IV107 Bioinformatika 1

- Dr. Matej Lexa, C505, lexa@fi.muni.cz
- Prednaska: Ut 10:00 - 11:50
- Konzultace: Ct 13:00 – 15:00

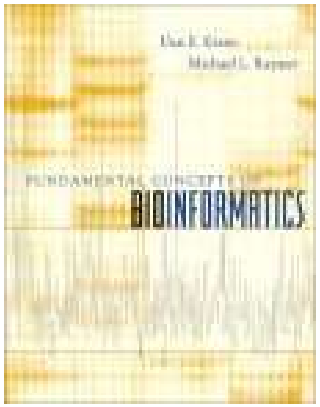
IV107 Bioinformatika 1

- NAVAZUJÍCÍ PŘEDMĚTY
  - IV105 – Seminar z bioinformatiky P (podzim)
  - IV106 – Seminar z bioinformatiky G (ut 13:00)
  - IV108 – Bioinformatika II (podzim)

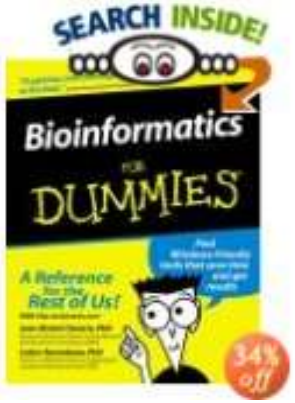
IV107 Dulezite informace

- Prednasky: **12x** (5.4. nebude)
- Exkurze: **1x** **(10.5.?)**
- Kviz: **28.3.**
- **Zkouska:** **24.5.**

IV107 Studijni materialy

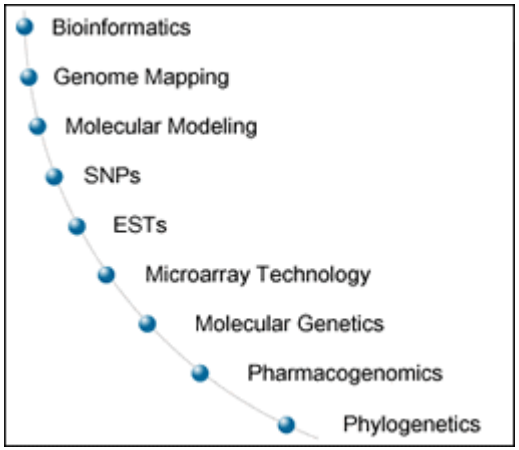


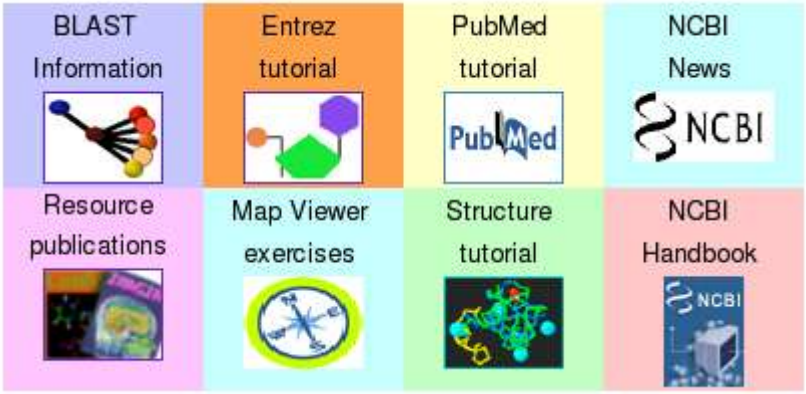
D.E.Krane and M.L.Raymer (2003).
Fundamental Concepts of Bioinformatics.
Benjamin Cummings, London, 320 s.
ISBN 0-8053-4633-3



J.-M.Claverie. (2003).
Bioinformatics for dummies.
Hoboken, Wiley Publishing, 452 s.
ISBN: 0-7645-1696-5

NCBI http://www.ncbi.nlm.nih.gov/Education/index.html

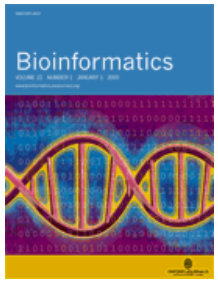




http://www.fi.muni.cz/~lexa/links.html

Briefings in Bioinformatics

Applied Bioinformatics

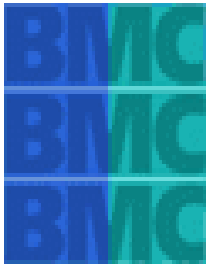


Bioinformatics

Theoretical Biology
and Medical Modelling

Journal of Bioinformatics
and Computational Biology

Genome Biology



BMC Bioinformatics

Science World

BMC Genomics

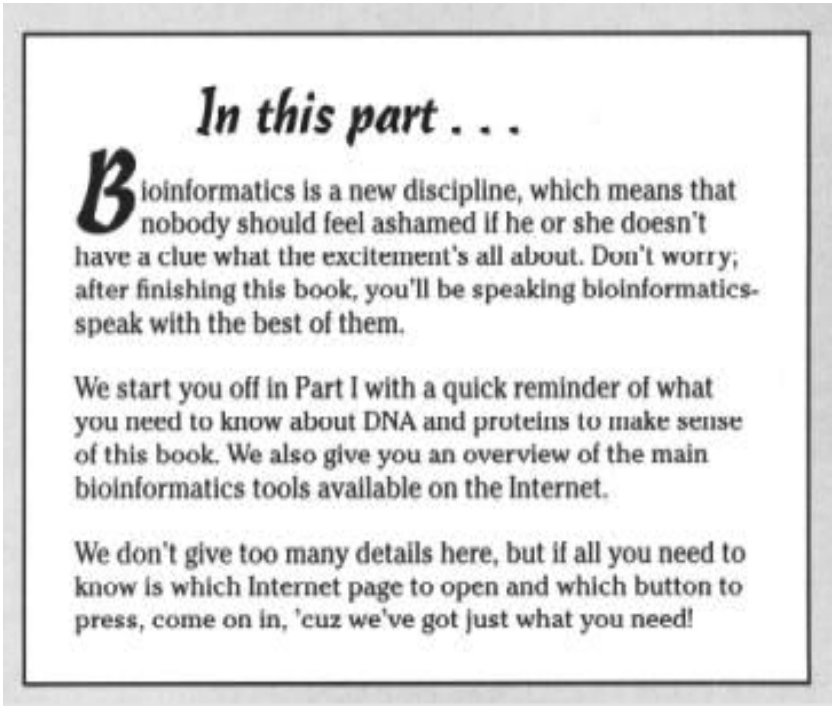
InSilico Biology

- SEMINARE A KONFERENCE V BRNE
  - IV106 Seminar z bioinformatiky G (nove algoritmy pro analyzu geneticke sekvence, ut 13:00 B411)
  - Setkani ceskych bioinformatiku v Telci (31.3-1.4.2006)

IV107 Klasifikace

- kviz: **nad 50%, max. 1x oprava**
- Zkouska:
  - A – 91-100 %
  - B – 81 - 90 %
  - C – 71 - 80 %
  - D – 61 - 70 %
  - E – 41 - 60 %
  - F – 0 - 40 %

In fact, teachers must cope with the fact that biology has its own catch-22: "Everything in biology is understandable as long as you know everything " says Gerald Audesirk. He recalls that he and his



In this part . . .

Bioinformatics is a new discipline, which means that nobody should feel ashamed if he or she doesn't have a clue what the excitement's all about. Don't worry; after finishing this book, you'll be speaking bioinformatics-speak with the best of them.

We start you off in Part I with a quick reminder of what you need to know about DNA and proteins to make sense of this book. We also give you an overview of the main bioinformatics tools available on the Internet.

We don't give too many details here, but if all you need to know is which Internet page to open and which button to press, come on in, 'cuz we've got just what you need!

IV107 Osnova

- Historie a zamereni bioinformatiky
- Zaklady molekularni biologie - Organizace zive hmoty - Struktura a funkce DNA - Struktura a funkce proteinu - Evoluce na urovni genu a proteinu
- Data v bioinformatice - Generovani dat - Bezne formaty dat
- Verejna sekvencni data a pristup k nim
- Analyza sekvence DNA
- Analyza sekvenci proteinu
- Strukturni a funkcni data
- Hodnoceni a vyhledavani podobnosti
- Jina data a analyzy
- Prace s expresnimi daty
- Stepeni proteinu a hmotnostni spektra
- Analyza dat v literature

Bioinformatika

metody pro shromazdovani a analyzu rozsahlych souboru biologickych dat

Vypocetni nebo matematicka biologie

matematicke pristupy k reprezentaci a zkoumani biologickych procesu, casto simulace

Lekarska informatika

prace s medicinskymi daty, prevazne zaznamy pacientu


Modern Life...
SEE? ISN'T IT GREAT TO GET AWAY FROM THE VIDEO MONITOR AND SEE THE REAL WORLD?
COOL GRAPHICS, DAD!
SO WHERE'S THE JOYSTICK? I WANNA BLOW STUFF UP!!
HORSEY

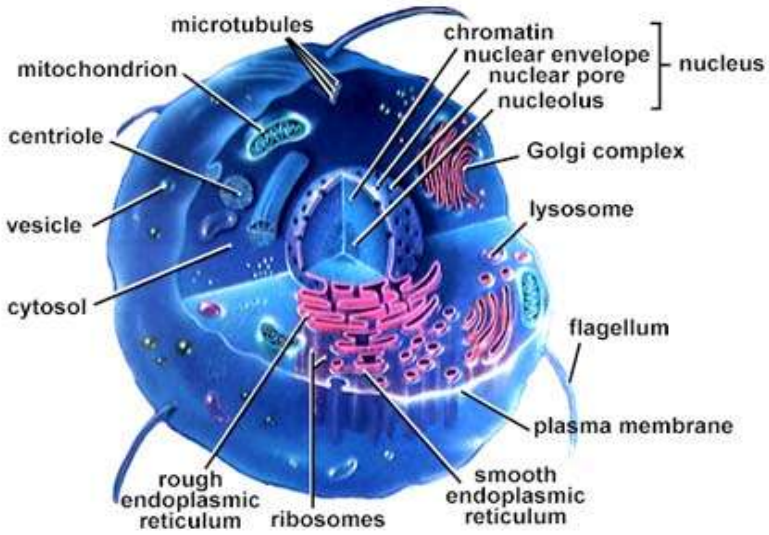
Bioinformaticka data

- Clovek se sklada z asi 1.00E14 bunek. Kazda obsahuje 3.00E09 vesmes stejnych bazi DNA, ktere obsahuji kolem 30000 genu. Kazda bunka aktivuje urcitou podmnozinu teto sady.
- Vysledkem je obrovske mnozstvi moznych stavu bunek, asi tak 2**30000 za predpokladu, ze geny muzou byt jenom aktivovany nebo deaktivovany.
- Samotne geny u jednotlivych organizmu jsou vybrane sady ze zhruba 4**1000 moznych sekvenci

Bunky

- Zakladni forma organizace zive hmoty

- Molekuly/geny/proteiny
- Proteinove komplexy/membrany
- Organely a jine substruktury
- **Bunka**
- Tkan/pletivo
- Organismy



Bioinformaticka data

- Sekvence DNA a RNA
- Sekvence proteinu
- Struktura proteinu
- Udaje o aktivite genu – DNA cip, „microarray“
- Udaje o expresi proteinu – 2-D gely + MS
- Mapy interakci mezi proteiny a DNA
- Mapy interakci mezi proteiny navzajem
- Literatura

Bioinformatik

- Biolog – uzivatel - navrh a interpretace
- Informatik – tvurce

Odhad: 90% rozsireného softwaru bylo vytvoreno biology, kteri se naucili programovat

Vysledek: Pro informatiky, kteri rozumí biologii zustava hodne prace

Co dela bioinformatik?

IN VINO VERITAS		162000
VENI VIDI VICI		132000
IN VIVO	= biolog	19100000
IN VITRO	= biochemik	12900000
IN SILICO	= bioinformatik	349000

Biochemists then recognized that a given type of protein (such as insulin or myoglobin) always contains precisely the same number of total amino acids (generically called *residues*) in the same proportion. Thus, a better formula for a protein looks like:

insulin = (30 glycine + 44 alanine + 5 tyrosine + 14 glutamine + . . .)

Finally, biochemists discovered that these amino acids are linked together as a chain, and that the true identity of a protein isn't only derived from its composition but also from the precise order of its constituent amino acids. The first amino-acid sequence of a protein — insulin — was determined in 1951. The actual recipe for human insulin, from which all its biological properties derive, is the following chain of 110 residues:

insulin = MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSH-
LVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKR-
GIVEQCCTSICSLYQLENYCN

More than 50 years later, analyzing protein sequences like these remains a central topic of bioinformatics in all laboratories throughout the world. Check

Co dela informatik

Because of the centrality of bioinformatics to cutting-edge developments in molecular biology, people from many different fields have been stumbling across the term in a variety of different contexts. If you're a biology, medical, or computer science student, a professional in the pharmaceutical industry, a lawyer or a policeman worrying about DNA testing, a consumer concerned about GMOs (Genetically Modified Organisms), or even a NASDAQ investor interested in start-up companies, you'll already have come across the word *bioinformatics*. If you're good at what you do, you'll want to know what all the fuss is about. This chapter, then, is for you.

Co dela bioinformatik

- Umi pracovat s velkymi datovymi soubory
- Moudrymi triky ovlada vykonne pocítace
- V datech hleda zajimave subsekvence
- Srovnava podobne sekvence
- Predpovida strukturu a funkci genu a proteinu
- Studuje vyvoj sekvencí a organizmu
- Data a vysledky analyz zobrazuje graficky

Co dela bioinformatik

- biologie
- informatika

- analyza sekvenci
- strukturni bioinformatika
- dynamicke modelovani
- analyza obrazu
- lingvistika
- neurologie

Zpusoby nahlizeni na data

KLASICKY

smes biologie, chemie, fyziky atd.

MECHANISTICKY

zive bunky jsou stroje, ktere chceme pochopit a ovladat

EVOLUCE A ZIVOT JAKO HRA

sekvence jsou definicni soubory hracu

GENETICKE INFORMACE JAKO JAZYKY

sekvence se skladaji z frazi a slov s urcitou funkci

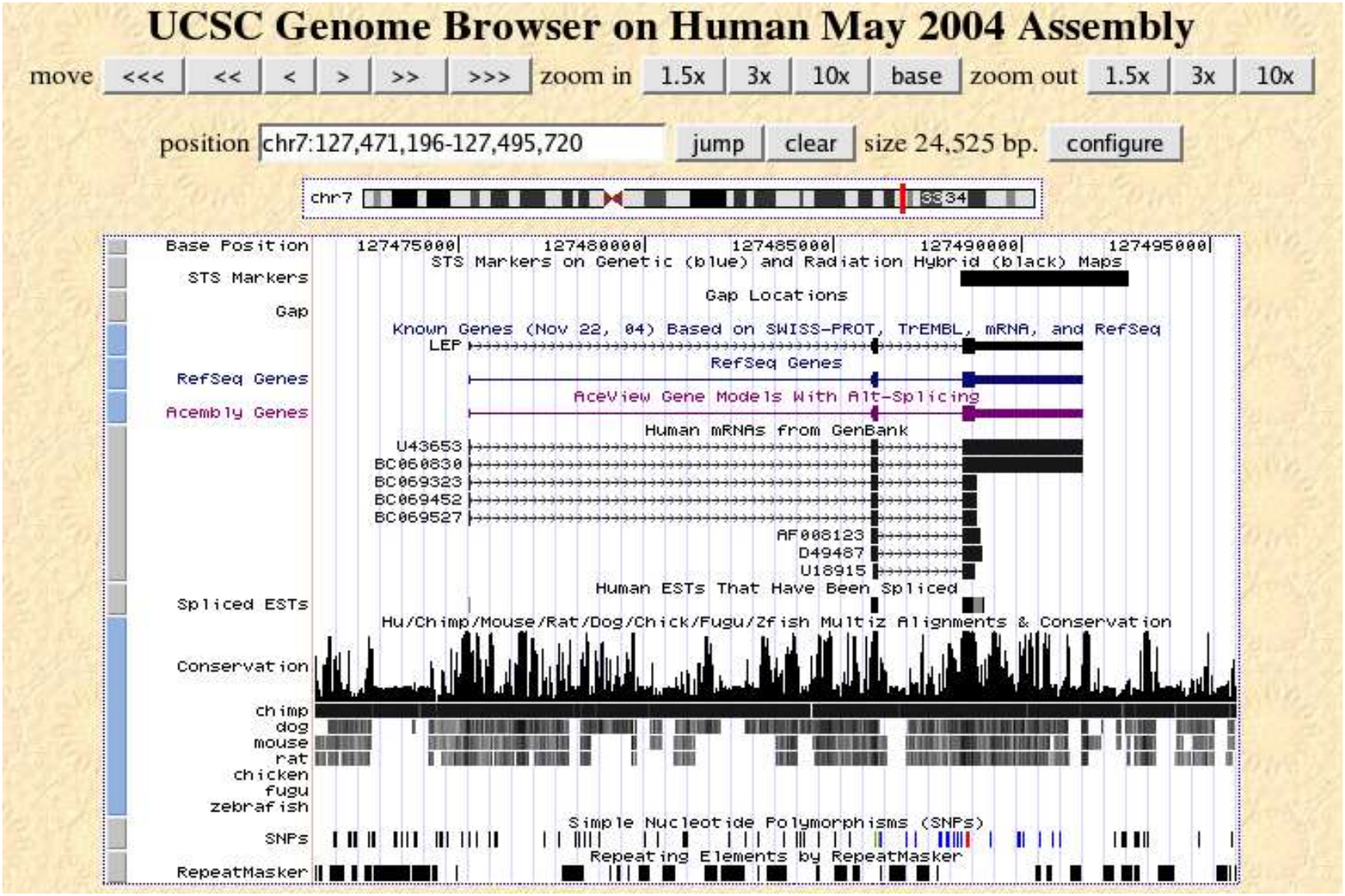
Jim Kent

- autor Aegis Animator, Cyber Paint a Autodesk Animator

- po shlednuti 12 CD-ROM vyvojového prostredi pro Windows 95 presedlava na bioinformatiku s oduvodnenim, ze lidsky genom se vejde na jedno CD

- autor Genome Browser

- sehrava dulezitou roli v honicce o pecteni a skompletovani lidskeho genomu (GigAssembler)


UCSC Genome Browser on Human May 2004 Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position chr7:127,471,196-127,495,720 jump clear size 24,525 bp. configure
chr7
Base Position
127475000
127480000
127485000
127490000
127495000
STS Markers
STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
Gap
Gap Locations
Known Genes (Nov 22, 04) Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
LEP
RefSeq Genes
RefSeq Genes
Acembly Genes
AceView Gene Models With Alt-Splicing
Human mRNAs from GenBank
U43653
BC060830
BC069323
BC069452
BC069527
AF008123
D49487
U18915
Spliced ESTs
Human ESTs That Have Been Spliced
Hu/Chimp/Mouse/Rat/Dog/Chick/Fugu/Zfish Multiz Alignments & Conservation
Conservation
chimp
dog
mouse
rat
chicken
fugu
zebrafish
SNPs
Simple Nucleotide Polymorphisms (SNPs)
RepeatMasker
Repeating Elements by RepeatMasker

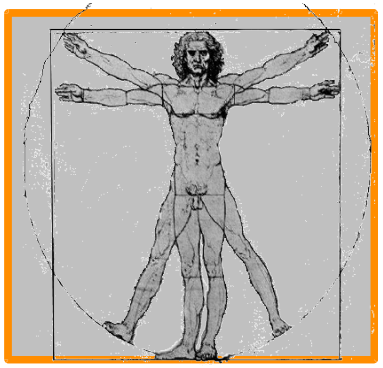
Human vs. Human

- A variation every 1000 nucleotides.
- 90% of human variation is within African populations.
- There are enough humans, and the mutation rate is high enough, that on average each base is mutated several times in each generation.
- Humans each carry hundreds of bad mutations. Most are recessive, only show up with inbreeding.

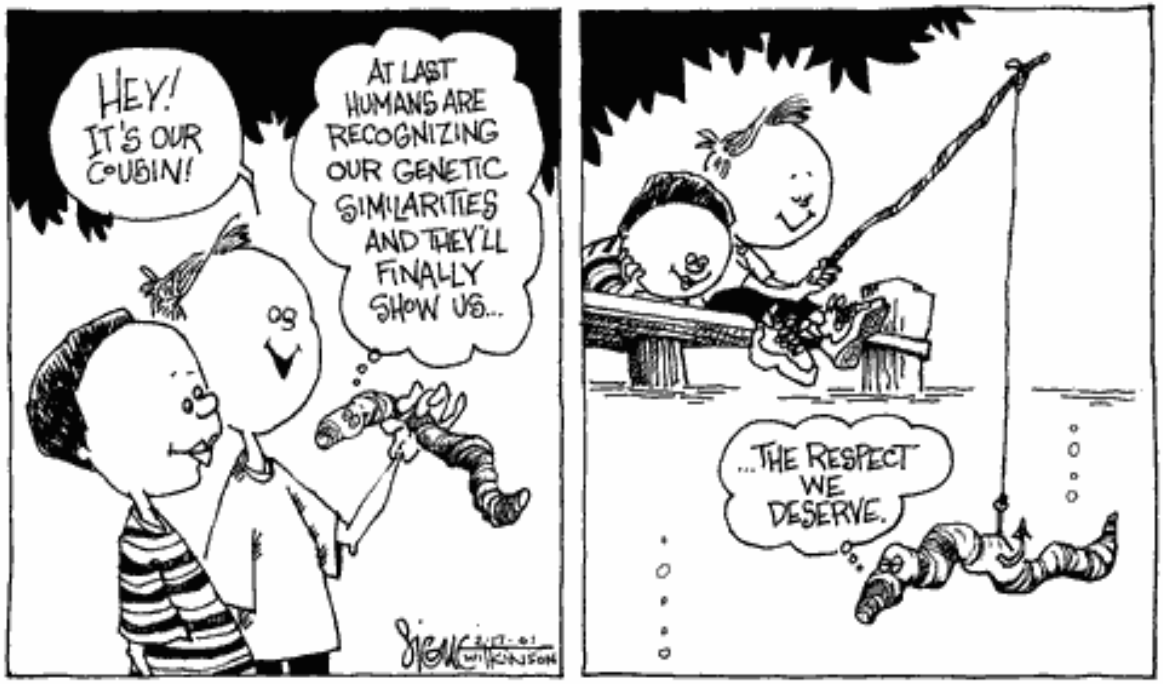
Human vs. Chimpanzee

- A difference every 100 bases.
- A new transposon every 50000 bases
- Two chromosome in one species fused compared to the other.

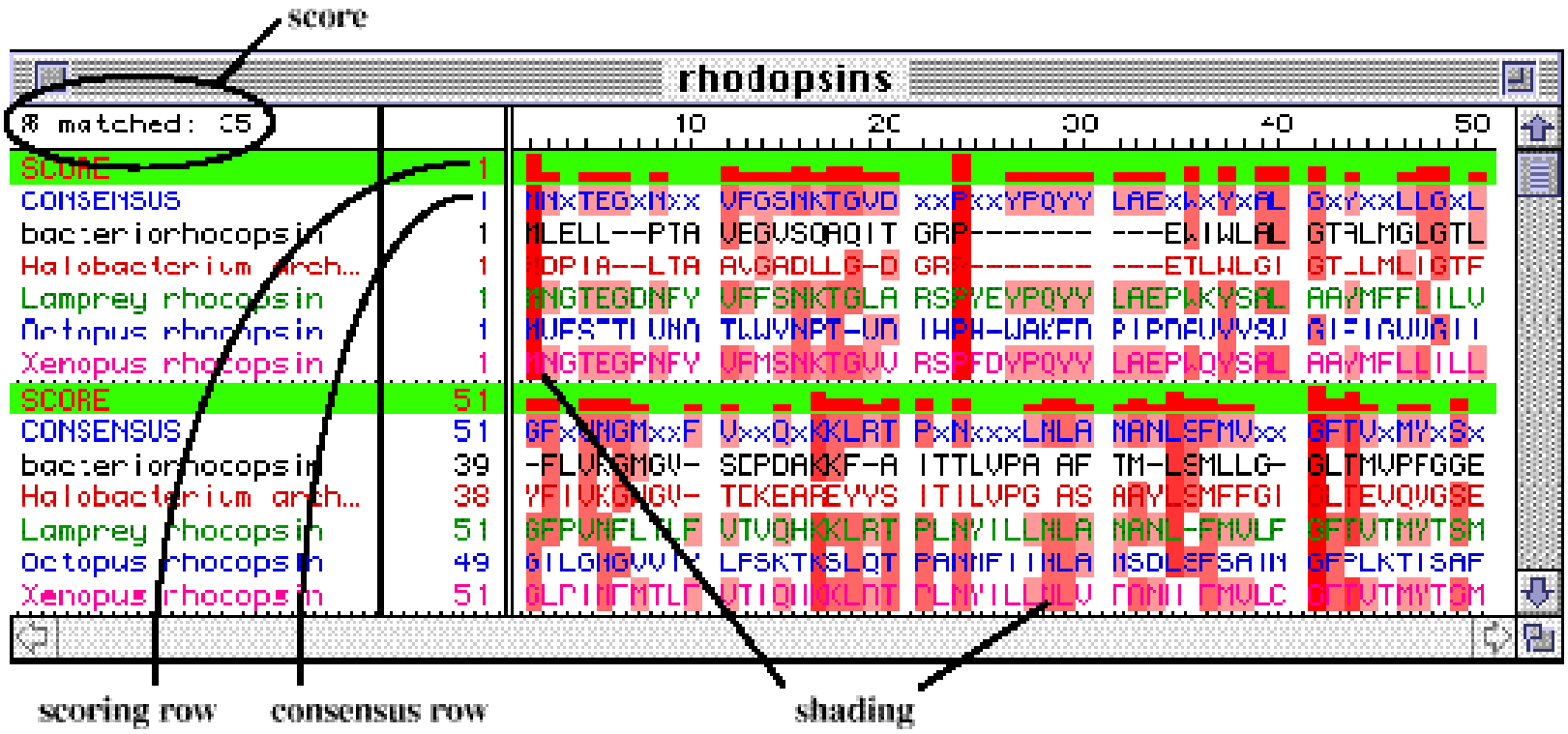
Human vs. Mouse



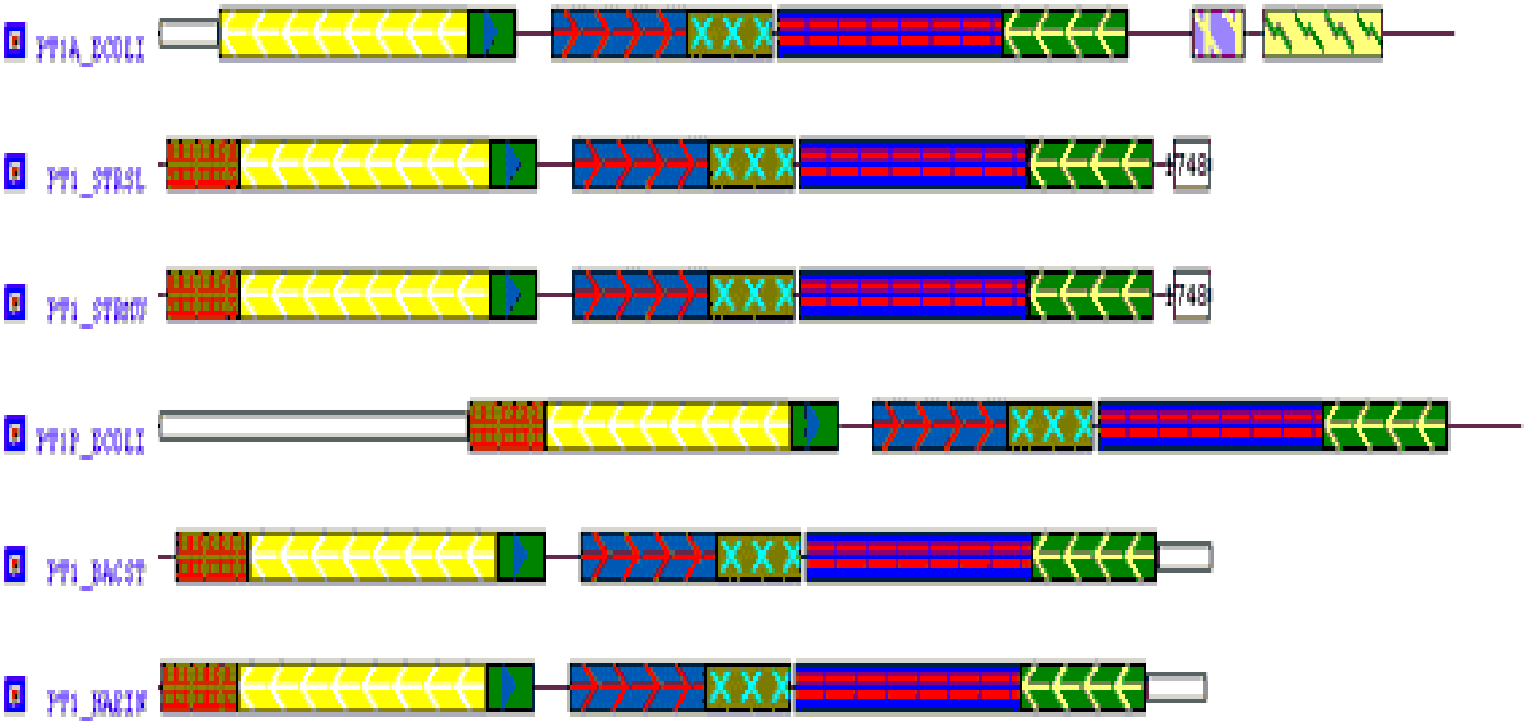
- In general 40% of bases have changed.
- In functional regions only 15% of bases have changed.
- Looking for conserved regions between human and mouse helps identify functional parts of human genome.


HEY! IT'S OUR COUSIN!
AT LAST HUMANS ARE RECOGNIZING OUR GENETIC SIMILARITIES AND THEY'LL FINALLY SHOW US...
...THE RESPECT WE DESERVE.

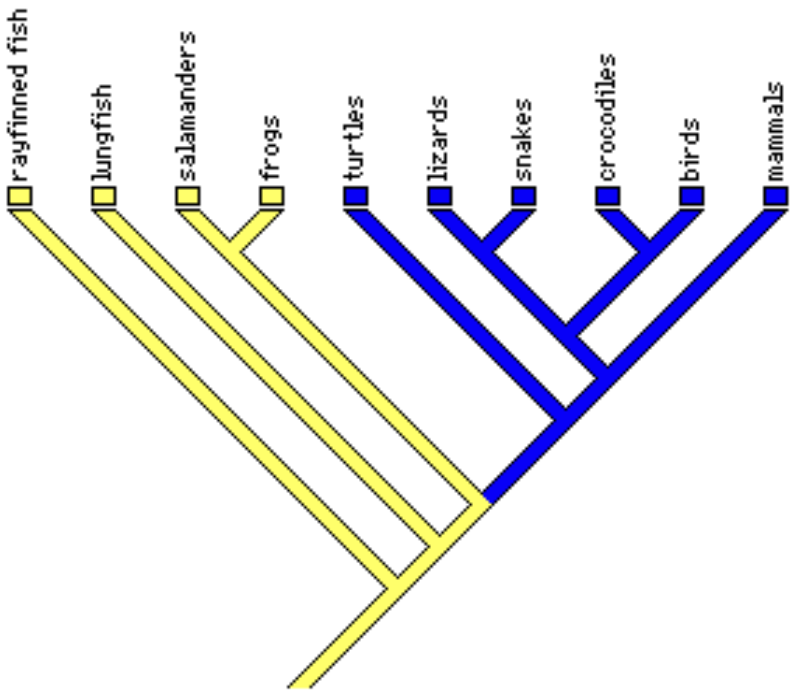
Co dela bioinformatik



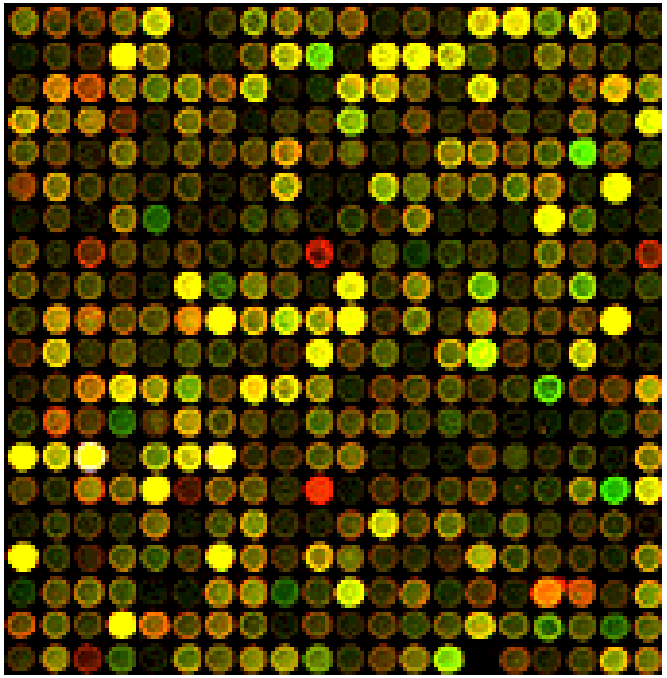
Co dela bioinformatik



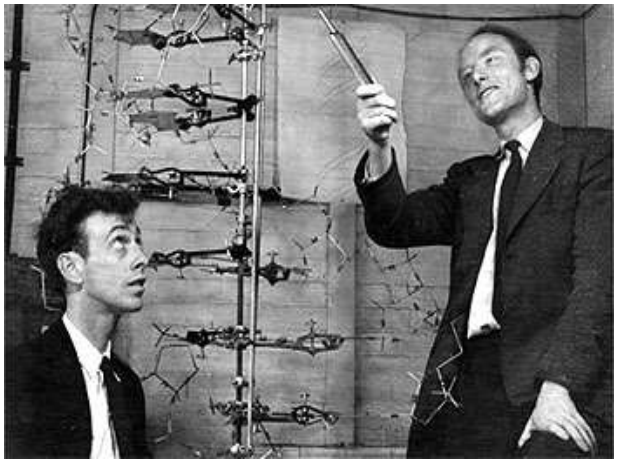
Co dela bioinformatik



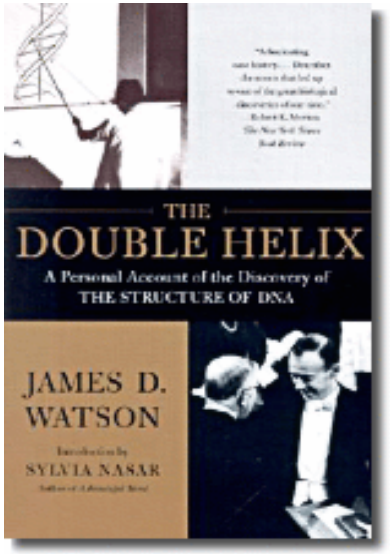
Co dela bioinformatik?



1953 – Watson, Crick, Franklin



We wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.



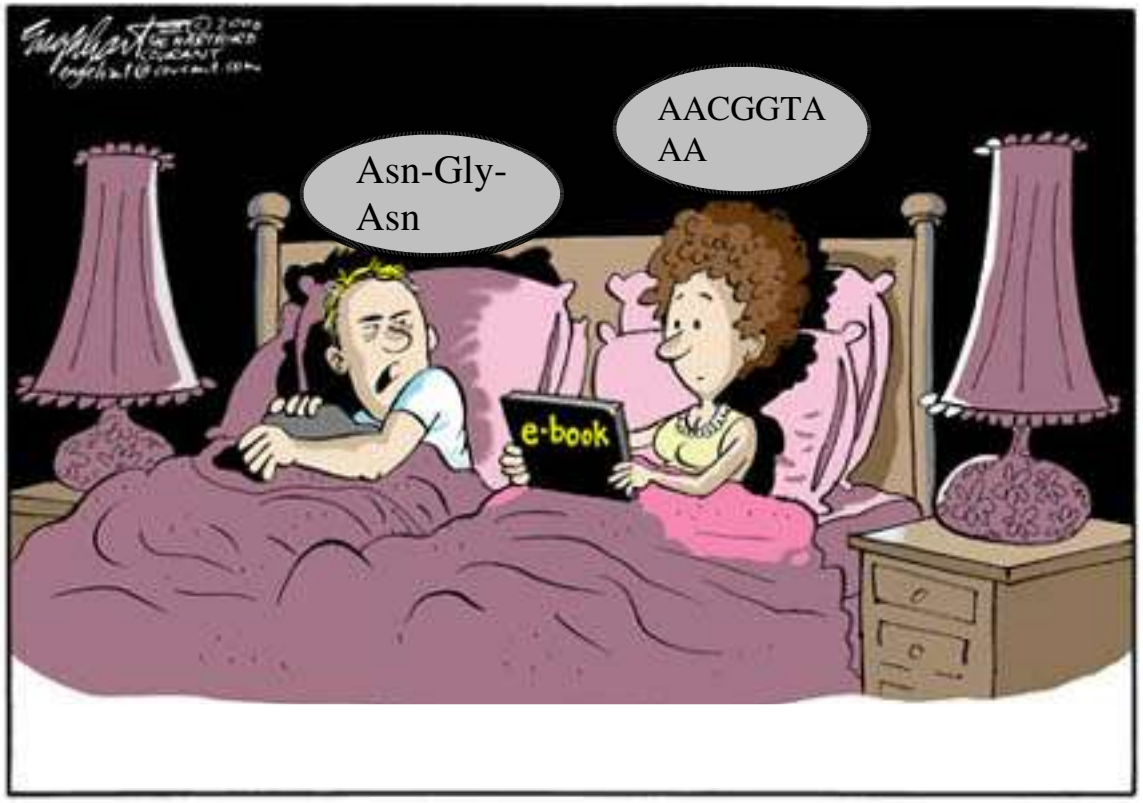
1951 – Pauling	struktura proteinu
1952 – Turing	chemicke zaklady vyvoje
1953 – Watson and Crick	struktura DNA
1956 – Gamow et al.	geneticky kod
1969 – Britten and Davidson	genova regulace
1959 – Chomsky	gramatiky
1962 – Shannon and Weaver	informacni teorie
1966 – Martin-Lof	nahodne retezce
1966 – Neumann	automata

Koreny BIOINFORMATIKY sahaji do 60. let

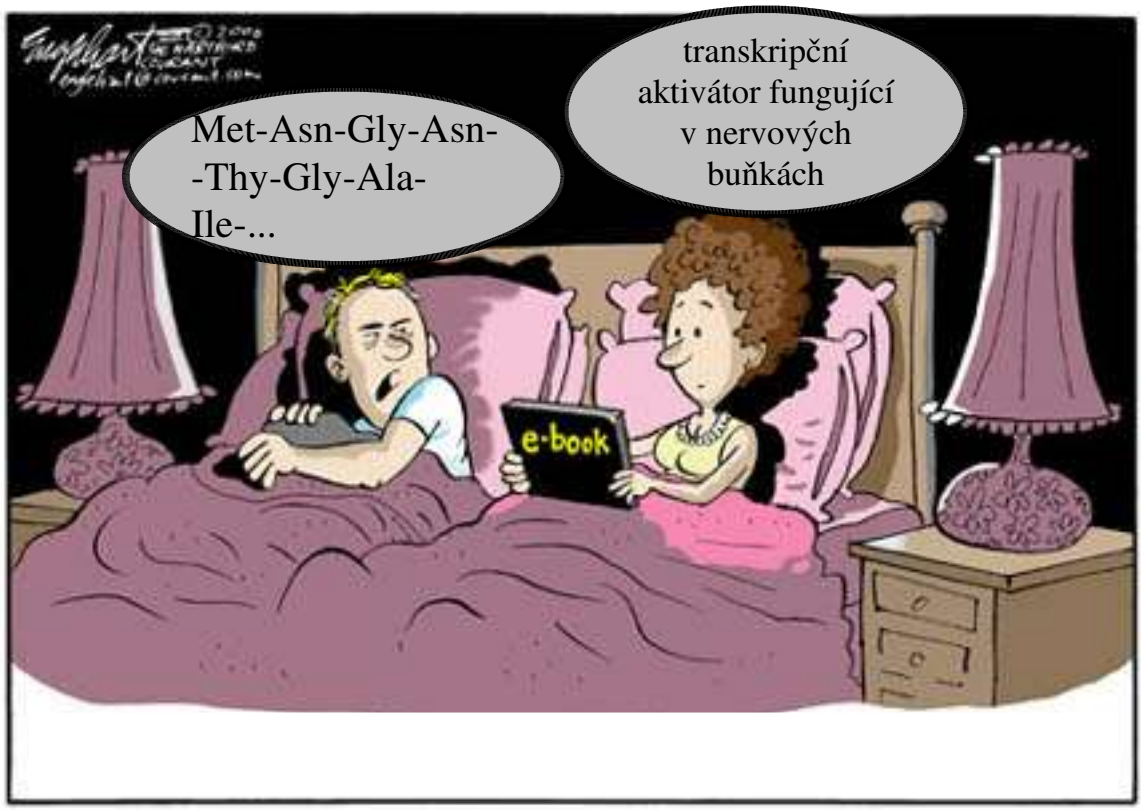
1965 – Zuckerkandl and Pauling	prvni pouziti sekvence v evolucni studii
1967 – Fitch and Margoliash	sestrojeni prvnich fylogenetickych stromu
1970 – Needleman and Wunsh	uziti dyn. programovani k zarovnavani
1974 – Chou and Fasman	predikce sekundarni struktury proteinu
1975 – Tanaka and Sheraga	simulace skladani proteinu
1978 – Dayhoff	prvni sbirka sekvenci proteinu
1981 – Smith and Waterman	modifikace algoritmu pro zarovnavani
1984 – Kabsch and Sander	modelovani struktury proteinu
1986 – Bilofsky et al.	GenBank
1986 – Hamm and Cameron	EMBL Data Library
1987 – Feng and Doolitle	mnohonasobne zarovnani sekvenci
1987 – Gribskov	analyza sekvencnich profilu
1990 – Altschul et al.	efektivni hledani lokalnich podobnosti

1998 – The journal Comp Appl Biosci becomes Bioinformatics

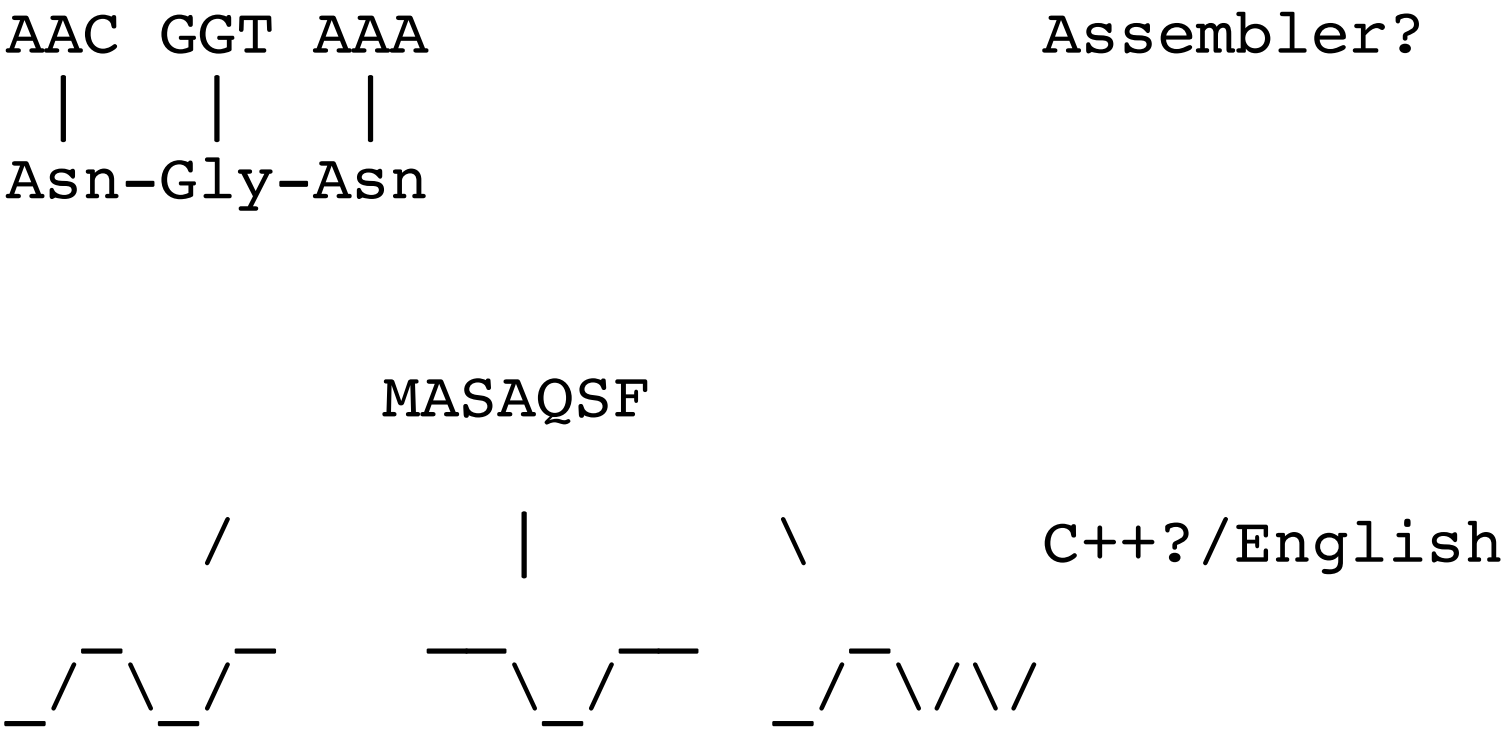
CENTRALNI DOGMA DNA – RNA – PROTEIN



CENTRALNI DOGMA 2? PROTEIN/GEN – STRUKTURA - FUNKCE



Aktualni problemy



Aktualni problemy

BIOLOGICKE SEKVENCE JAKO JAZYK

PROTEIN/GEN	STRUKTURA	FUNKCE
VETA	SYNTAX	VYZNAM

Aktualni problemy

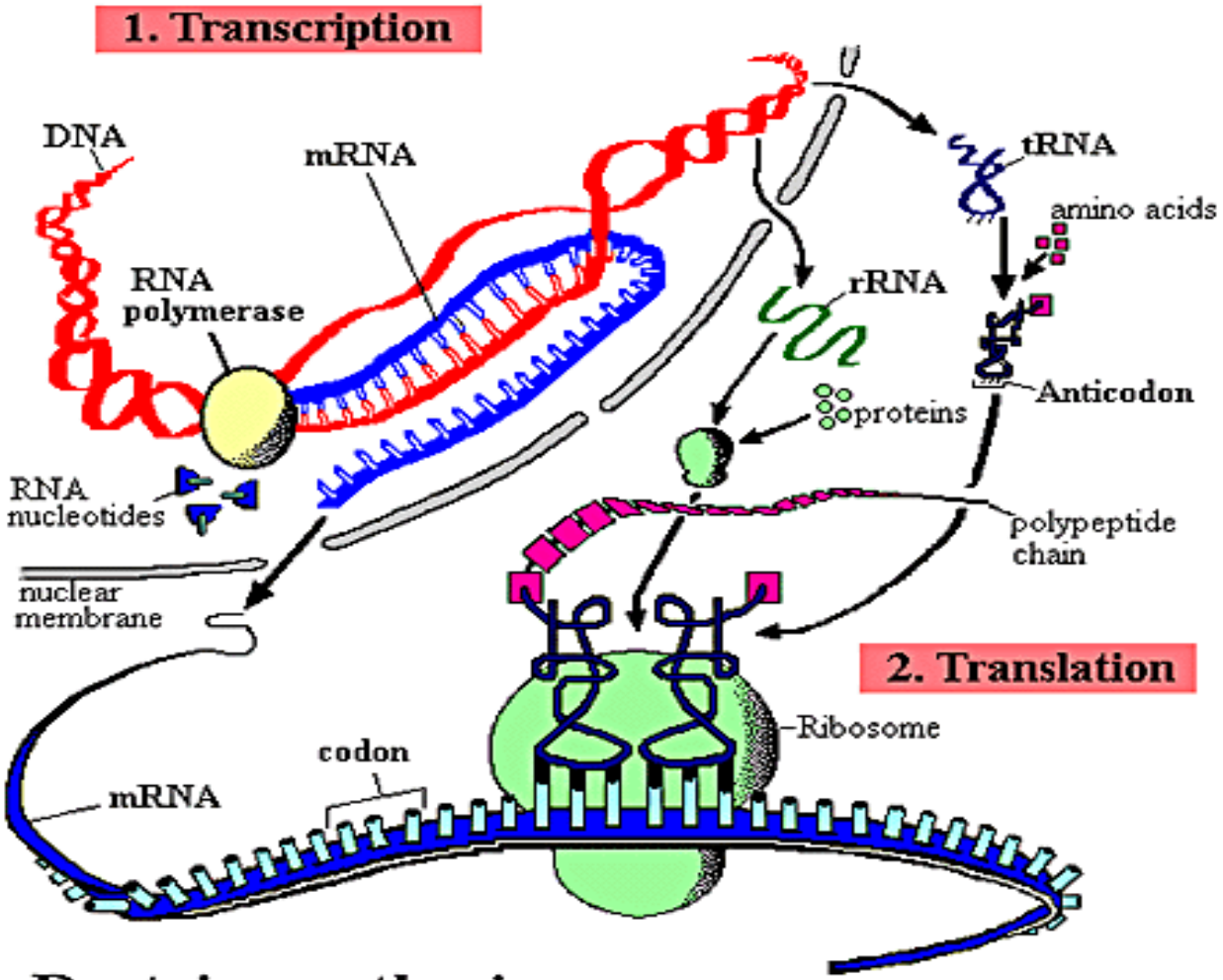
```
Mam z toho velkou radost.
Mam toho kocoura dost.

Mamztohovelk__ouradost.
::: ::::   :  ::::::::
Mam_toho___kocouradost.
```

Aktualni problemy

- Clovek se sklada z asi 1.00E14 bunek. Kazda obsahuje 3.00E09 vesmes stejnych bazi DNA, ktere obsahuji kolem 30000 genu. Kazda bunka aktivuje urcitou podmnozinu teto sady.
- Vysledkem je obrovske mnozstvi moznych stavu bunek, asi tak 2**30000 za predpokladu, ze geny muzou byt jenom aktivovany nebo deaktivovany.
- Samotne geny u jednotlivych organizmu jsou vybrane sady ze zhruba 4**1000 moznych sekvenci

Aktualni problemy

01000101001000001111
11010100100101010010
10101010010100100100
10101001010100010010
10101010010101001010
10101010010101010010



Protein synthesis

Centralni dogma

- DNA -> RNA -> PROTEIN