.1 GI:1322283
KEYWORDS   .
SOURCE     Drosophila melanogaster (fruit fly)
  ORGANISM   Drosophila melanogaster
           Eukaryota; Metazoa; Arthropoda; Hexapoda; Insecta; Pterygota;
           Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha;
           Ephydroidea; Drosophilidae; Drosophila; Sophophora.
REFERENCE  1 (bases 1 to 2881)
  AUTHORS  Lavoie,C.A., Lachance,P.E., Sonenberg,N. and Lasko,P.
  TITLE    Alternatively spliced transcripts from the Drosophila eIF4E gene
           produce two different Cap-binding proteins
  JOURNAL J. Biol. Chem. 271 (27), 16393-16398 (1996)
  PUBMED 8663200
...
ORIGIN
     1 cggttgcttg ggttttataa catcagtcag tgacaggcat ttccagagtt gccctgttca
    61 acaatcgata gctgcctttg gccaccaaaa tcccaaactt aattaaagaa ttaaataatt
   121 cgaataataa ttaagcccag taacctacgc agcttgagtg cgtaaccgat atctagtata
   181 catttcgata catcgaaatc atggtagtgt tggagacgga gaaggtaaga cgatgataga
   241 cggcgagccg catgggttcg ...
```

Formát FASTA

- Alternativní formát dat: FASTA
  - Jeden z nejjednodušších formátů
  - Textový formát: Hlavička (1 řádek) + Data
  - Hlavička: uvozená znakem >, zdroj, ID, krátký popis
  - Data: 60 znaků na řádek
- Příklad:

```
>gb|U54469.1|DMU54469 Drosophila melanogaster eukaryotic initiation factor
CGGTTGCTTGGGTTTTATAACATCAGTCAGTGACAGGCATTTCCAGAGTTGCCCTGTTCAACAATCGATA
GCTGCCTTTGGCCACCAAAATCCCAAACTTAATTAAAGAATTAAATAATTCGAATAATAATTAAGCCCAG
TAACCTACGCAGCTTGAGTGCGTAACCGATATCTAGTATACATTTCGATACATCGAAATCATGGTAGTGT
TGGAGACGGAGAAGGTAAGACGATGATAGACGGCGAGCCGCATGGGTTCGATTTGCGCTGAGCCGTGGCA
CTATTGTCTACGTCATAGCTATCGCTCATCTCTGTCTGTCTCTATCAAGCTATCTCTCTTTCGCGGTCAC
…
```

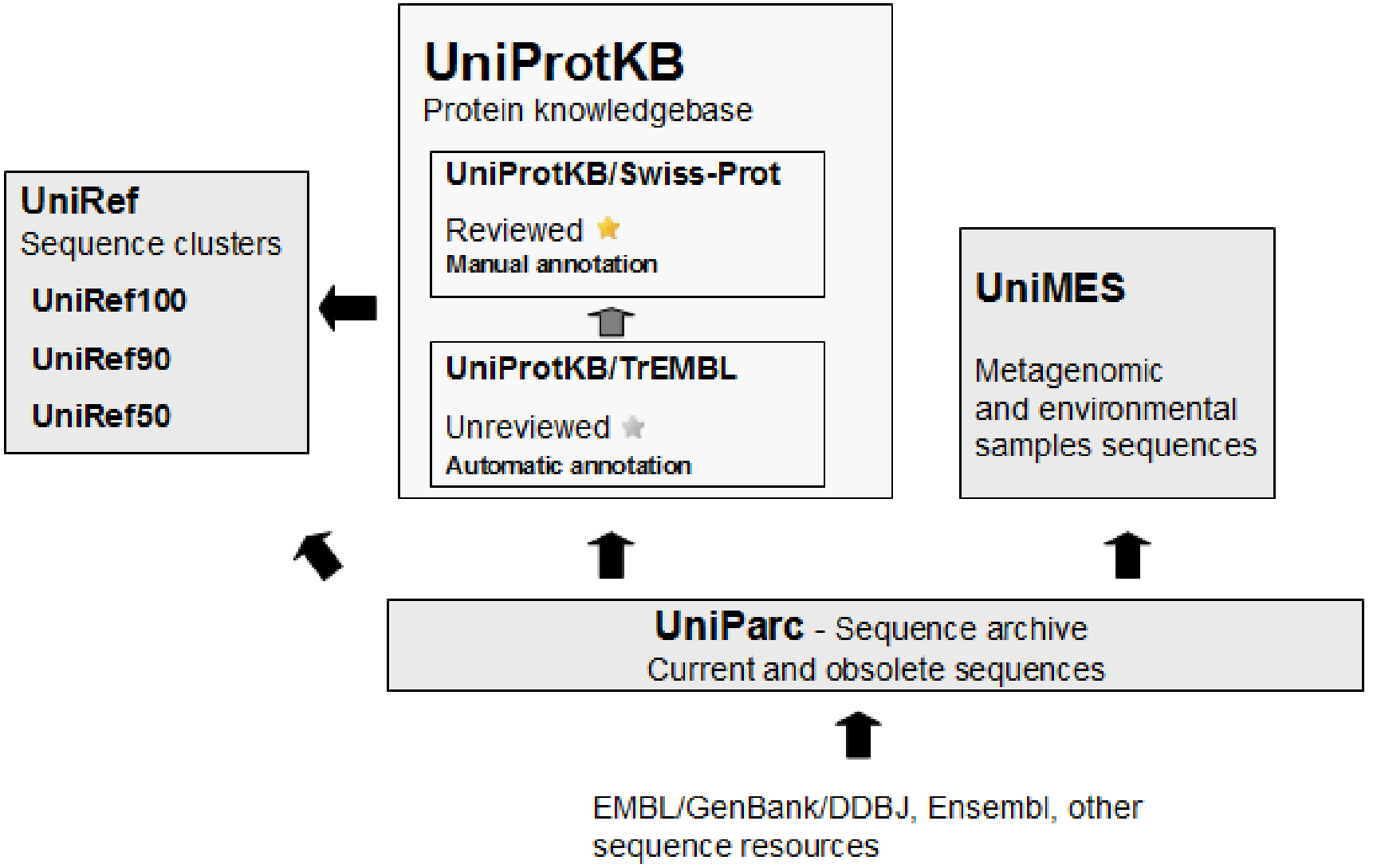
Sekvence proteinů - NCBI

- NCBI
  - GenePept
    - Kolekce proteinových sekvencí odvozených automatizovaně od kódujících oblastí nukleotidových sekvencí v GenBank
    - Záznamy nejsou příliš dobře provázány
  - RefSeq
    - Kolekce ověřených proteinových sekvencí
    - Pečlivě provázána informace DNA – mRNA – Protein

Sekvence proteinů - UniProt

- UniProt
  - Nejvýznamnější databáze proteinových sekvencí
  - Integruje data z:
    - European Bioinformatics Institute (EBI)
    - Swiss Institute of Bioinformatics (SIB)
    - Protein Information Resource (PIR)
  - Vytváří:
    - UniProtKB/Swiss-Prot – manuálně anotované ověřené záznamy
    - UniProtKB/TrEMBL – automaticky anotované neověřené záznamy
    - UniRef – shluky podobných proteinových sekvencí
    - UniMes – metagenomické a enviromentální vzorky

UniProt - schéma



UniProt – Formát záznamu

- Formát záznamu sekvence v databázi UniProt
  - Textový formát záznamu obsahuje vlastní sekvenci + doplňující info
  - Každý řádek začíná unikátním 2-znakovým kódem, který se může opakovat
    - ID – identifikační řádek
    - AC – accession number – identifikace záznamu (protože ID se může měnit)
    - DT – date entry - datum vytvoření/změny záznamu
    - DE – description – jména proteinu
    - GN – gene name – odpovídající gen
    - OS – organism species - identifikace organismu
    - OC – organism classification
    - OX – odkazy do taxonomické databáze
    - OG – organelle line – identifikace organely
    - RN – reference (odkazy na literaturu)
    - DR – database cross-references
    - FT – Feature table (detailní informace o specifických místech v sekvenci)
    - SQ – vlastní sekvence aminokyselin

UniProt – Příklad záznamu

```
ID   ROA1_HUMAN Reviewed; 372 AA.
AC   P09651; A8K4Z8; Q3MIB7; Q6PJZ7;
DT   01-JUL-1989, integrated into UniProtKB/Swiss-Prot.
DT   09-FEB-2010, sequence version 5.
DT   25-JAN-2012, entry version 168.
DE   RecName: Full=Heterogeneous nuclear ribonucleoprotein A1;
DE   Short=hnRNP A1; DE AltName: Full=Helix-destabilizing protein;
DE   AltName: Full=Single-strand RNA-binding protein;
DE   AltName: Full=hnRNP core protein A1;
GN   Name=HNRNPA1; Synonyms=HNRPA1;
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
OC   Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
...
SQ   SEQUENCE   372 AA;  38747 MW;  A06683571C6C109F CRC64;
     MSKSESPKEP EQLRKLFIGG LSFETTDESL RSHFEQWGTL TDCVVMRDPN TKRSRGFGFV
     TYATVEEVDA AMNARPHKVD GRVVEPKRAV SREDSQRPGA ...
```

Formát FASTA

- Alternativní formát dat: FASTA
  - Textový formát: Hlavička (1 řádek) + Data
  - Hlavička: uvozená znakem >, zdroj, ID, krátký popis
  - Data: 60 znaků na řádek
- Příklad:

```
>sp|P09651|ROA1_HUMAN Heterogeneous nuclear ribonucleoprotein A1
MSKSESPKEPEQLRKLFIGGLSFETTDESLRSHFEQWGTLTDCVVMRDPNTKRSRGFGFV
TYATVEEVDAAMNARPHKVDGRVVEPKRAVSREDSQRPGAHLTVKKIFVGGIKEDTEEHH
LRDYFEQYGKIEVIEIMTDRGSGKKRGFAFVTFDDHDSVDKIVIQKYHTVNGHNCEVRKA
LSKQEMASASSSQRGRSGSGNFGGGRGGGFGGNDNFGRGGNFSGRGGFGGSRGGGGYGGS
GDGYNGFGNDGGYGGGGPGYSGGSRGYGSGGQGYGNQGSGYGGSGSYDSYNNGGGGFGG
...
```

Osnova

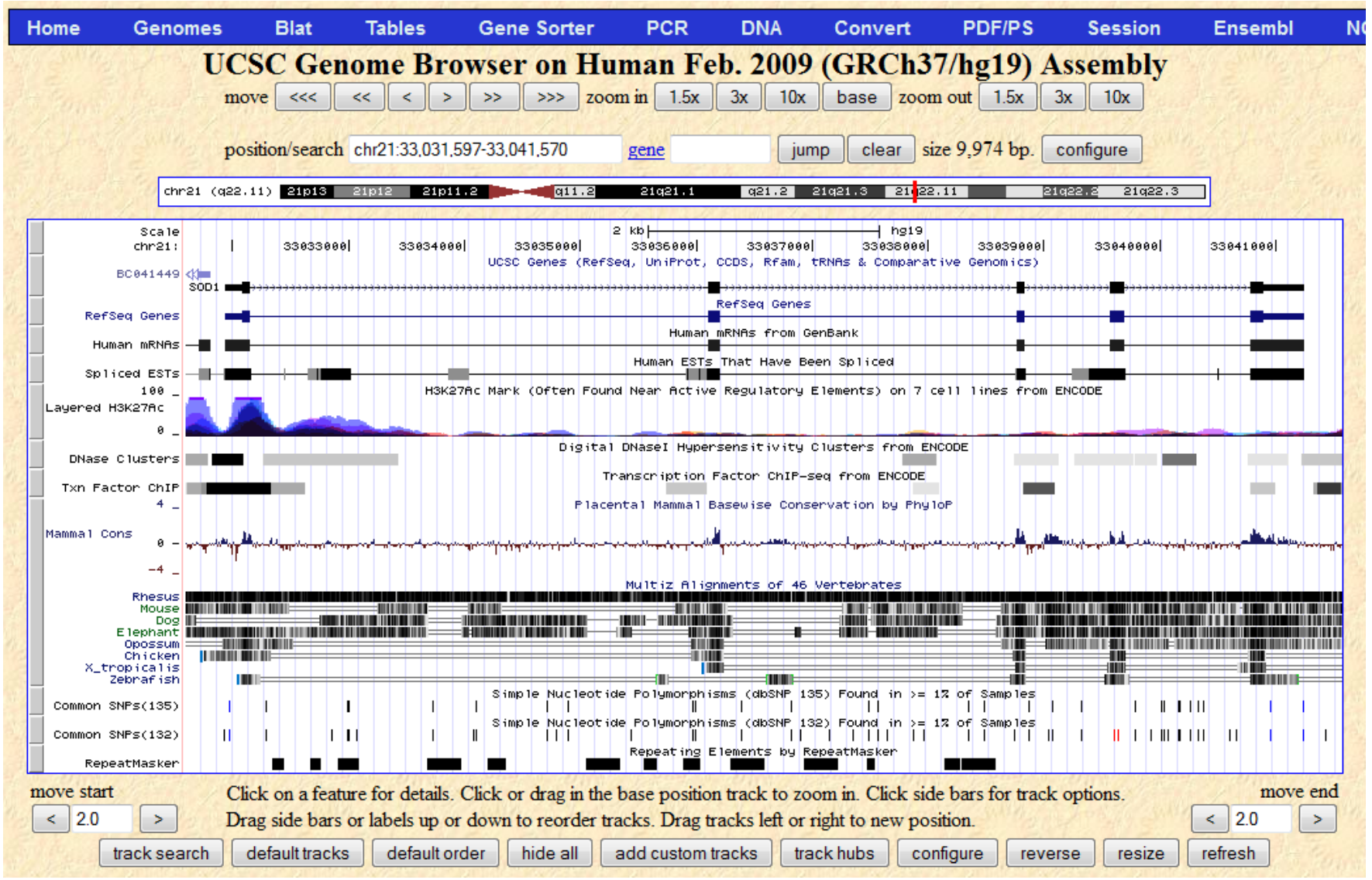
- Motivace
- Databáze sekvencí
- Genomové databáze
- Databáze struktur
- Ostatní databáze
- Integrační nástroje
- Shrnutí

Genomové databáze/prohlížeče

- Zachycují polohu významných elementů na chromozomech genomu
- Mezi tyto elementy patří např.:
  - Transkripty (mRNA, EST, tRNA, …)
  - Variabilní oblasti (SNP, Repeats, HapMap)
  - Regulační oblasti (ENCODE, Vazební místa, CpG, …)

- Nejvýznamnější databáze:
  - UCSC Genome Browsers
  - NCBI Map Viewer
  - EBI Ensembl

Genomové databáze/prohlížeče



Genomové databáze/prohlížeče

- Elementy uloženy ve formě stop (Tracků)
- Prohlížeče umožňují:
  - Snadný pohyb v rámci genomu (posun, zoom)
  - Zapnutí/vypnutí jednotlivých stop + vložení vlastních stop
  - Prohlížení elementů ve formě tabulek a jejich export (BED, GFF, …)
- Formát BED - Na každém řádku uloženy položky:
  - Chrom – označení chromozomu
  - ChromStart – počáteční pozice elementu
  - ChromEnd – koncová pozice elementu
  - Name – název elementu
  - Score – bodové hodnocení (forma zvýraznění v prohlížeči)
  - Strand – orientace vlákna
  - …

Formát BED - Příklad

- Příklad:
  - Seznam genů na lidském chromozomu 1
  - Položky v pořadí:
    - Chrom, ChromStart, ChromEnd, Name, Score, Strand

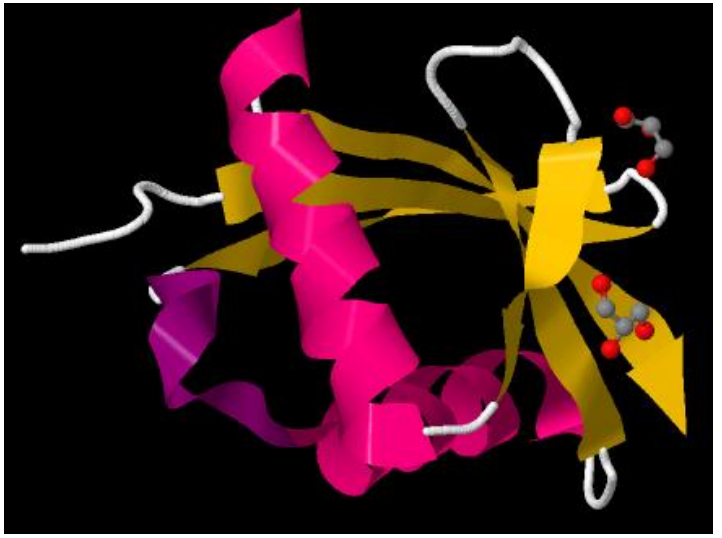
```
chr1    11873   14409   uc001aaa.3 0   +
chr1    11873   14409   uc010nxr.1 0   +
chr1    11873   14409   uc010nxq.1 0   +
chr1    14361   16765   uc009vis.3 0   -
chr1    16857   17751   uc009vjc.1 0   -
chr1    15795   18061   uc009vjd.2 0   -
chr1    14361   19759   uc009vit.3 0   -
chr1    14361   19759   uc009viu.3 0   -
chr1    14361   19759   uc001aae.4 0   -
...
```

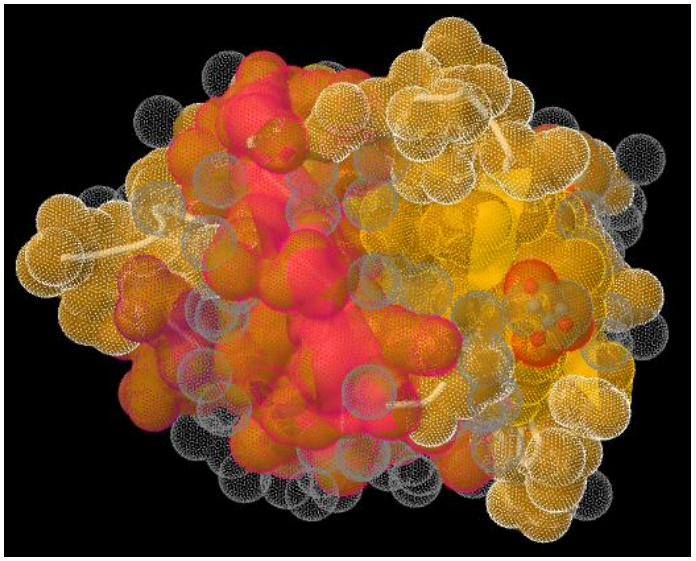
Osnova

- Motivace
- Databáze sekvencí
- Genomové databáze
- Databáze struktur
- Ostatní databáze
- Integrační nástroje
- Shrnutí

Protein Data Bank (PDB)

- Kolekce 3D struktur biologických molekul (nukleových kyselin, proteinů)
- Informace o 3D struktuře získány experimentálně technikami:
  - Rentgenové strukturní analýzy (X-ray crystallography)
  - Spektroskopie nukleární magnetické resonance (NMR spectroscopy)
- Integruje informace z:
  - PDBe – Protein Data Bank Europe
  - PDBj – Protein Data Bank Japan
  - RSCB - Research Collaboratory for Structural Bioinformatics





Protein Data Bank (PDB)

- Aktuální obsah databáze (únor 2012)

Experimental Method	Proteins	Nucleic Acids	Protein/NA complexes	Other	Total
X-ray diffraction	64758	1340	3194	2	69294
NMR	8117	966	186	7	9276
Electron microscopy	277	22	101	0	400
Hybrid	42	3	2	1	48
Other	139	4	5	13	161
Total:	73333	2335	3489	23	**79180**

PDB – formát záznamu

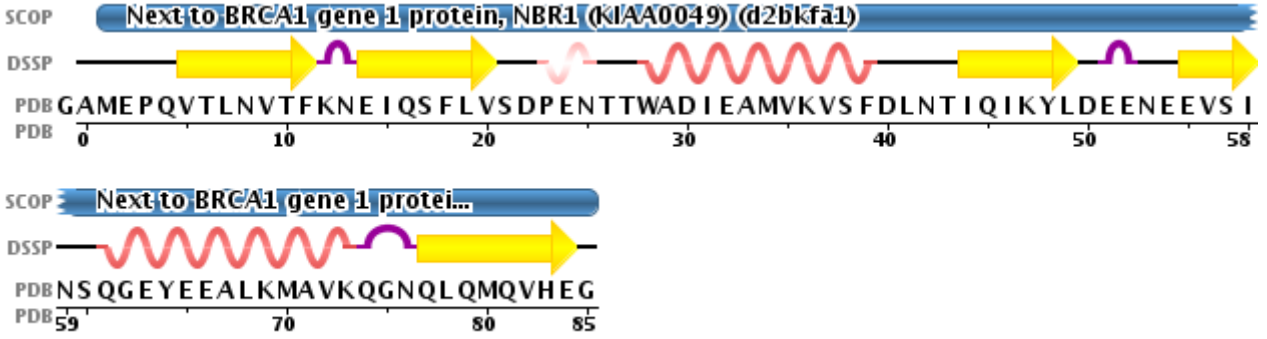
- Záznam obsahuje:
  - sekvenci aminokyselin + souřadnice jednotlivých atomů + doplňující info
- Každý řádek začíná klíčovým slovem (max. 7 znaků)
  - COMPND – z jakých molekul se protein (komplex) skládá
  - REMARK
    - odkazy na literaturu
    - informace o rozlišení se kterým byl model struktury získán
    - metoda použitá pro zjištění struktury
    - mnoho dalších informací
    - odkazy do PDB na další molekuly
  - SEQRES – sekvence aminokyselin (nukleotidů) – 3-písmenný kód
  - HELIX, SHEET, TURN – oblasti sekundární struktury
  - HET, FORMUL – popis nestandardních atomů
  - ATOM, HETATM – souřadnice jednotlivých atomů

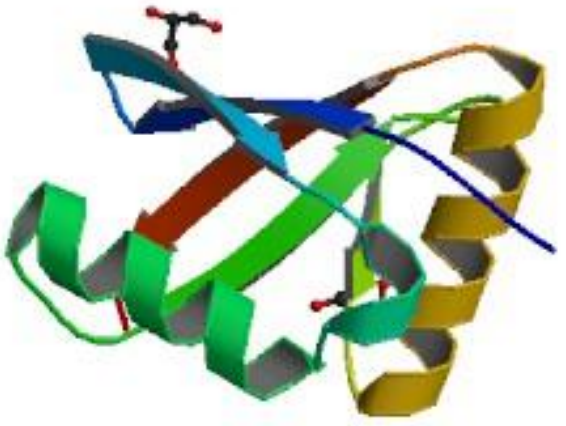
PDB – příklad záznamu

```
HEADER    ZINC-FINGER PROTEIN                      16-FEB-05   2BKF
TITLE     STRUCTURE OF THE PB1 DOMAIN OF NBR1
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: ZINC-FINGER PROTEIN NBR1 (NEXT TO BREAST CANCER 1);
COMPND   3 CHAIN: A;
COMPND   4 FRAGMENT: RESIDUES 1-85 (PB1 INTERACTION DOMAIN);
COMPND   5 SYNONYM: PB1 DOMAIN OF NBR1, NEXT TO BRCA1 GENE 1 PROTEIN, 1A1-3B;
COMPND   6 ENGINEERED: YES
...
SEQRES   1 A   87  GLY ALA MET GLU PRO GLN VAL THR LEU ASN VAL THR PHE
SEQRES   2 A   87  LYS ASN GLU ILE GLN SER PHE LEU ...
...
ATOM      1  N   ALA A   0      23.870  73.458  19.493  1.00 62.65           N
ATOM      2  CA  ALA A   0      22.457  73.932  19.574  1.00 62.81           C
ATOM      3  C   ALA A   0      22.372  75.381  19.117  1.00 61.55           C
ATOM      4  O   ALA A   0      23.329  76.134  19.344  1.00 62.58           O
ATOM      5  CB  ALA A   0      21.958  73.813  21.020  1.00 62.99           C
ATOM      6  N   MET A   1      21.283  75.825  ...
```

PDB – příklad záznamu

- Textová forma záznamu je vhodná zejména pro počítačové zpracování
- Web nabízí uživatelsky příjemnější rozhraní zahrnující
  - Interaktivní prohlížeč 3D struktury
  - Grafické schéma rozdělení proteinu na domény s vyznačením sekundárních struktur (Helix, Sheet, Turn)
  - Podobnost s ostatními proteiny
  - Odkazy na literaturu, apod.



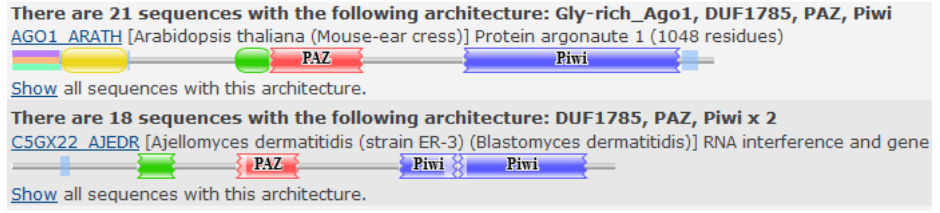


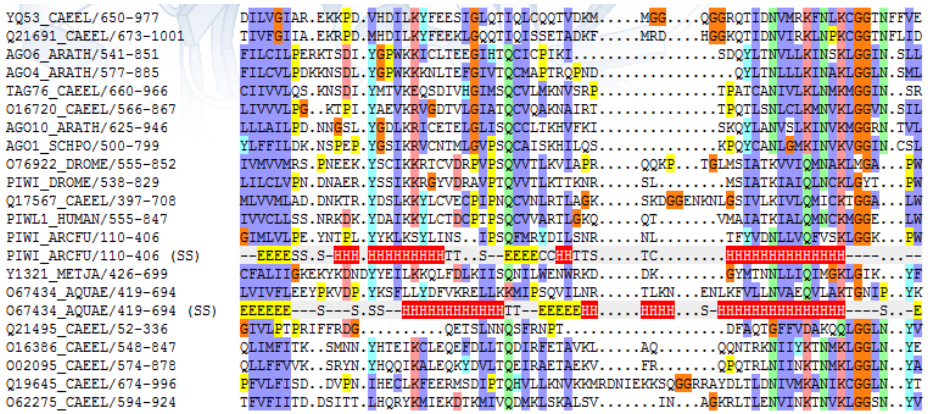
Osnova

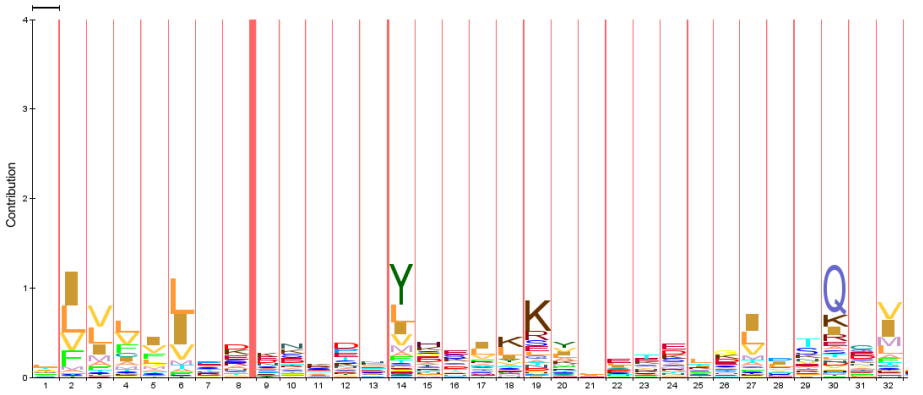
- Motivace
- Databáze sekvencí
- Genomové databáze
- Databáze struktur
- Ostatní databáze
- Integrační nástroje
- Shrnutí

Pfam – Protein families

- Databáze obsahující informace o rodinách proteinů
- Rodina proteinu – daný protein z různých organizmů
- Pfam nabízí:
  - Rozdělení na domény
  - Vícenásobné zarovnání rodiny
  - Fylogenetický strom
  - HMM motif (vyhledání pomocí HMMER)
- Data rozdělena na
  - Manuálně upravené
  - Automaticky generovaná

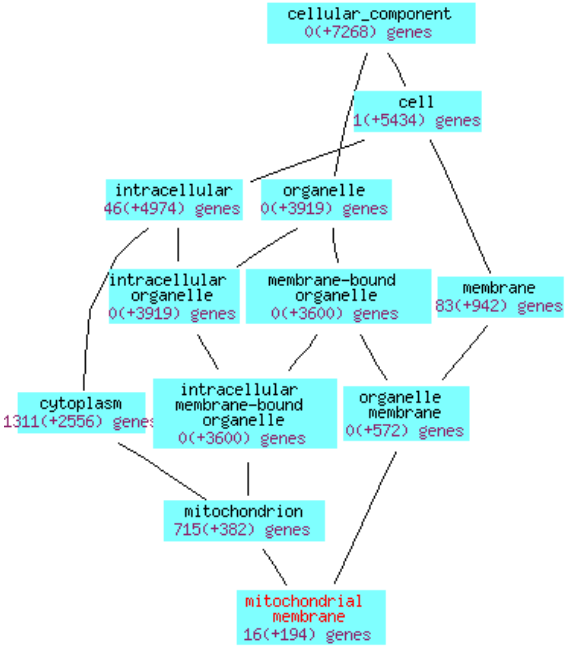


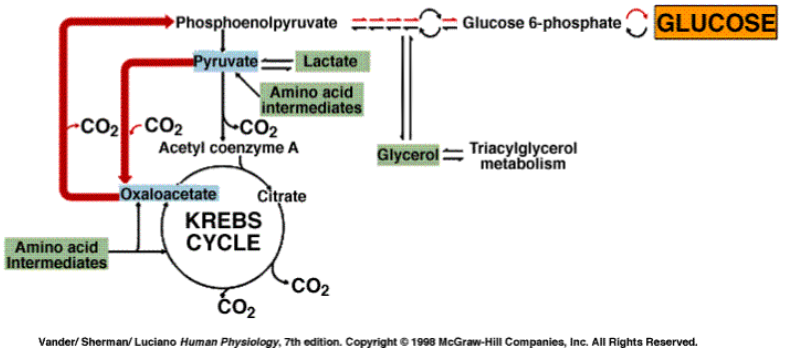




Gene Ontology

- Ontologie
  - formální reprezentace definující objekty a vztahy mezi nimi
- Gene ontology
  - Databáze popisující geny, jejich produkty a vztahy s ostatními geny
- Pro každý gen udržuje info o:
  - Molekulární funkci (co dělá)
  - Biologickém procesu (v jakém procesu)
  - Buněčné komponentě (kde)
- Data uložena ve formě acykl. grafu
- Data získaná manuálně nebo automatizovaně





Další databáze

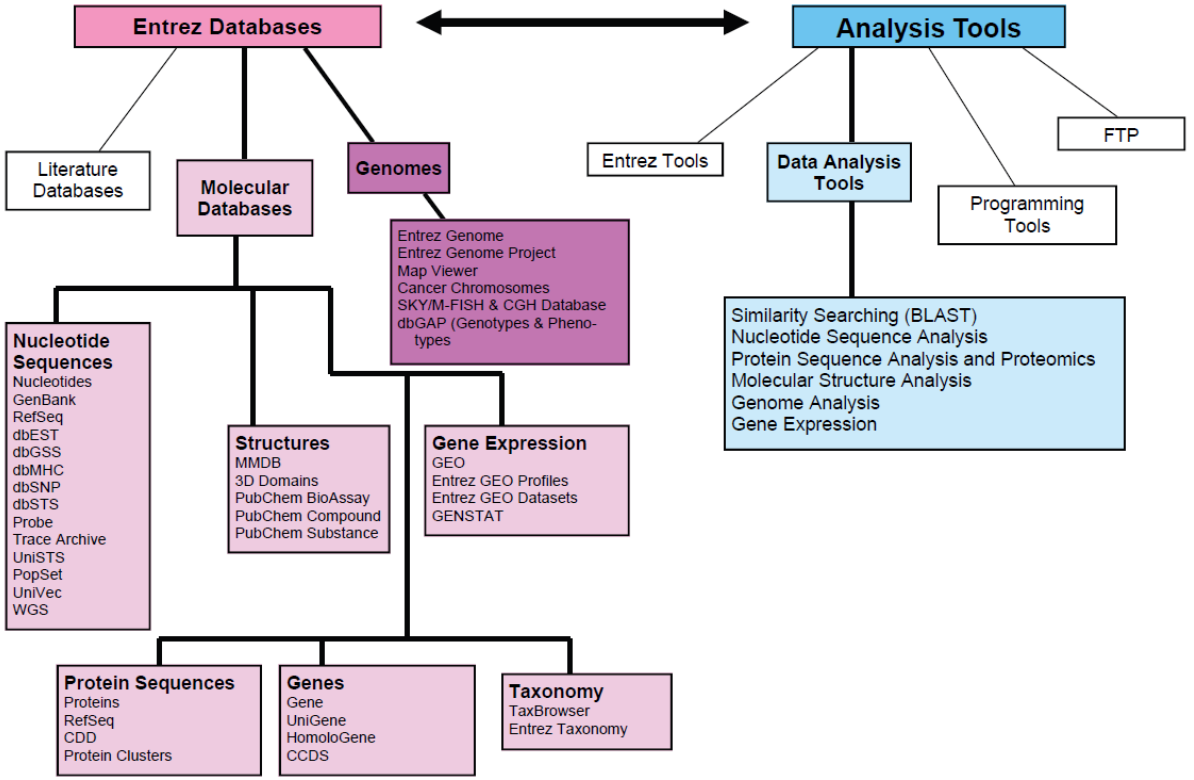
- Databáze biologické literatury:
  - PUBMED
  - MEDLINE
- Databáze lidských genů a fenotypů (genetická onemocnění):
  - OMIM – Mendeliánské choroby
  - COSMIC – mutace způsobující rakovinu
- Databáze metabolických drah:
  - KEGG
- Databáze údajů o genové expresi:
  - ArrayExpress

Osnova

- Motivace
- Databáze sekvencí
- Genomové databáze
- Databáze struktur
- Ostatní databáze
- Integrační nástroje
- Shrnutí

Integrační nástroje

- Nástroje integrující data z různých databází
- Umožňují získat co nejvíce informací na jednom místě
  - EBI SRS
  - NCBI Entrez
- Integrace probíhá nejen na úrovni databází, ale i na úrovní softwarových nástrojů pracujících s daty



Osnova

- Motivace
- Databáze sekvencí
- Genomové databáze
- Databáze struktur
- Ostatní databáze
- Integrační nástroje
- Shrnutí

Shrnutí

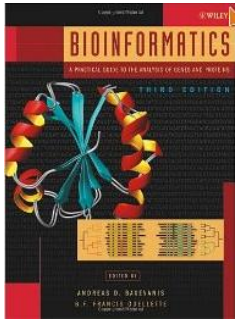
- Existuje velké množství databází pro uložení biologických dat na různých úrovních
  - Sekvence, Struktury, Genomy, Ontologie, ...
- Při práci s daty je potřeba rozlišovat, zda se jedná o experimentálně získaná a manuálně upravená data, nebo o predikovaná data získaná automatizovaně
- Pro počítačové zpracování jsou data k dispozici v textových souborech různých formátů
- Pro běžné uživatele nabízí obvykle databáze graficky přívětivé webové rozhraní

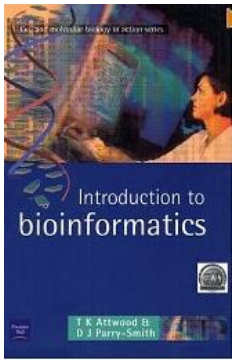
Shrnutí

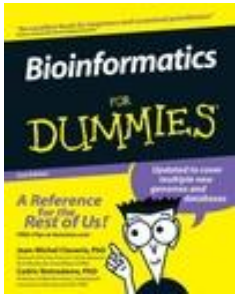
- Primární organizace spravující data (NCBI, EBI, NIG) obvykle kooperují a data si navzájem sdílí
- Vznikají také nástroje, které se snaží integrovat a propojit různá biologická data (NCBI Entrez, EBI SRS)
- Integrace probíhá nejen na úrovni databází, ale i na úrovni softwarových nástrojů pracujících s daty

Literatura

- Knihy:
  - A. D. Baxevanis, B. F. F. Ouellette: *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*
  - T. K. Attwood, D. J. Parry-Smith: *Introduction to Bioinformatics*
  - Jean-Michel Claverie, Cedric Notredame: *Bioinformatics for Dummies*







Konec

Děkuji za pozornost