

# Úvod do bioinformatiky

Bioinformatika

Tomáš Martínek

[martinto@fit.vutbr.cz](mailto:martinto@fit.vutbr.cz)

# Informace o kurzu

---

- **Webová stránka:**
  - <http://www.fit.vutbr.cz/study/courses/BIF/>
  - <https://www.fit.vutbr.cz/study/courses/BIF/private/>
- **Garant předmětu:**
  - Prof. Ing. Lukáš Sekanina, Ph.D., [sekanina@fit.vutbr.cz](mailto:sekanina@fit.vutbr.cz)
- **Přednášející a cvičící:**
  - Ing. Tomáš Martínek, Ph.D., [martinto@fit.vutbr.cz](mailto:martinto@fit.vutbr.cz)
  - Ing. Jaroslav Bendl, [ibendl@fit.vutbr.cz](mailto:ibendl@fit.vutbr.cz)
  - Ing. Ivan Vogel, [ivogel@fit.vutbr.cz](mailto:ivogel@fit.vutbr.cz)
- **Bodové hodnocení:**

– Cvičení:	12 bodů
– Projekt:	14 bodů
– Půlsemestrální zkouška:	16 bodů
– Semestrální zkouška:	58 bodů

# Přednášky

---

- **Program:**

1. Úvod do bioinformatiky
2. Základy molekulární biologie, nástroje mol. biologie
3. Databáze biologických dat
4. Zarovnání sekvencí, dynam. prog., BLAST, FASTA
5. Vícenásobné zarovnání sekvencí
6. Predikce sekundární struktury RNA
7. **Půlsemestrální zkouška (20. 3. 2012)**
8. Konstrukce fylogenetických stromů
9. Konstrukce fylogenetických stromů
10. Sekvenování DNA
11. Genomika a rozpoznávání genů
12. Proteiny a predikce jejich struktury
13. Proteomika, regulační sítě

# Cvičení

---

- Program:

1. Databáze biologických dat
2. Analýza genomových sekvencí
3. Zarovnání sekvencí
4. Konstrukce fylogenetických stromů
5. Analýza struktury proteinů
6. Analýza struktury proteinů

- Termíny:

- Bude upřesněno

# Literatura

---

- **Knihy:**

- Dan K. Krane, Michael L. Raymer: *Fundamental Concepts of Bioinformatics*
- Jean-Michel Claverie, Cedric Notredame: *Bioinformatics for Dummies*
- Neil C. Jones, Pavel A. Pevzner: *An Introduction to Bioinformatics Algorithms*
- Marketa Zvelebil, Jeremy O. Baum: *Understanding Bioinformatics*

- **Odkazy na Internetu:**

- [NCBI](#)
- [Bioinformatics Algorithms](#)
- [EMBL- EBI](#)

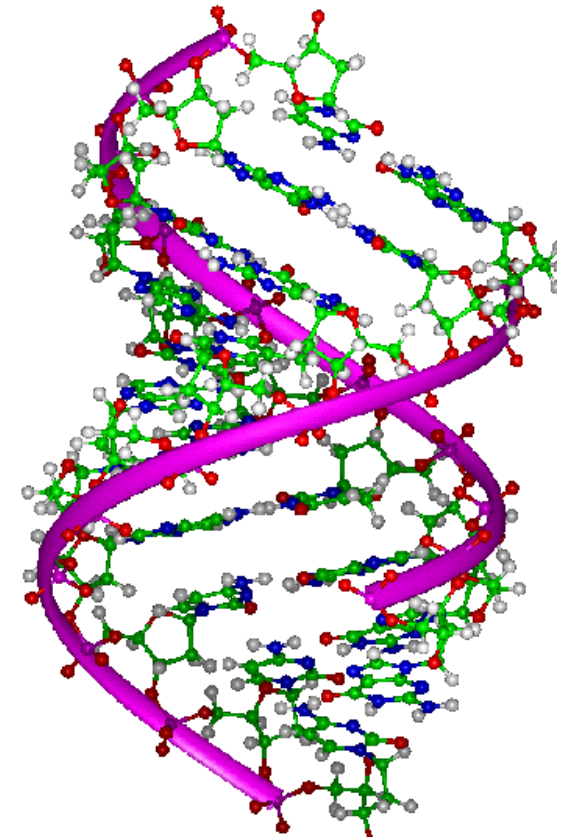
# Proč bioinformatika?

---

- Jakými problémy se zabývá molekulární biologie?
  - Genom, DNA
  - Centrální dogma molekulární biologie
  - Budoucnost biologie
- Jak získat DNA?
  - Technologie sekvenování
  - Významná sekvenovací centra
- Jak interpretovat informace uložené v DNA?
  - Výzkumná oblast
  - Klinická praxe
  - Komerční sféra
- Proč bioinformatika?

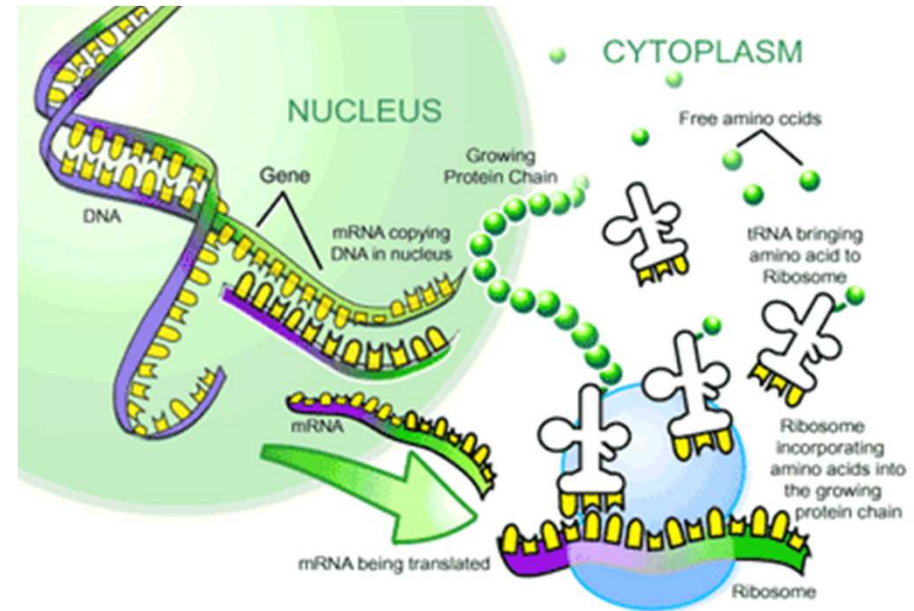
# Genom, DNA

- **Genom:**
  - Obsahuje kompletní genetickou informaci daného organismu (jaderná DNA, mitochondriální DNA, ...)
- **DNA:**
  - Dvoušroubovice složená z nukleotidů **Adenim**, **Guanin**, **Thymin** a **Cytosin** (A, C, G, T)
  - DNA molekula člověka obsahuje zhruba **3G bází**
- **Rozdíly v DNA:**
  - člověk-člověk: **0.1%**
  - člověk-šimpanz: **1%**
- **Uložení DNA:**
  - Každá báze lze zakódovat do **2 bitů** – pro uložení celého genomu stačí **750 MB**
  - Změny oproti referenční DNA – **100-vky kB**
- **Možné využití:**
  - Doktor, zaměstnavatel, pojišťovna :-)



# Centrální dogma molekulární biologie

- DNA obsahuje specifické podřetězce zvané **geny**, které dokáží syntetizovat **proteiny**
- **Proteiny**
  - účastní se většiny biologických procesů uvnitř buňky
  - řídí biochemické reakce (**enzymy**)
  - zastávají strukturní nebo mechanické funkce
  - aktivují nebo potlačují funkci jiných proteinů
- Pouze cca **1,5%** lidské DNA tvoří **geny kódující proteiny**
- Konzervovaná oblast genomu člověka reprezentuje cca **10%** - obsahuje převážně **regulační oblasti**



obr. Transkripce a Translace

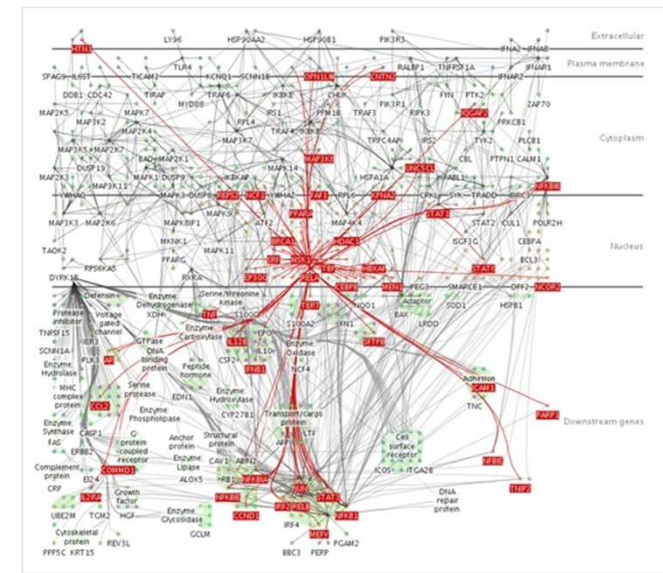
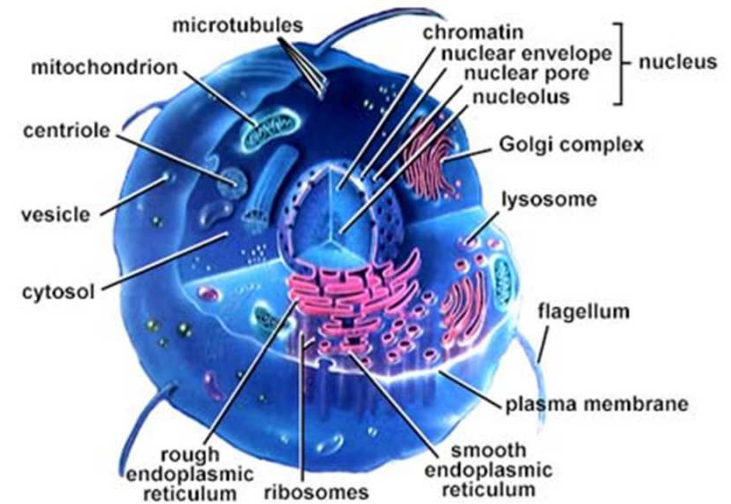


obr. Reprezentace proteinu



# Budoucnost biologie – výzkumná část

- Výzkumná část
  1. Na-sekvenování vzorku tkáně
  2. Identifikace všech genů, proteinů a predikce jejich 3D struktury a funkce
  3. Identifikace regulačních oblastí a predikce jejich aktivity v různých buňkách a vývojových stádiích
  4. Rekonstrukce modelu buňky a sestavení simulačního modelu
  5. Rekonstrukce modelu celého organismu a sestavení simulačního modelu



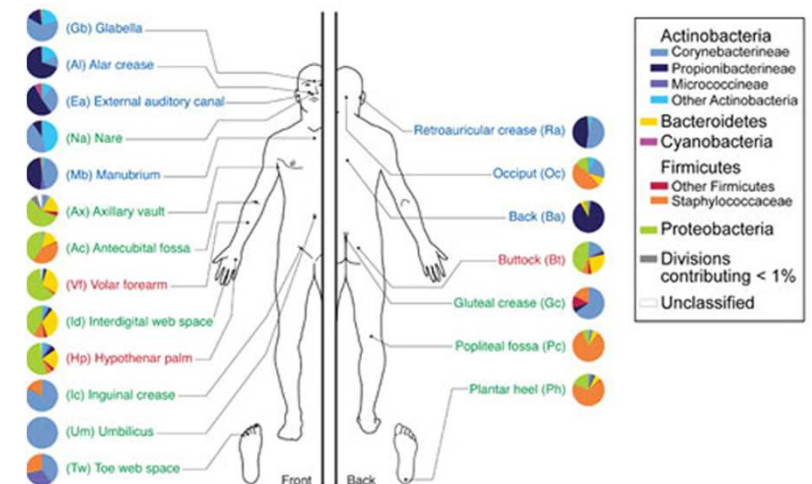
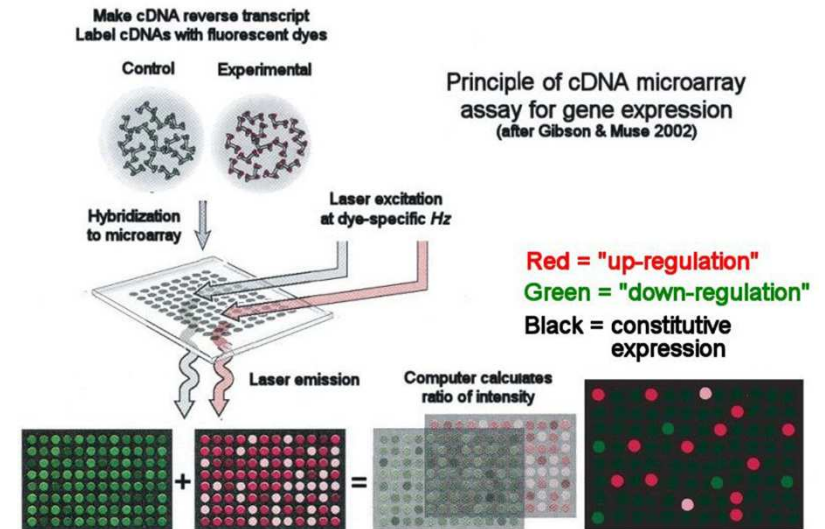
# Budoucnost biologie – klinická praxe

- ## Klinická praxe

1. Sestavení diagnózy na základě:

- **Genomu** (analýza souboru mutací a predikce jejich vlivu na chování organismu)
- **Aktuálního stavu organismu** (interakce na molekulární úrovni, míra exprese genů v různých částech, regulační aktivita)
- **Imunitního repertoáru** (čtením vzorku B a T buněk, sledování reakcí na vakcínu)
- **Mikrobiomu** (soubor mikrobů, které ovlivňují činnost organismu)

2. Aplikace léčby s ohledem na potřeby konkrétního pacienta



# Proč bioinformatika?

---

- Jakými problémy se zabývá molekulární biologie?
  - Genom, DNA
  - Centrální dogma molekulární biologie
  - Budoucnost biologie
- Jak získat DNA?
  - Technologie sekvenování
  - Významná sekvenovací centra
- Jak interpretovat informace uložené v DNA?
  - Výzkumná oblast
  - Klinická praxe
  - Komerční sféra
- Proč bioinformatika?

# Technologie sekvenování

- Výrazným způsobem ovlivňuje vývoj v celé oblasti biologie a biotechnologií
- Pro masivní nasazení je potřeba se dostat na hranici **\$1000 za genom**

## 1. Generace

- Sangerova metoda
- Cena genomu: \$1M

## 2. Generace (současný stav)

- 454 Life Science, Roche
- HiSeq2000, Illumina
- SOLiD 5500 System, Ion Torrent Systems
- Cena genomu: \$10.000 (akt. \$4000)

## 3. Generace ( fáze prototypu)

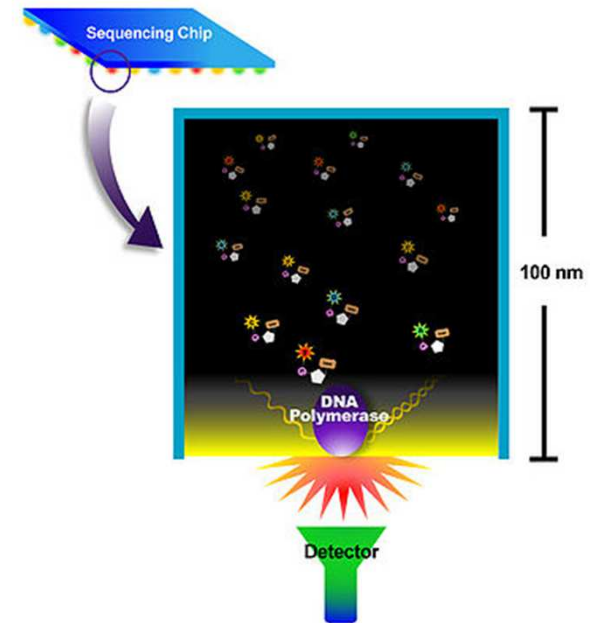
- Pacific Biosciences, SMRT
- Oxford Nanopore Technologies
- Cena genomu: \$100
- **Očekávané komerční nasazení 2014**



# Sekvenovací technologie 3. generace

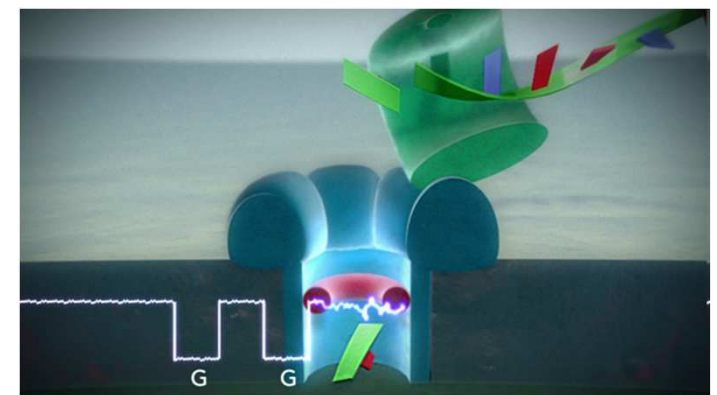
- **Pacific Biosciences**

- Využívá DNA Polymerázu, která je schopna replikovat celou molekulu DNA v rámci minut
- Fluorescenčně obarvené nukleotidy: každý typ nukleotidu jinou barvičkou.
- Jakmile se nukleotid naváže na vlákno barvivo se uvolní a vytvoří intenzivní záblesk
- Celý proces probíhá v nádobce o šířce 70nm, uzavřené sklíčkem a snímané citlivým senzorem



- **Oxford Nanopore Technologies**

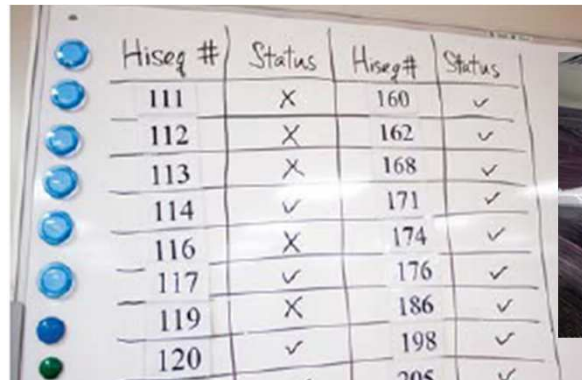
- Molekula DNA prochází přes úzké místo (nanopore) o velikosti jednotek nanometrů
- Nanopore měří velikost nápoje, která je pro jednotlivé nukleotidy různá





# Beijing Genomics Institute (BGI)

- “We will sequence all the genetics in the world”
- Největší centrum pro sekvenování genomů na světě
- Vznik: rok 2010
- 150 přístrojů HiSeq2000
- 5000 zaměstnanců
- Plný provoz
  - 5TB/den  
(cca 1500 lidských genomů)
  - 50.000 CPU
  - 200 TB RAM
  - 1PB - 1EB data storage
  - \$10M/ročně jenom na elektřině



Hiseg #	Status	Hiseg #	Status
111	X	160	✓
112	X	162	✓
113	X	168	✓
114	✓	171	✓
116	X	174	✓
117	✓	176	✓
119	X	186	✓
120	✓	198	✓
		205	✓



# Proč bioinformatika?

---

- Jakými problémy se zabývá molekulární biologie?
  - Genom, DNA
  - Centrální dogma molekulární biologie
  - Budoucnost biologie
- Jak získat DNA?
  - Technologie sekvenování
  - Významná sekvenovací centra
- Jak interpretovat informace uložené v DNA?
  - Výzkumná oblast
  - Klinická praxe
  - Komerční sféra
- Proč bioinformatika?

# Interpretace genomových dat

- *„We are close to having a \$1,000 genome sequence, but this may be accompanied by a \$1,000,000 interpretation.“*



Bruce Korf  
President of the American College of Medical Genetics

- *„The cost of DNA sequencing might not matter in a few years. But the interpretation of the data will be keeping us busy for the next 50 years.“*



Russ Altman  
Stanford's Chair of Bioengineering



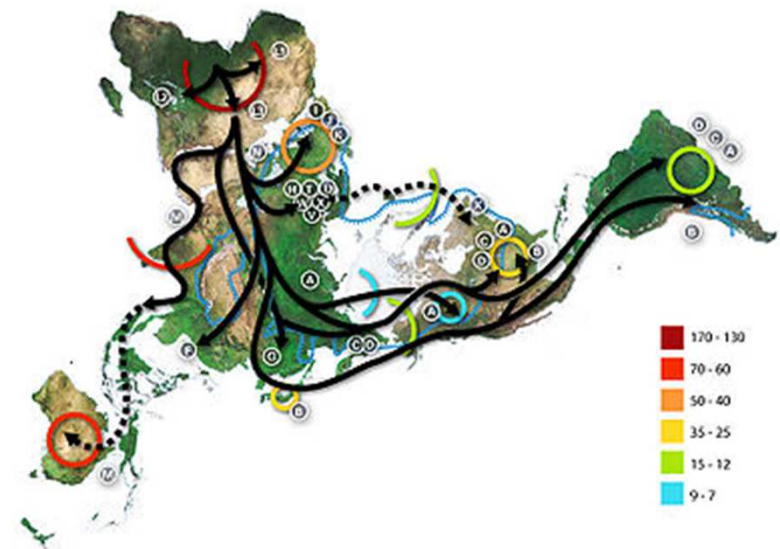
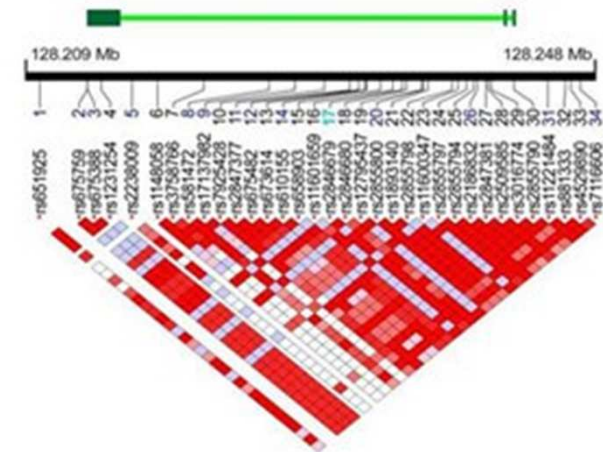
# Oblast výzkumu

---

- **Anatomie a fyziologie lidského genomu**
  - Hlavní část byla doposud věnována kódujícím oblastem (cca 1,5% genomu)
  - Bylo identifikováno cca 21 000 kódujících genů (u většiny z nich prozatím nevíme co dělají)
  - Analýzou konzervovanosti však bylo zjištěno, že funkční část genomu tvoří cca 10%
  - Hlavní část nekódujících konzervovaných oblastí se podílí na regulaci celého procesu
- **Projekt ENCODE** (the ENCyclopedia Of DNA Elements)
  - Konsorcium výzkumných pracovišť, jejichž cílem je systematické zmapování všech funkčních elementů lidského genomu
  - 2003 – 2007: pilotní nasazení
  - 2007 – současnost: produkční fáze

# Oblast výzkumu

- **Genomové variace**
  - Studium změn mezi genomy (v rámci populace stejných i různých organismů)
  - Sledování jednobodových **SNP** mutací i statisticky sdružených skupin – **haplotypy**.
  - V lidské populaci **500k – 1 milión** SNP reprezentuje cca **90%** genetické variace v populaci
  - **Pomáhá:**
    - Pochopit funkce některých oblastí genomu
    - Při identifikaci mutací způsobujících choroby
    - Odhalit historii vývoje lidstva



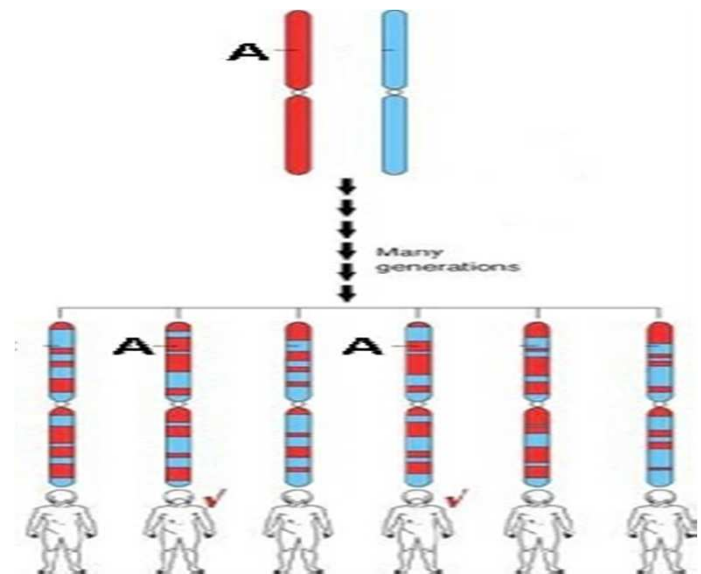
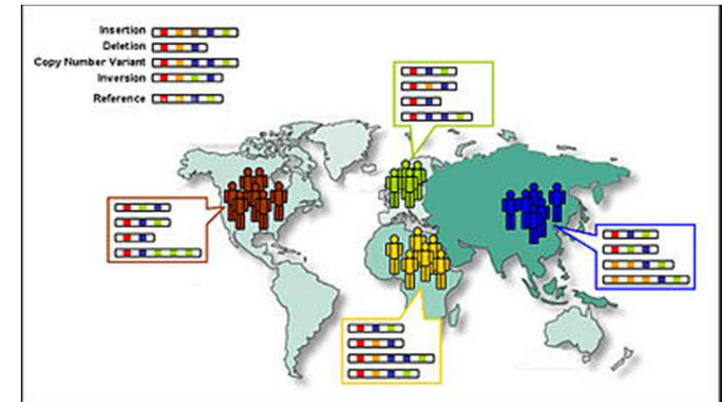
# Oblast výzkumu

- 1000 Genomes Project

- <http://www.1000genomes.org/>
- Výzkum variací v genomech v souvislosti s lidskými chorobami, chováním a evolucí
- Snaha najít všechny varianty se zastoupením  $>1\%$  v genomu a  $>0,1\%$  v kódujících oblastech
- 2008 - 2010: pilotní nasazení - 1000 genomů (sekvenování exomů)
- 2010 – současnost: produkční fáze: 2000 genomů

- HapMap Project

- <http://hapmap.ncbi.nlm.nih.gov/>
- Cílem je zmapovat všechny haplotypy v rámci lidského genomu a nalézt genetické varianty, které ovlivňují zdraví, nemoci, reakce na léky a vliv prostředí
- Eviduje až 3 milióny SNP



# Oblast medicíny

- Klasifikace chorob

- Podmíněné prostředím

- Způsobené vnějšími vlivy: bakterie, viry
    - **Příklady:** Chřipka, HIV, různé druhy infekcí

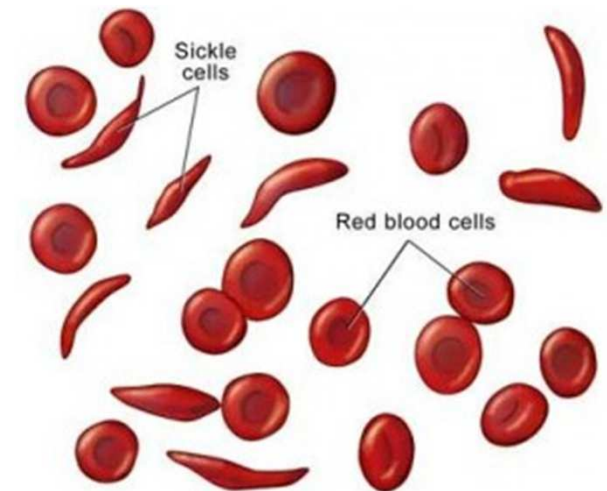
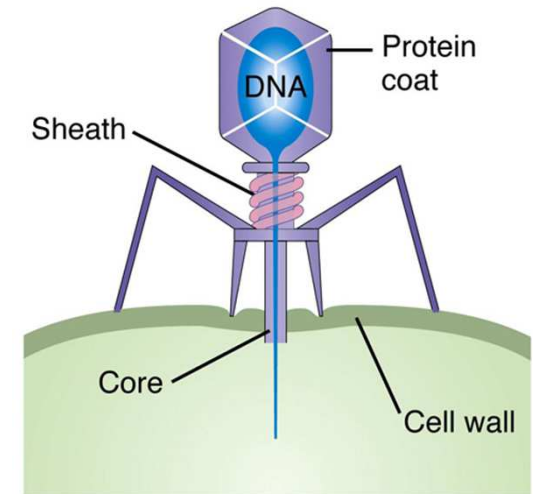
- Podmíněné geneticky

- Mendelovské choroby

- Vznikají poruchou jednoho genu
      - Dědičné
      - **Příklady:** Cystická fibróza, Huntingtonova nemoc, Hemofilie, srpková anémie, apod.

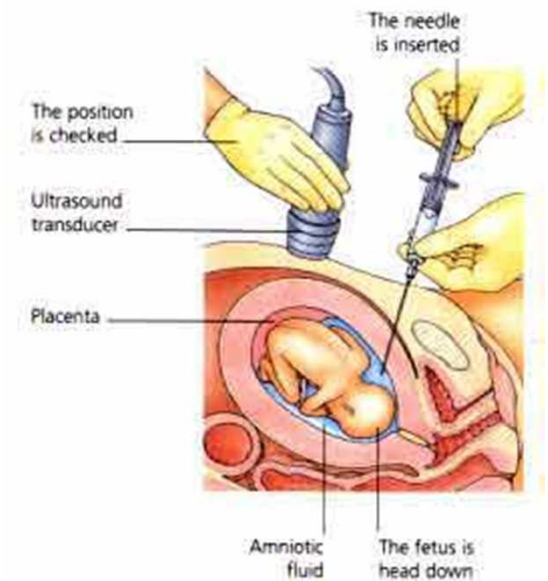
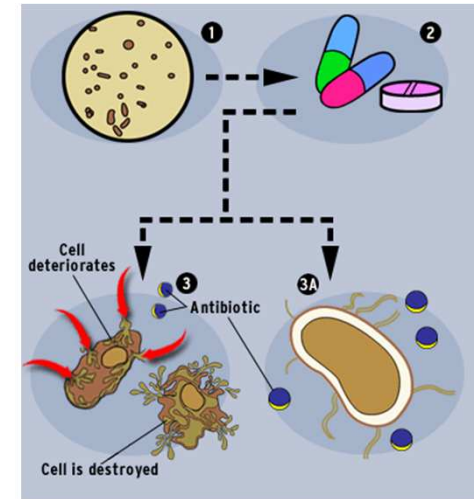
- Polygenetické choroby

- Mají základ v genech, ale projevují se pouze za určitých podmínek
      - Obvykle nejsou dědičné nebo ji nelze potvrdit
      - **Příklady:** Rakovina, cukrovka 2. typu, náchylnost na srdeční zástavu, Kronova nemoc, autoimunitní choroby, schizofrenie, apod.



# Oblast medicíny

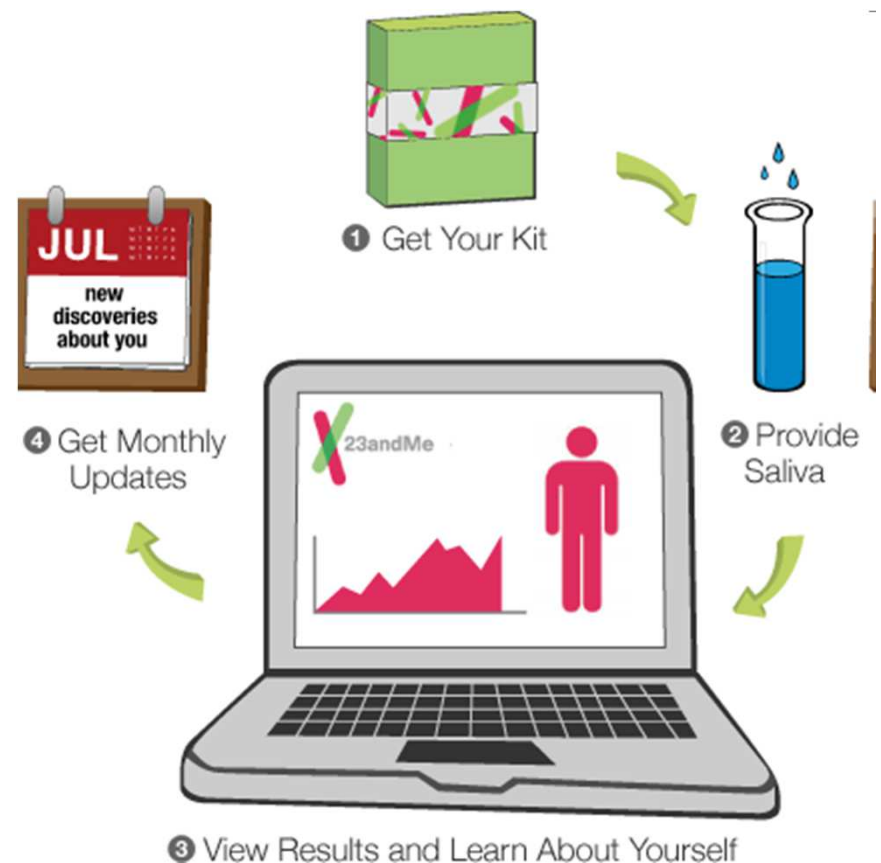
- Léčba chorob
  - **Viry, bakterie**
    - Důležité je pochopit funkci dané viru nebo bakterie
    - Vývoj a výroba antibiotik a antivirotik
  - **Mendelovské choroby**
    - Prozatím neumíme moc léčit
    - Důležité je však i diagnostika (např. v prenatálním stádiu)
    - V současnosti známe cca 2850 genů způsobujících tyto choroby, dalších min. 1800 čeká na objasnění
  - **Polygenetický choroby**
    - Velmi obtížně se detekují – pomáhá výzkum genomová variace
    - Prozatím známe cca 1100 pozic ovlivňujících 165 chorob, další desítky tisíc bude potřeba objasnit



# Komerční sféra

- **Pesonální genomika**

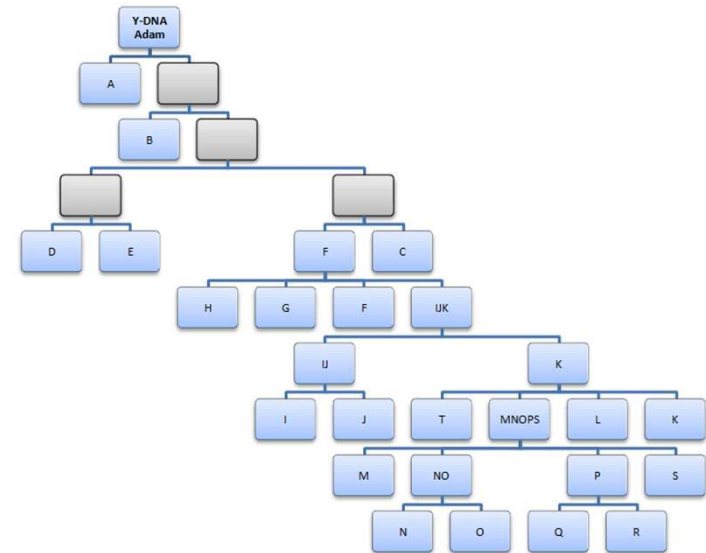
- Pravidlené zasílání informací, náchylnosti k chorobám, genetickém původu apod.
- **23andMe**
  - Spoluzakladatelem je Anne Wojcicki manželka jednoho z majitelů Google
  - <https://www.23andme.com>
- **Navigenics**
  - <http://www.navigenics.com>
- Google má tendenci využít výsledky z obou firem a integrovat je do svého nástroje **Google Health**





# Komerční sféra

- Informací o genetickém původu
  - **Mateřská linie:** analýzou mitochondriální DNA
  - **Otcovská linie:** analýzou Y chromosomu
  - **Postup:**
    - Analýza vzorku DNA
    - Identifikace haploskupiny
    - Sestavení osobního rodokmenu
  - **Komerční společnosti:**
    - Genomia: <http://www.genomia.cz>
    - Genomac: <http://www.genomac.cz>



# Proč bioinformatika?

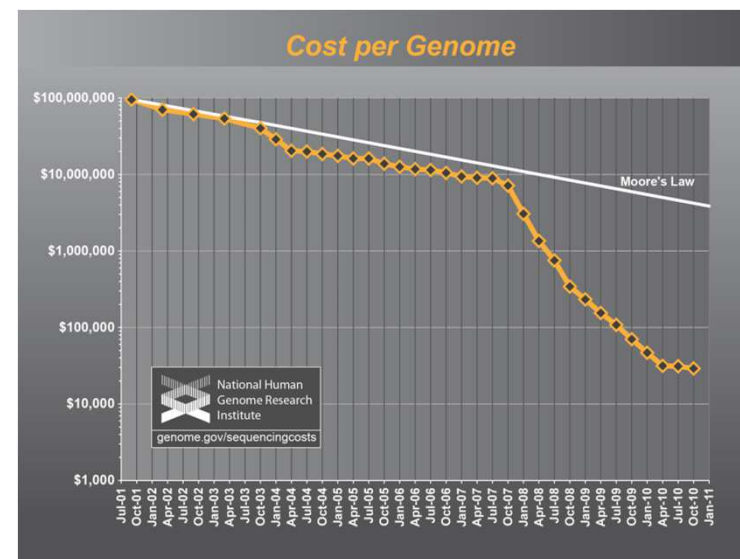
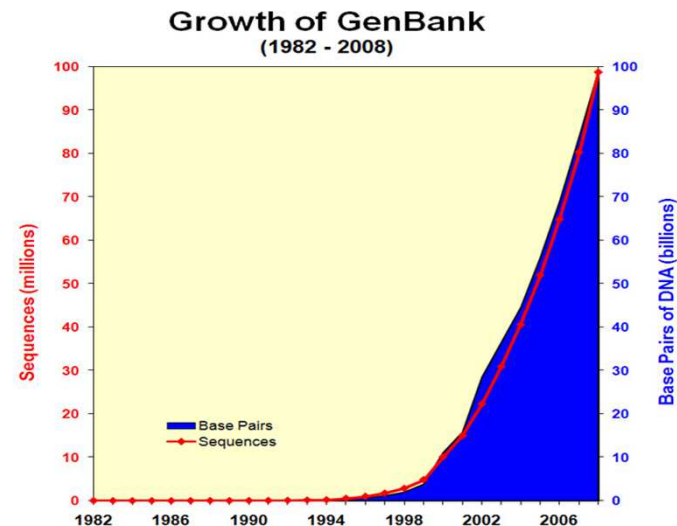
---

- Jakými problémy se zabývá molekulární biologie?
  - Genom, DNA
  - Centrální dogma molekulární biologie
  - Budoucnost biologie
- Jak získat DNA?
  - Technologie sekvenování
  - Významná sekvenovací centra
- Jak interpretovat informace uložené v DNA?
  - Výzkumná oblast
  - Klinická praxe
  - Komerční sféra
- Proč bioinformatika?



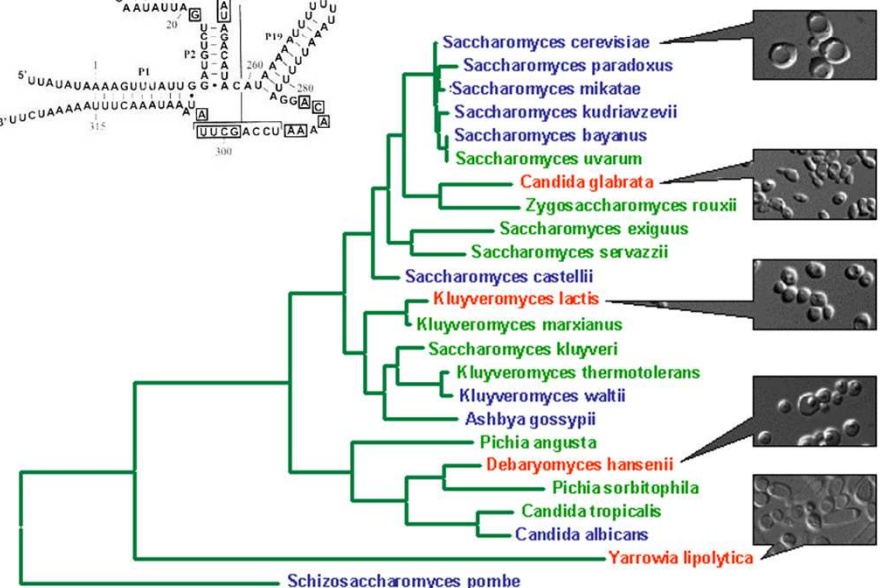
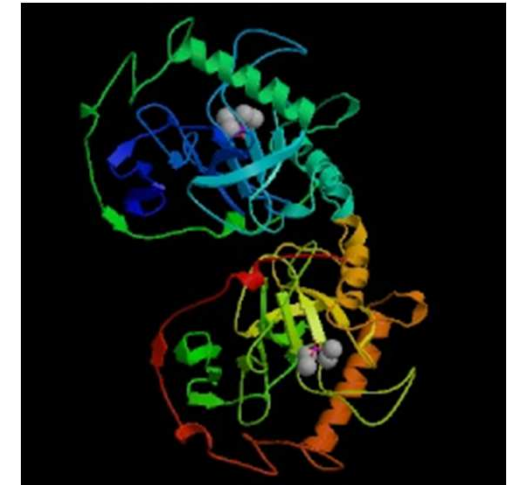
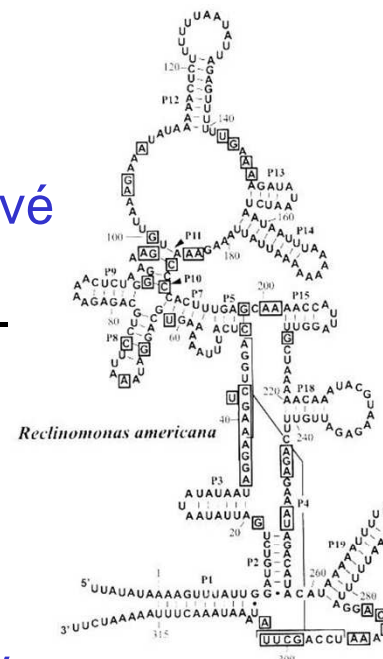
# Proč bioinformatika?

- Množství biologických dat
  - Zdvojnásobuje se každých 6 měsíců
- Charakter biologických dat
  - Obsahují chyby ve formě mutací – zvyšují časovou složitost algoritmů pro jejich analýzu
- Analýza biologických dat
  - Často NP těžké problémy nebo jejich kombinace
  - Pro každé pravidlo existuje výjimka
- Jak se s tímto vyrovnat?
  - Využití znalostí matematických modelů a návrh efektivních algoritmů pro zpracování biologických dat



# Proč bioinformatika?

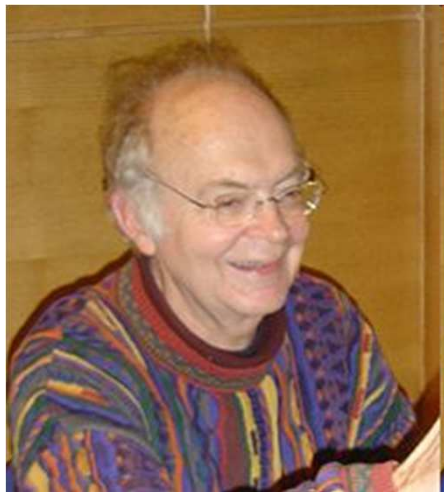
- Příklady aplikací
  - Zarovnání sekvencí - **optimalizační algoritmy**
  - Sestavování fragmentů DNA - **grafové algoritmy**
  - Konstrukce fylogenetických stromů - **metody shlukování**
  - Detekce genů - **skryté Markovovy modely**
  - Predikce struktury RNA - **pravděpodobnostní gramatiky**
  - Predikce struktury proteinů - **metody počítačového učení**
  - Vyhodnocení dat z microarray čipů - **statistické metody**
  - Simulace interakcí mezi proteiny - **molekulová dynamika**
  - Systémová biologie - **Bayesovské sítě**



# Proč bioinformatika?

---

- *„Major discoveries in computer science are unlikely to occur as frequently as they did in the past few decades. On the other hand, Biology easily has 500 years of exciting problems to work on. . .“*



Donald Ervin Knuth, 1993

# Konec

---

Děkuji za pozornost