IV107 Bioinformatika I

Přednáška 1

Katedra informačních technologií
Masarykova Univerzita Brno

Podzim 2024

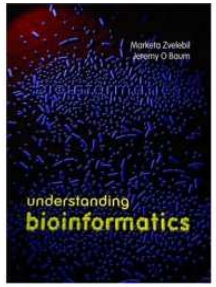
Outline

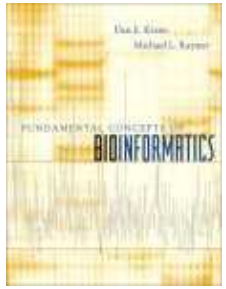
Kontaktní údaje

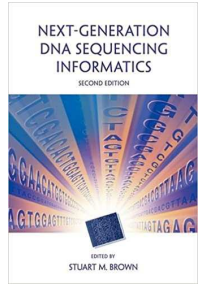
- Dr. Matej Lexa, C202 (lexa@fi.muni.cz)
- Přednáška Po 14:00-15:50 (někdy 16:40) (A318)
- Konzultace St 14:30-15:30
- https://www.muni.cz/lide/31298-matej-lexa/vyuka

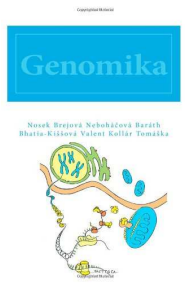
Studijní literatura

1. Zvelebil and Baum (2007).
   Understanding bioinformatics, Garland Science, Oxford, 772 s. (ISBN: 0-8153-4024-9)
2. Krane and Raymer (2005).
   Fundamental concepts in bioinformatics, Benjamin Cummings, London, 320 s. (ISBN 0-8053-4633-3)
3. Nosek et al. (2013).
   Genomika, CreateSpace Independent Publishing Platform, Bratislava, 276 s. (ISBN: 978-1493731336)
4. Stuart M. Brown (2015).
   Next-Generation DNA Sequencing Informatics, 2nd edition, CSHL Press, 402 s. (ISBN: 978-1621821236)









Vědecké časopisy

- Bioinformatics
- BMC Bioinformatics
- J. of Bioinformatics and Computational Biology
- Briefings in Bioinformatics
- Evolutionary Bioinformatics
- GigaScience
- InSilico Biology
- Více na https://en.wikipedia.org/wiki/List_of_bioinformatics_journals

Bioinformatika na FI

- Bakalářská úroveň jako zaměření a magisterská jako specializace
- Předpokládá se vypracování bioinformatické závěrečné práce
- Od roku 2025 nový studijní program FI+PřírF
- https://bioinf.pages.fi.muni.cz/
- https://is.muni.cz/auth/kruh/biotika Bioinformatika@FI Muni
- Další vyučující: doc.Vít Nováček, PhD - doc.RNDr.David Šafránek, PhD - doc.RNDr.Barbora Kozlíková, PhD

Navazující předměty FI

- IV108 - Bioinformatika II (St 12:00 A219)
- IV105/IV106 - Seminář z bioinformatiky Bc/Mgr (St 16:00 A319+MS Teams)
- IV110/IV114 - Projekt z bioinformatiky a systémové biologie (Po 10:00 B410+CEITEC)
- PB051 - Výpočetní metody v bioinformatice a systémové biologii (jaro)
- PV269 - Pokročilé metody bioinformatiky (jaro)

Příbuzné předměty FI

- IV109 - Modelování a simulace
- IV117/8 - Systémová biologie
- PB172 - Seminář ze systémové biologie
- PA183 - Projekt ze systémové biologie
- PV287 - Artificial Intelligence and Machine Learning in Healthcare
- PV251 - Visualization
- PA055 - Vizualizace komplexních dat

Harmonogram kurzu

- Rychlý úvod do molekulární biologie (do poloviny října)
- Semestrální test (konec října)
- Základní oblasti bioinformatiky, datová a algoritmická řešení v nich

Klasifikace

- Hodnotí se
  - Semestrální test 20 bodů
  - Zkouška 80 bodů
- Klasifikační stupnice
  - A 90 - 100
  - B 80 - 89
  - C 70 - 79
  - D 60 - 69
  - E 50 - 59
  - F méně než 50

Outline

Definice bioinformatiky

Bioinformatika
Studuje metody shromáždďování, spřístupňování a analýzy rozsáhlých souborů biologických dat, zejména molekulárně – biologických.

Další disciplíny

- Výpočetní nebo matematická a systémová biologie
  matematické přístupy k reprezentaci a zkoumání biologických procesů, často simulace
- Lékařská informatika
  práce s medicínskými daty, převážně záznamy pacientů
- Genomika
  Experimentální zjišťování sekvencí DNA celých genomů
- Proteomika
  Experimentální zjišťování složení a funkce souborů proteinů

Předmětem zájmu nebo používanými metodami se bioinformatika prolíná s

1. molekulární biologií
2. genomikou a proteomikou
3. genetikou
4. výpočetní biologií
5. matematickou či teoretickou biologií
6. systémovou biologií
7. biomedicínskou informatikou
8. biomedicínským inženýrstvím
9. výpočetní chemií
10. informatikou
11. výpočetní lingvistikou

Částečně převzato z http://cz.wikipedia.org/wiki/Bioinformatics 26.2.2018

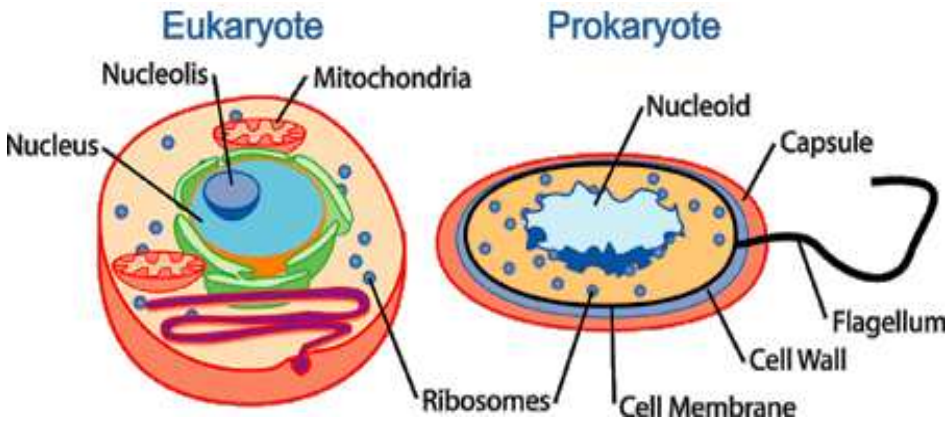
Typické okruhy problémů

- Analýza sekvencí
- Sekvenování a anotace genomů
- Evoluční bioinformatika
- Studium biodiverzity / metagenomika
- Analýza exprese genů
- Analýza genové regulace
- Analýza proteomu
- Odhad struktury proteinů
- Srovnávací genomika
- Modelování biologických systémů
- Analýza obrazu
- Studium strukturních interakcí proteinů

Částečně převzato z http://en.wikipedia.org/wiki/Bioinformatics 26.2.2018

Outline

Buňka – základní forma organizace živé hmoty



- Molekuly (DNA, proteiny, sacharidy, lipidy)
  Geny (abstraktní pojem)
- Proteinové komplexy/membrány
- Organely a jiné substruktury
- Buňka
- Tkáň/pletivo
- Organizmus

Složitost biologických systémů na molekulární úrovni

Člověk: cca 10^{14} buněk.

Genom buňky: 3×10^9 párů nukleotidů DNA (A:T a C:G).

Nukleotidy: vytváří sřetězenými kombinacemi cca 20000 genů (a statisíce jiných funkčních míst)

Geny: kódují (a aktivitou vytváří) staticíce molekul (proteinů a RNA)

Aminokyseliny: vytváří sřetězenými kombinacemi statisíce proteinů

Buňka: aktivuje v daném momentu určitou podmnožinu této sady

Výsledek: obrovské množství možných stavů buněk (2^{20000} je velmi podceňující odhad)

Geny: evolucí vybrané sady z cca 4^{1000} možných sekvencí DNA (1000 nukl./gen)

Outline

Bioinformatická data

- Sekvence DNA a RNA
- Sekvence proteinů
- Struktura proteinů
- Fenotypy (mutantů), klinická data
- Údaje o aktivitě genů microarray, RNA-Seq
- Údaje o stavu chromatinu (metylace, 3D) (ChIP-seq, Hi-C)
- Údaje o expresi proteinů imunodetekce, 2-D gely, hmotn.spektrometrie (MS)
- Mapy interakcí mezi proteiny a DNA - Chip-Seq
- Mapy interakcí mezi proteiny navzájem - ”yeast two-hybrid”
- Literatura

Sekvenční data

AUGACAGUUGACGAGUGCA

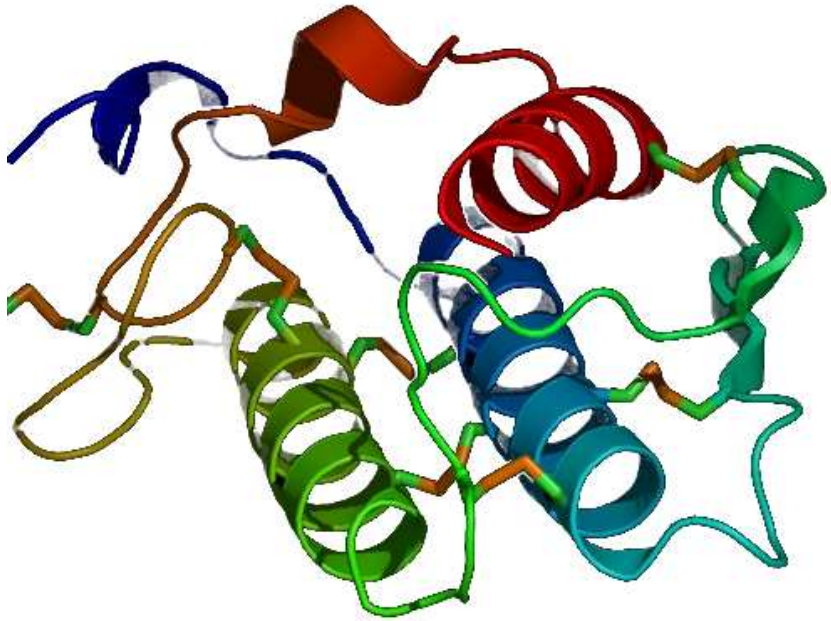
ATAGCAGTGCGCATGCAGT

MASAQSFYLLMDDHLAVFM

Sekvenční data

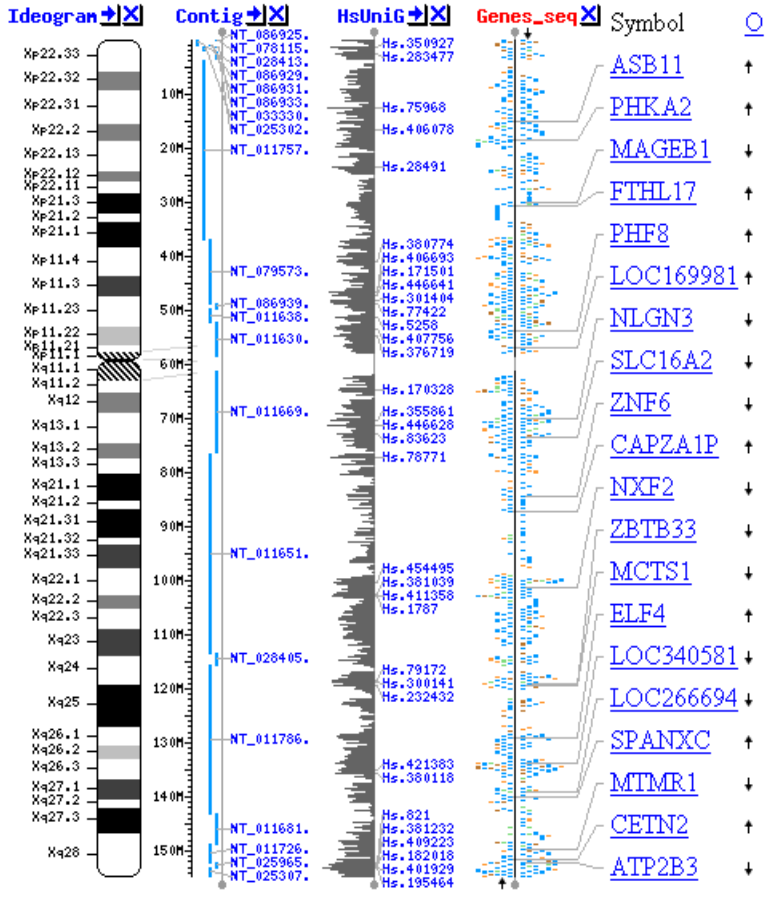
DNA	ATAGCAGTGCGCATGCAGT
RNA	AUGACAGUUGACGAGUGCA
Protein	MASAQSFYLLMDDHLAVFM

Strukturní data



Zobrazení struktury proteinu

Spřístupnění dat uživatelům – NCBI Genome Viewer



Zobrazení informací o genech na chromozomu

Spřístupnění dat vývojářům

- Grafika je sekundární. Prvořadá je rychlost a možnost automatizace manipulace s daty. API velkých portálů jako NCBI, KEGG, ENSEMBL, UNIPROT.
- Datový sklad BioMart
- BioJava, BioPerl, BioPython, Bioconductor (R) a další knihovny pro většinu jazyků a prostředí
- servery poskytující syrová data (holý text, obrázky, XML a jiné struktury přes HTTP, SOAP, ODBC, DAS, REST)
  https://www.sciencedirect.com/science/article/pii/S2001037015000471
- Beacon API
  https://www.ga4gh.org/news/new-release-of-ga4gh-beacon-expands-genomic-and-clinical-data-access/
- Data obohacena o semantiku (Ontologie, RDF triples/grafové databáze)

Outline

Stopy bioinformatiků v latině

et tu brutus *in vino veritas* *veni vidi vici*	
in vivo *in vitro* *in silico*	biolog biochemik bionformatik

Práce bioinformatika

- Umí pracovat s velkými datovými soubory
- Moudrými triky ovláda výkonné počítače
- V datech hledá zajímavé vzory nebo subsekvence
- Srovnává podobné vzory a sekvence
- Skládá genomy z kratších fragmentů
- Předpovídá strukturu a funkci genů a proteinů
- Studuje vývoj sekvencí a organizmů
- Vytváří intuitivní nástroje a reprodukovatelné výpočetní postupy
- Data a výsledky analýz zobrazuje graficky

Způsob nahlížení na data

KLASIK	směs biologie, chemie, fyziky atd.
MECHANIK	živé buňky jsou stroje, které chceme pochopit a ovládat
HRÁČ	sekvence jsou definiční soubory hráčů
SEMIOTIK	život je signalizace a interpretace signálů
LINGVISTA	sekvence se skládají z modulů (slov) s určitou funkcí vykazujících gramatické uspořádání
INFORMATIK	Buňky jsou počítače s hardwarem (molekulární stavba = proteom a další molekuly a organely) a softwarem (genetická informace = genom)

Outline

Kořeny a zdroje bioinformatiky

1951	Pauling	struktura proteinů
1952	Turing	chem. základy vývoje
1953	Watson, Crick, Franklin	struktura DNA
1956	Gamow et al.	genetický kód
1959	Chomsky	gramatiky
1962	Shannon a Weaver	informační teorie
1966	Martin-Lof	náhodné řetězce
1966	Neumann	automata
1969	Britten a Davidson	génová regulace

Historie bioinformatiky do sformování disciplíny

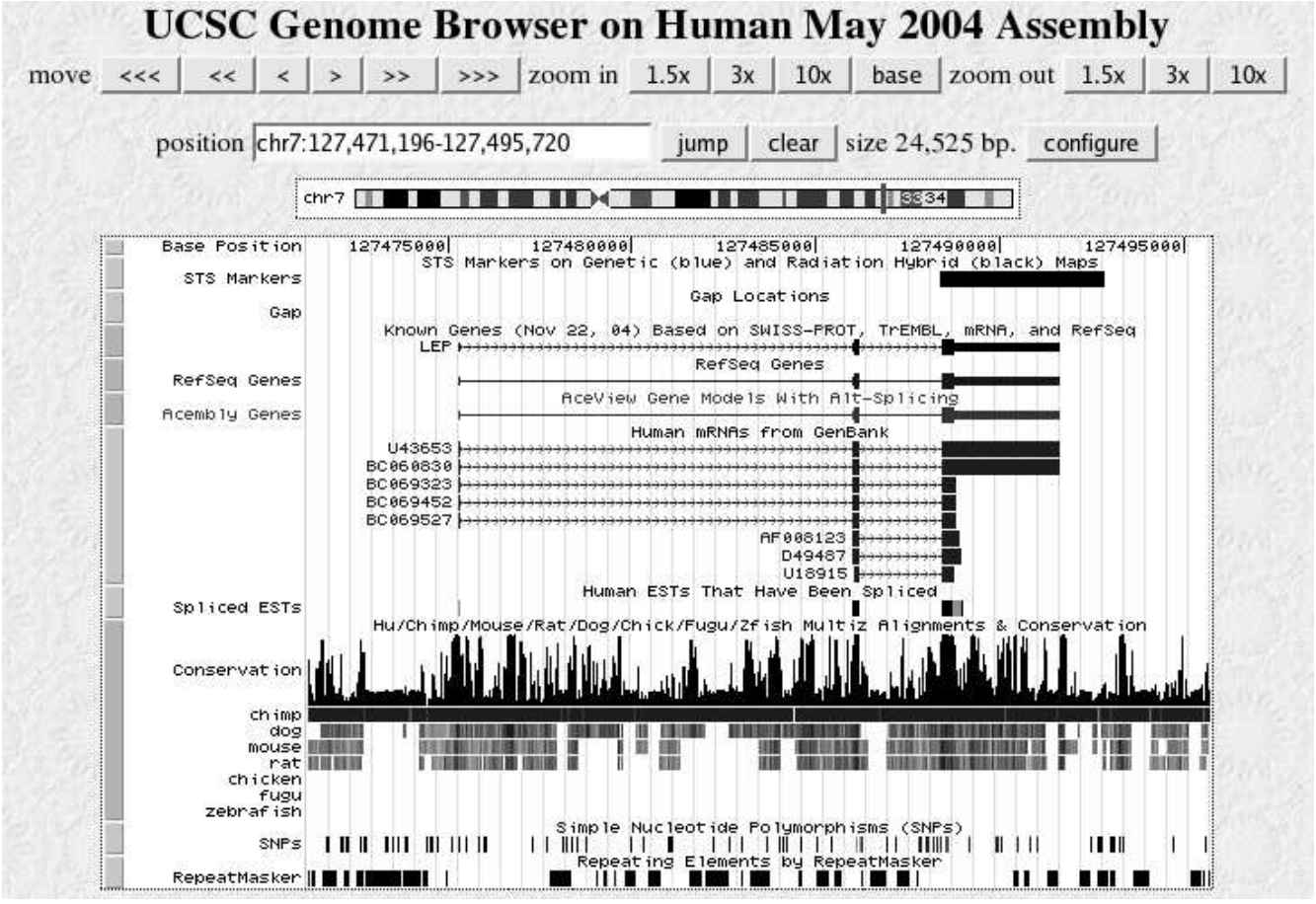
- 1967 Fitch and Margoliash: sestrojení prvních fylogenetických stromů z biologické sekvence
- 1970 Needleman and Wunsh: zarovnání dvou sekvencí
- 1974 Chou and Fasman: predikce sekundární struktury proteinů
- 1978 Dayhoff: první sbírka sekvencí proteinů
- 1981 Kabsch and Sander: modelování struktury proteinů
- 1987 Feng and Doolitle: mnohonásobné zarovnání sekvencí
- 1990 Altschul et al.: efektivní hledání lokálních podobností
- 1998 The Journal Comp Appl Biosci se přejmenovává na Bioinformatics

Outline

- Jim Kent – autor Aegis Animator, Cyber Paint a Autodesk Animator
- po shlédnutí 12-ti CD vývojového prostředí Windows 95 přechází k bioinformatikům s posteskem, že lidský genom se vejde na jedno CD
- autor webové aplikace Genome Browser
- sehrává důležitou roli v honičce o přečtení a skompletování lidského genomu (program GigAssembler)

Převzato z Jim Kent: ”The Genes, the Whole Genes, and Nothing But the Genes”, BioCon 2003.

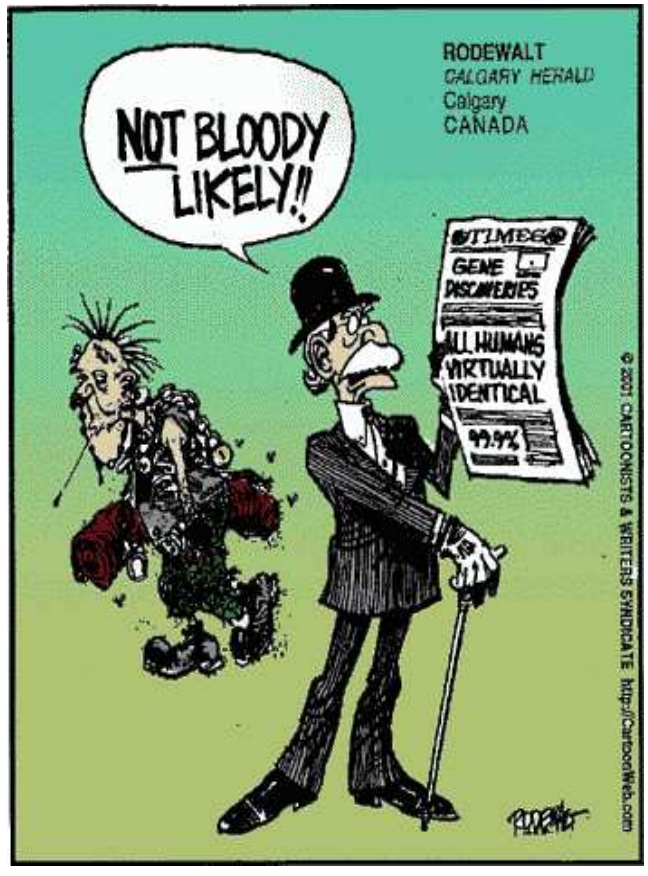
UCSC Genome Browser



Flexibilní nástroj určen k interaktivnímu prohlížení genomů

Homo/Homo

- rozdíl každých 1000 nukleotidů
- 90% variace je mezi africkými populacemi
- na Zemi je tolik lidí a četnost mutací je tak vysoká, že každý ze jmenovaných nukleotidů je v dané generaci mutován několik krát



- lidský genom obsahuje stovky nepříjemných mutací. Většina je recesivních, projeví se jenom ojediněle, pokud je mají oba rodiče

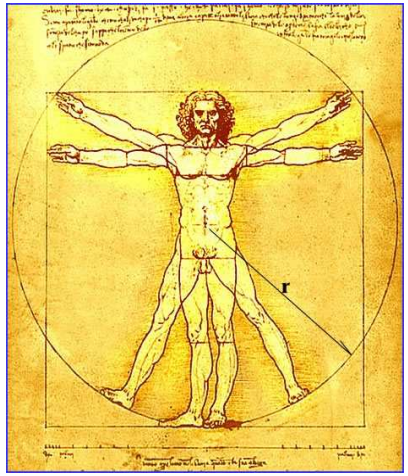
Převzato z Jim Kent: ”The Genes, the Whole Genes, and Nothing But the Genes”, BioCon 2003.

Homo/Pan

- rozdíl každých 100 nukleotidů
- transpozon každých 50000 nukleotidů
- dva chromozomy spojené, jinak podobná struktura

Podle Jim Kent: ”The Genes, the Whole Genes, and Nothing But the Genes”, BioCon 2003.

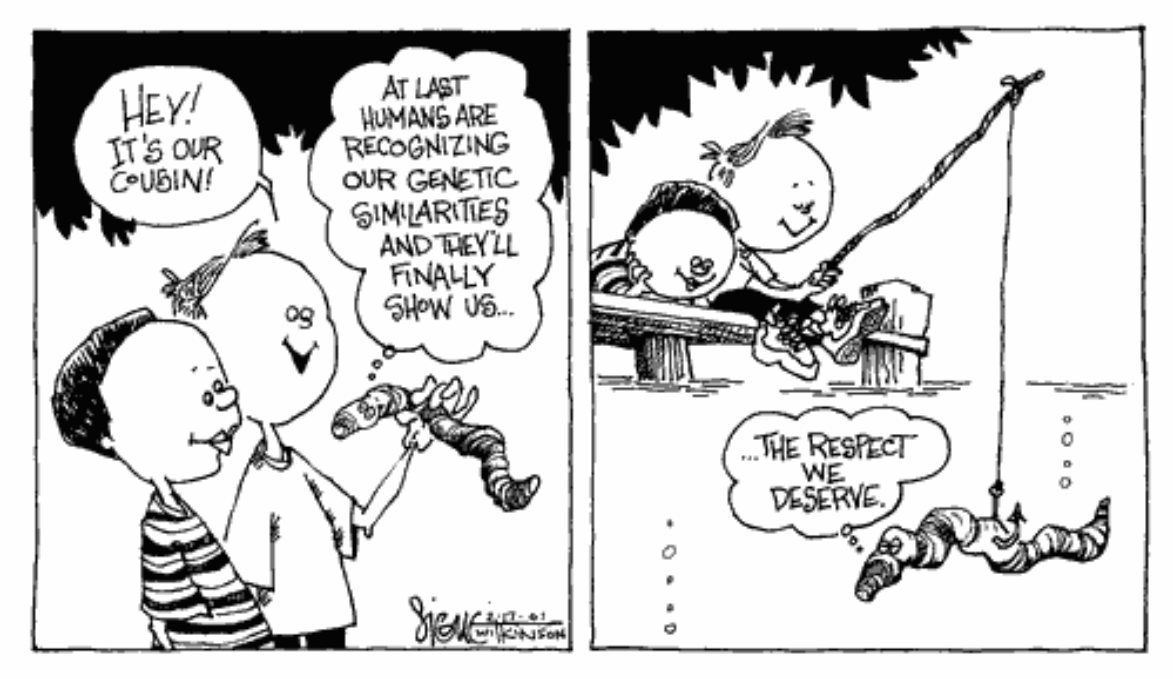
Homo/Mus



- 40% nukleotidů byli od dob společného předka změněny
- Ve funkčních oblastech se změnilo jenom 15% nukleotidů
- úseky podobnosti mezi genomy člověka a myši jsou kandidáti na biologické funkce

Převzato z Jim Kent: ”The Genes, the Whole Genes, and Nothing But the Genes”, BioCon 2003.

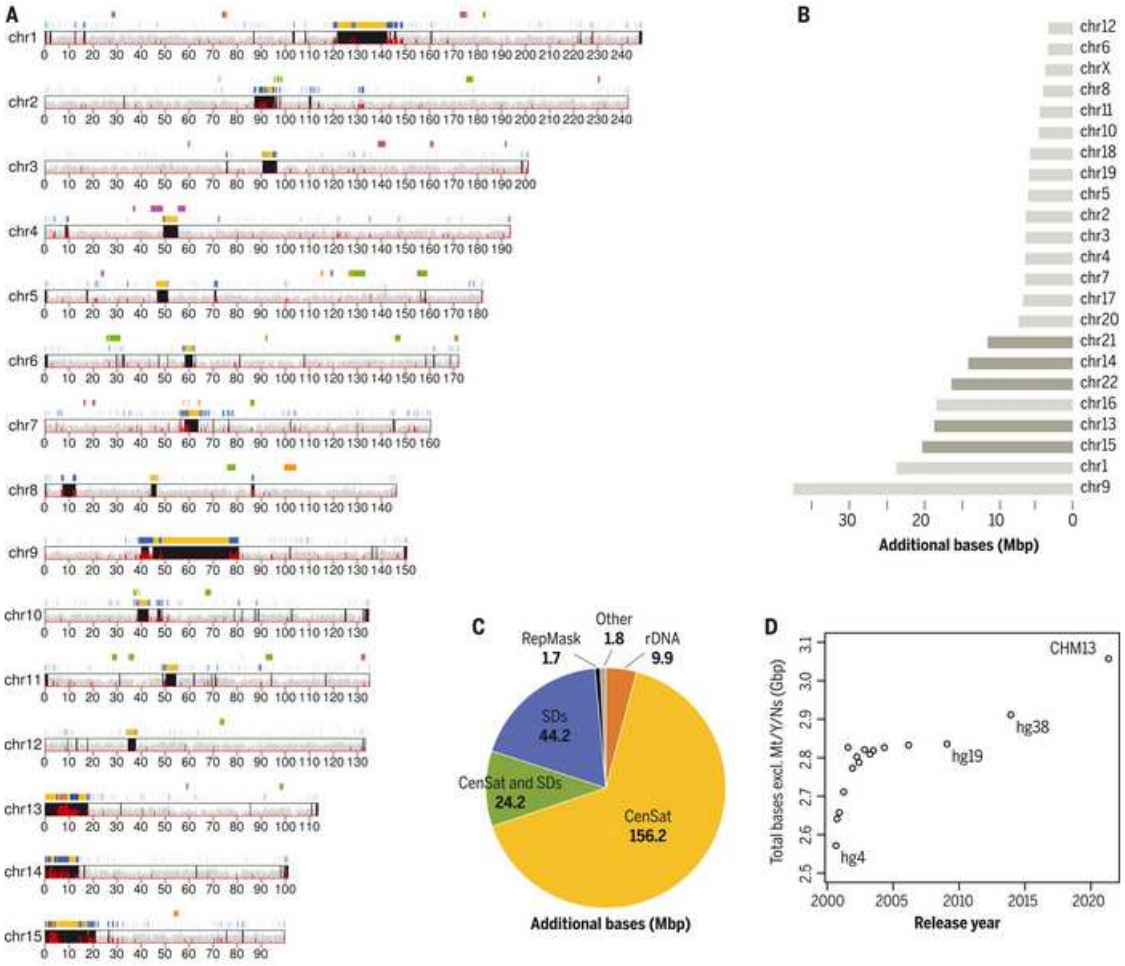
Homo/Caenorhabditis



Asi 80% nukleotidů změněno (35% ve funkčních oblastech)

Převzato z Jim Kent: ”The Genes, the Whole Genes, and Nothing But the Genes”, BioCon 2003.

Telomere to Telomere (T2T) human genome



Nové metody sekvenace (nanopórová, PacBio Hifi, Hi-C) umožňují skládat kompletní chromozomy

https://www.science.org/doi/10.1126/science.abj6987

Outline

Objem dat bude nadále narůstat

- Základní výzkum
- Medicína a jiné aplikace
- Bezpečnost na molekulární úrovni
- Komerční data

V současnosti např. nastupuje ”osobní genomika”, ”sekvenování jednotlivých buněk”, ”3-D genomika”

HT-Seq/NGS technologie

How a Genealogy Website Led to the Alleged Golden State Killer

Powerful tools are now available to anyone who wants to look for a DNA match, which has troubling privacy implications.

SARAH ZHANG APR 27, 2018

Press conference announcing the capture of Joseph DeAngelo (FRED GREAVES)

When the East Area Rapist broke into the home of his first victim in 1976, human DNA had not yet been sequenced. When he reemerged as the Original Night Stalker and began a spree of murders in 1979, the World Wide Web still did not exist. For decades, the Golden State Killer—as he is now best known—got away with it all.

Then DNA and the internet appear to have caught up. Reporting from *The Sacramento Bee* and *Mercury News* indicates that police arrested Joseph James DeAngelo based on DNA found at crime scenes that partially matched the DNA of a relative on the open-source genealogy website GEDmatch. Previous searches of law-enforcement DNA databases had turned up no matches.

MORE STORIES

Solving a Murder Mystery With Ancestry Websites
CIARA O'ROURKE

The False Promise of DNA Testing
MATTHEW SHAER

1

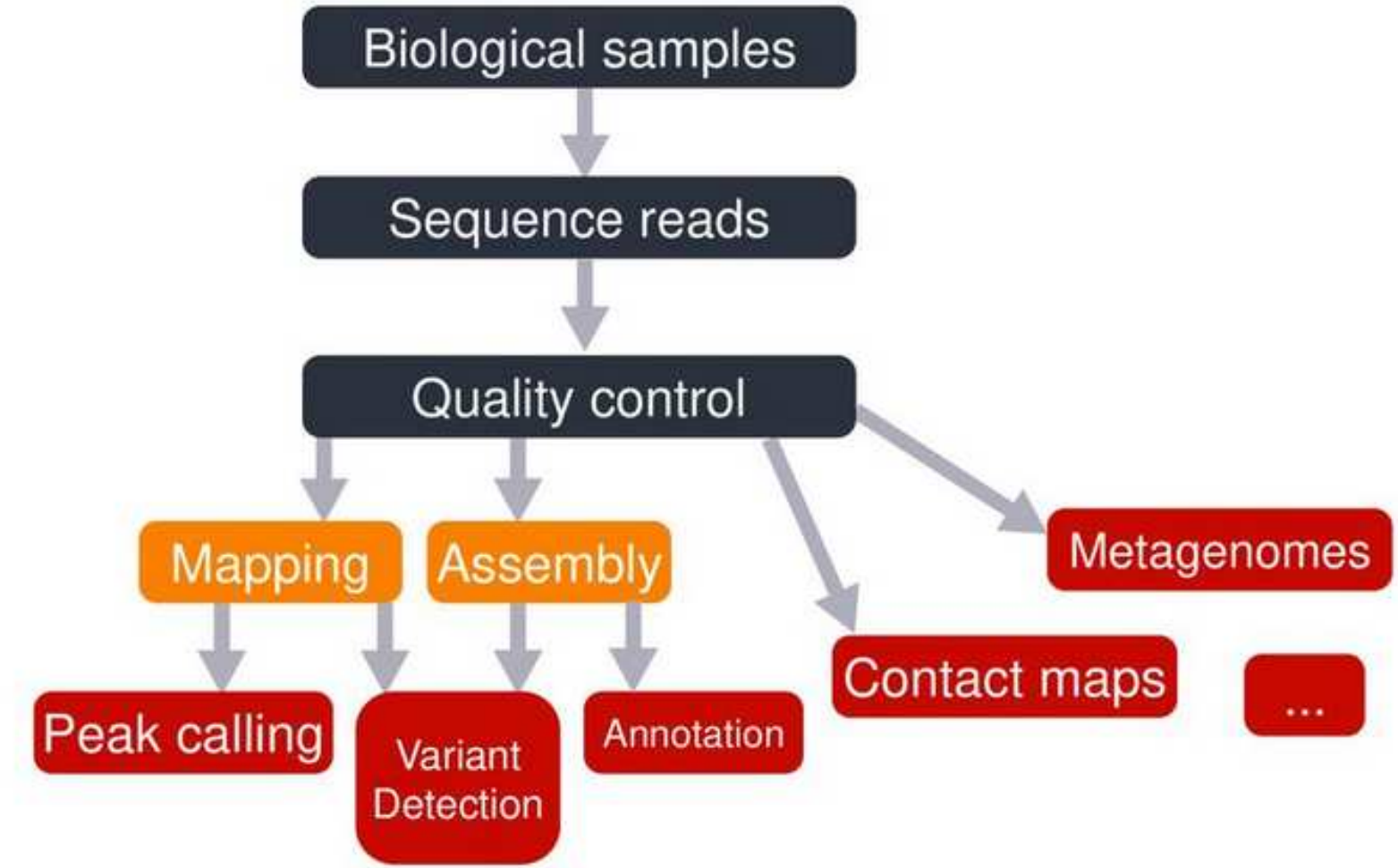
[1] https://www.theatlantic.com/science/archive/2018/04/golden-killer-east-area-rapist-dna-genealogy/559070/

HT-Seq/NGS technologie

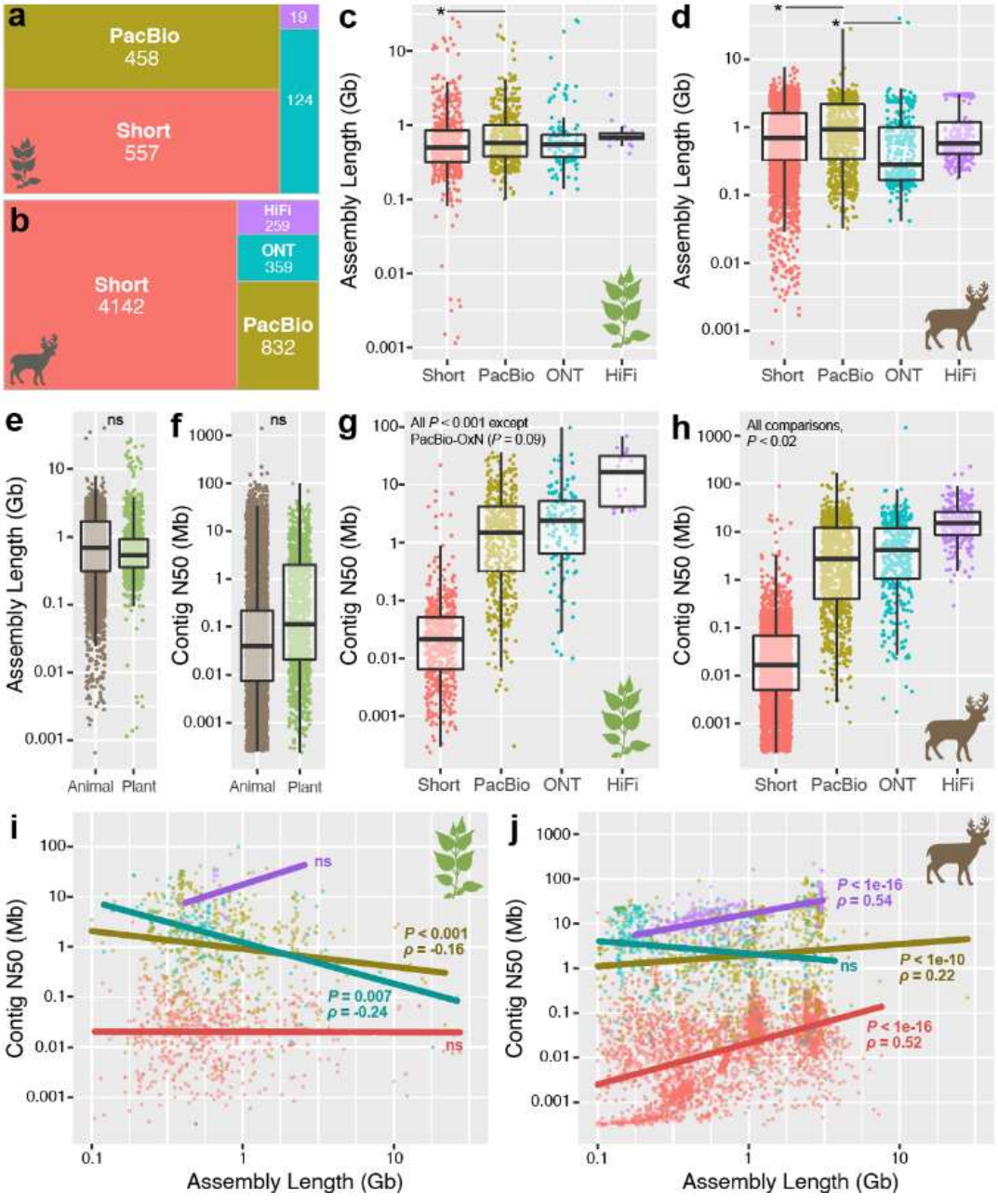
- **Solexa pyrosequencing (Illumina)**
- 454 (Roche)
- SOLiD (Life Technologies)
- Heliscope (Helicos, mrtvá technologie)
- Ion Torrent
- Polonator (Dover/Danaher Motion, otevřená platforma)
- Max-Seq (Intelligent Biosystems/Dover/Azco Biotech)
- **Zero-mode waveguide sequencing, te HiFi (Pacific Biosciences)**
- Nanoball sequencing (CompleteGenomics, jen jako služba)
- FRET sequencing (Visigen)
- **Nanopore sequencing (Oxford Nanopore)**

http://cen.acs.org/articles/92/i33/Next-Gen-Sequencing-Numbers-Game.html
https://whatisbiotechnology.org/index.php/science/summary/nanopore

NGS technologie - bioinformatické zpracování



NGS technologie - srovnání (doi:10.1101/2022.07.10.499467)



Hotaling et al. (2022)

Porovnávání sekvencí

```
>P11633 NONHISTONE CHROMOSOMAL PROTEIN 6B.
Score = 54.8 bits (155), Expect = 1e-10 Identities = 19/43
(46%), Positives = 24/43 (62%)
Query: 2 TKKFKDPNRPPSAFFLFCSEYRKIKGEHPGLSIGDVAKKLGEM 52
:         T : KDPNR  SA:  F :E R I  E:P :: G V : LGE
Sbjct: 5 TTRKKDPNRGLSAYMFFANENRDIRSENPDVTFGQVGRILGER 55
```

Analogie biosekvence – jazyk

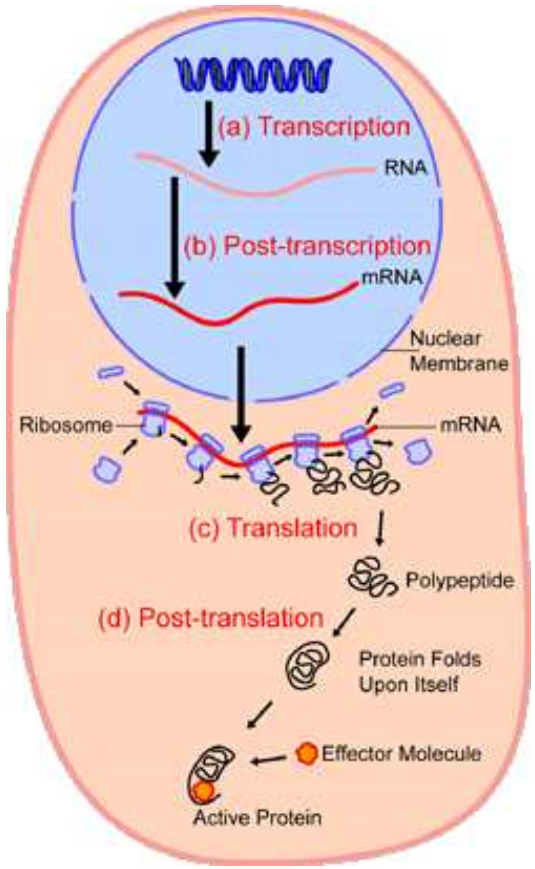
1. Mam z toho velkou radost.
2. Mam toho kocoura dost.

```
Mamztohovelk  ouradost.
::: ::::   :  ::::::::
Mam toho   kocouradost.
```

Outline

Informace v sekvenci proteinů se neodráží v sekvenci DNA nebo RNA

Informace v DNA určuje existenci proteinů v buňce



Příště struktura DNA a proteinů

- Struktura DNA
- Struktura proteinů
- Přenos genetické informace

Outline

Outline

FACULTAS ARTIS INFORMATICÆ · UNIVERSITAS MASARYKIANA

Outline