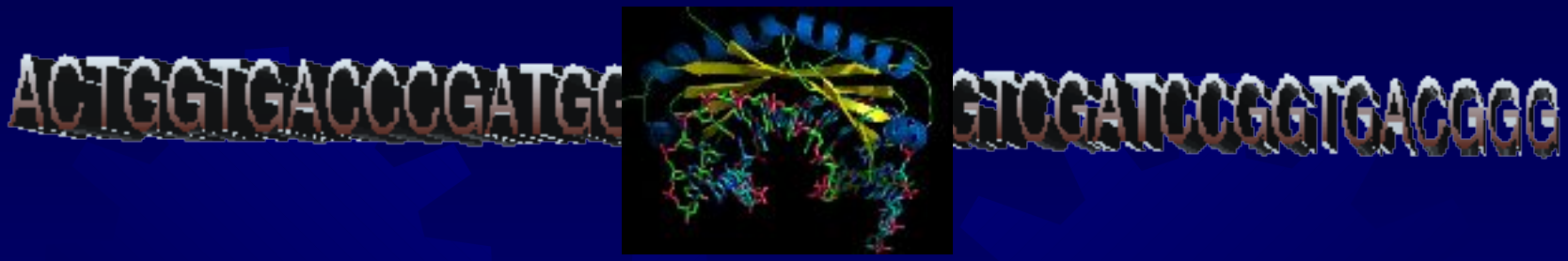
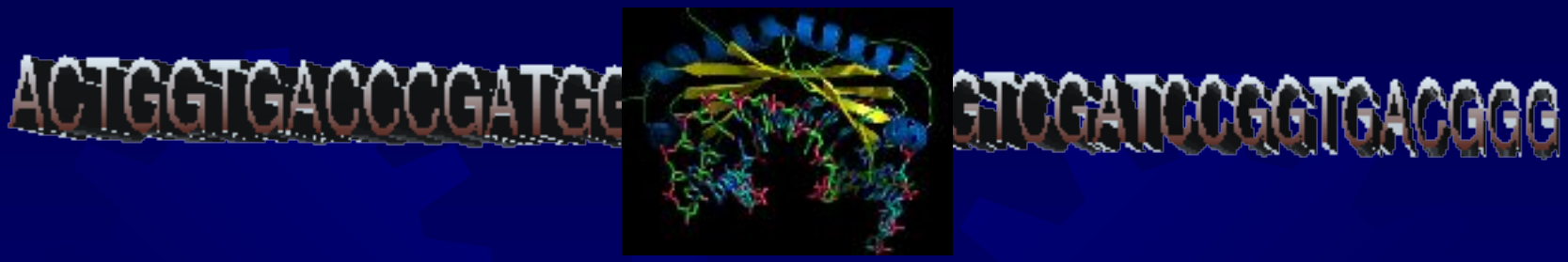


IV107 Bioinformatika 1

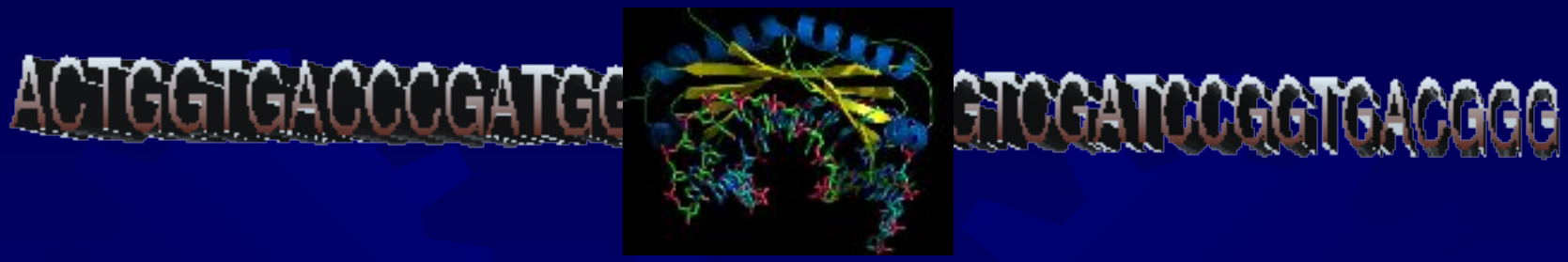
- Dr. Matej Lexa, C505, lexa@fi.muni.cz

- Přednáška: Út 8:00 - 9:50
- Konzultace: Čt 13:00 – 15:00



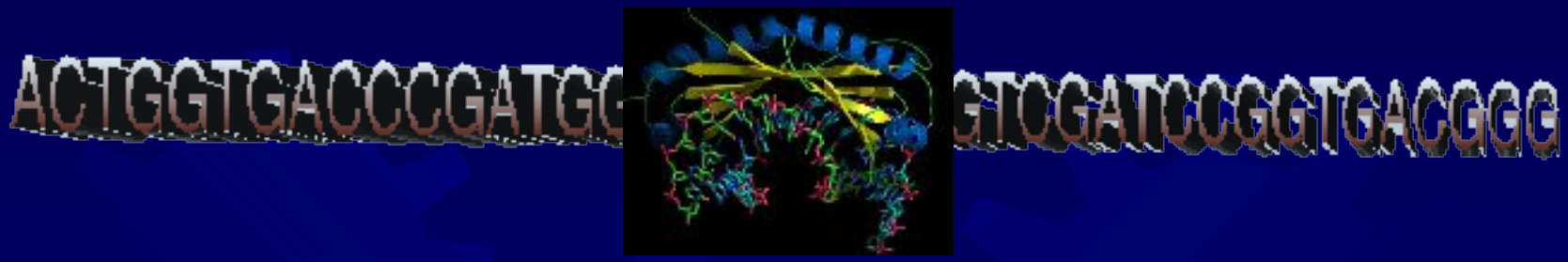
IV107 Bioinformatika 1

- NAVAZUJÍCÍ PŘEDMĚTY
  - IV105 – Seminář z bioinformatiky P (podzim)
  - IV106 – Seminář z bioinformatiky G (út 12:00)
  - IV108 – Bioinformatika II (podzim)
  - IV110 – Projekt z bioinformatiky (podzim)



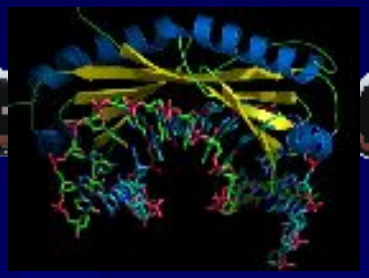
Obor Bioinformatika

- Lze zvolit i v průběhu studia
- Kromě základních předmětů FI
  - Biochemie (LF)
  - Molekulární biologie (PřF)
  - Bioinformatika (FI)
  - Počítačová chemie (FI)

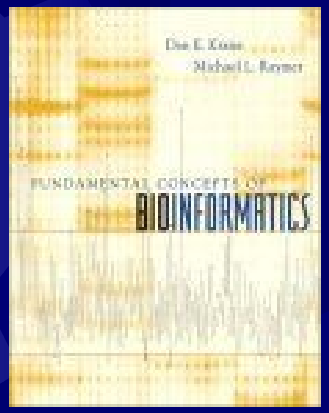


IV107 Důležité informace

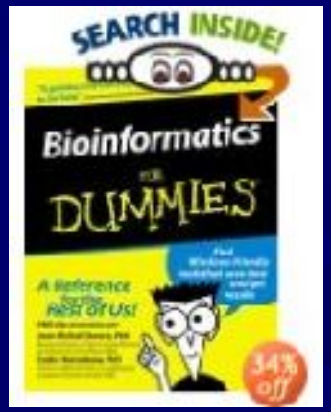
- Přednášky: **13x**
- Kvíz: **27.3.**



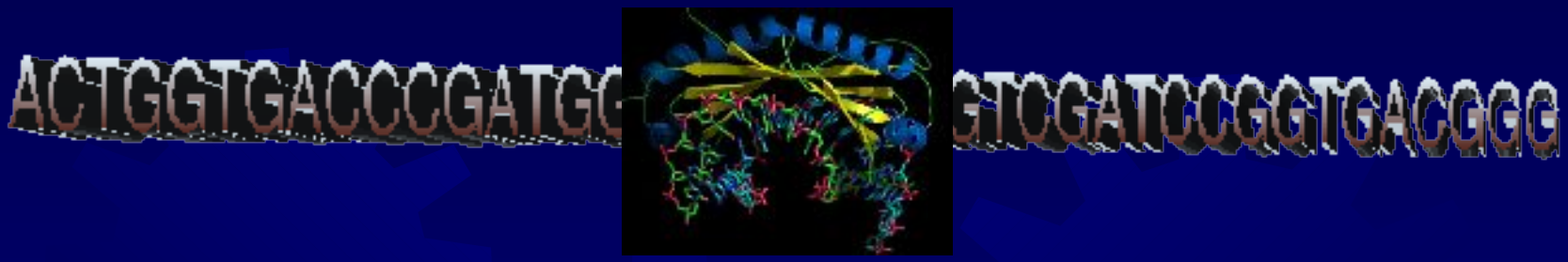
IV107 Studijní materiály



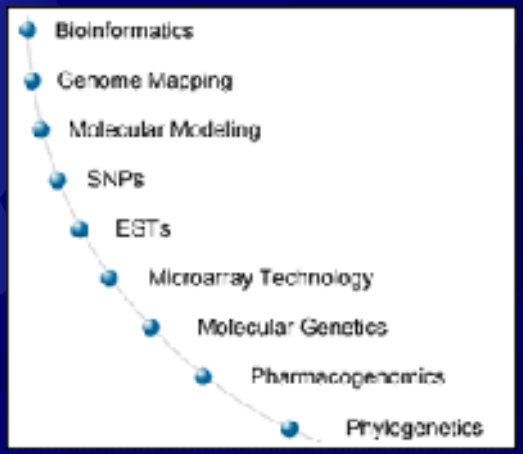
D.E.Krane and M.L.Raymer (2003).
Fundamental Concepts of Bioinformatics.
Benjamin Cummings, London, 320 s.
ISBN 0-8053-4633-3



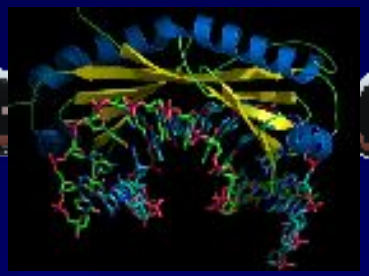
J.-M.Claverie. (2003).
Bioinformatics for dummies.
Hoboken, Wiley Publishing, 452 s.
ISBN: 0-7645-1696-5



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



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



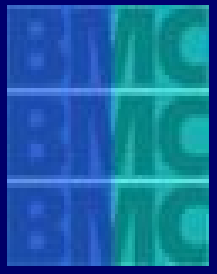
Briefings in Bioinformatics
Bioinformatics

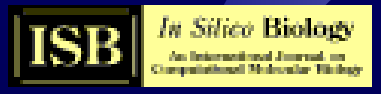
Applied

Bioinformatics
Theoretical Biology

and Medical Modelling

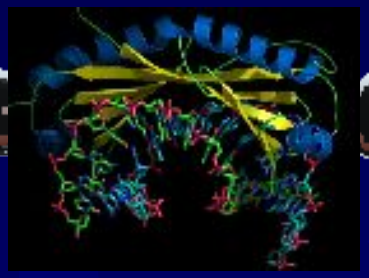
Journal of Bioinformatics
Genome Biology
and Computational Biology





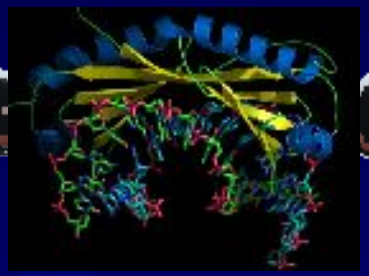
BMC Bioinformatics

Science



IV107 Klasifikace

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

ACTGGTGACCCGATGG  GTCGATCCGGTGACGGG

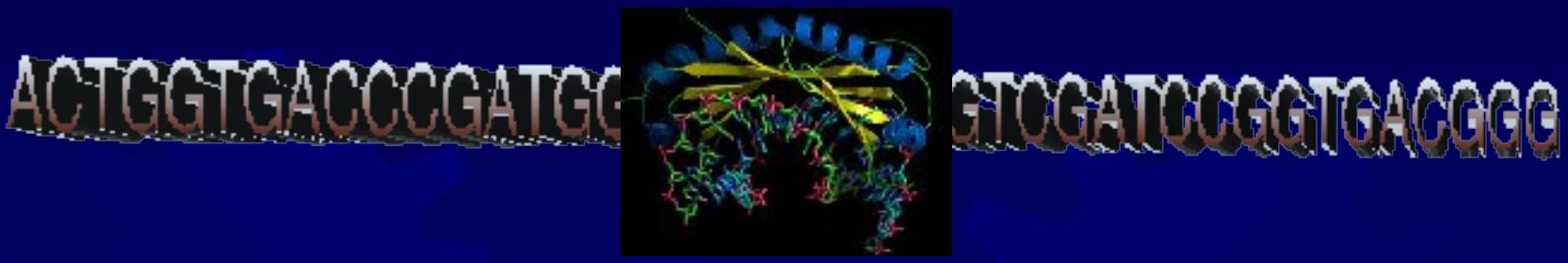
In fact, teachers must cope with the fact that biology has its own catch 22: "Everything in biology is understandable as long as you know everything," says Gerald Audesirk. He recalls that he and his

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.

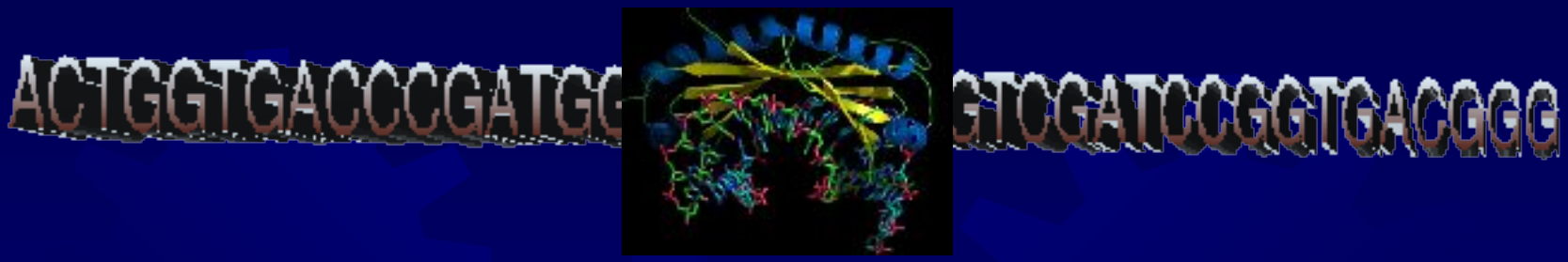
We start you off in Part I with a quick reminder of what you need to know about DNA and proteins to make sense of this book. We also give you an overview of the main bioinformatics tools available on the Internet.

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 zaměření bioinformatiky
- Základy molekulární biologie - Organizace živé hmoty - Struktura a funkce DNA - Struktura a funkce proteinů - Evoluce na úrovni genů a proteinů
- Data v bioinformatice - Generování dat - Běžné formáty dat
- Veřejná sekvenční data a přístup k nim
- Analýza sekvence DNA
- Analýza sekvencí proteinů
- Strukturní a funkční data
- Hodnocení a vyhledávání podobností
- Jiná data a analýzy
- Práce s expresními daty
- Štěpení proteinů a hmotnostní spektra
- Analýza dat v literatuře



Bioinformatika

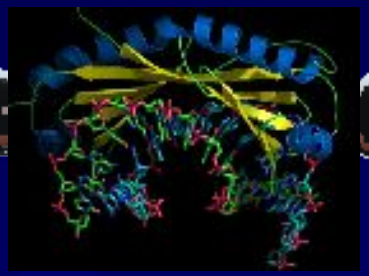
metody pro shromažďování a analýzu rozsáhlých souborů biologických dat

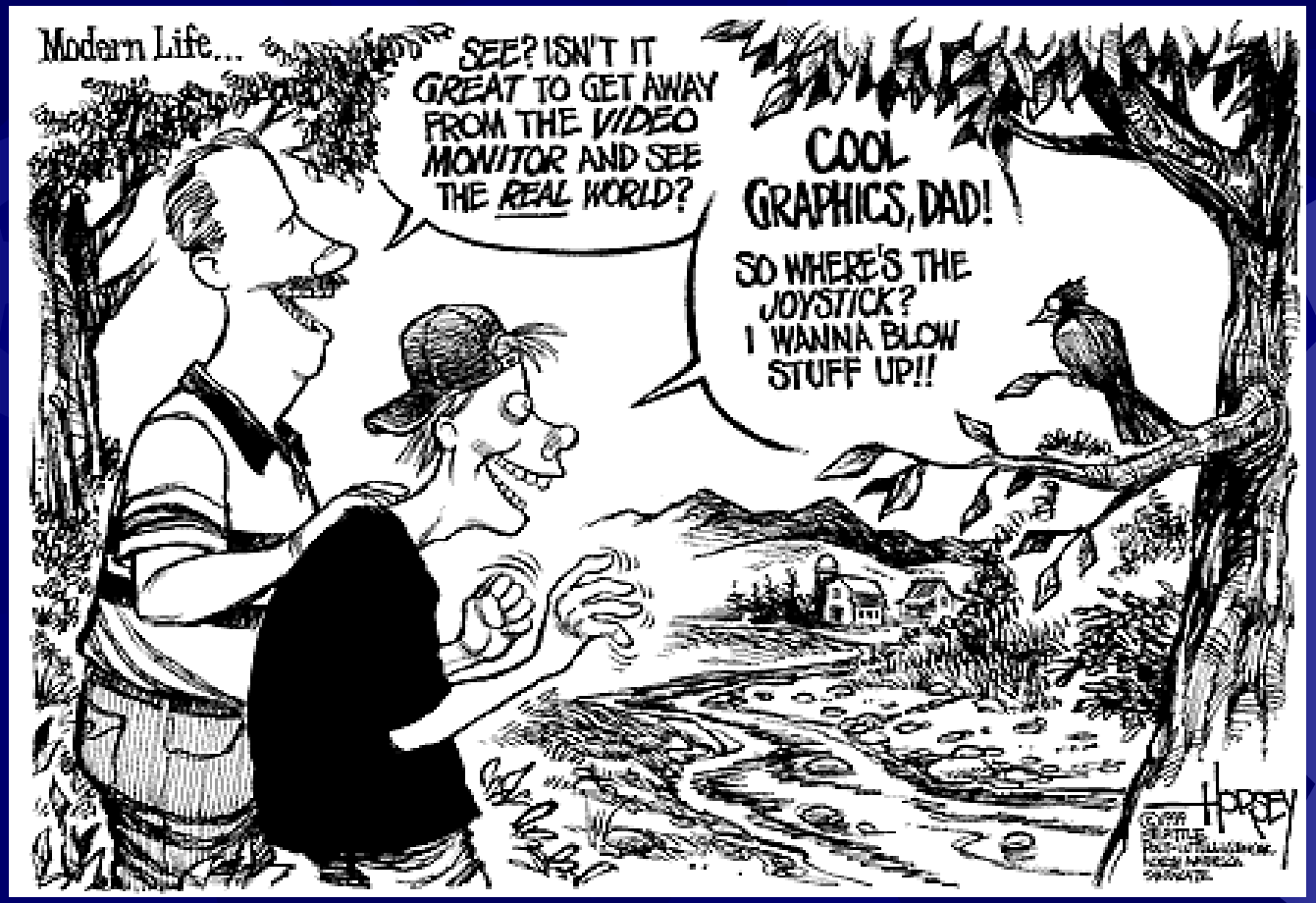
Výpočetní nebo matematická biologie

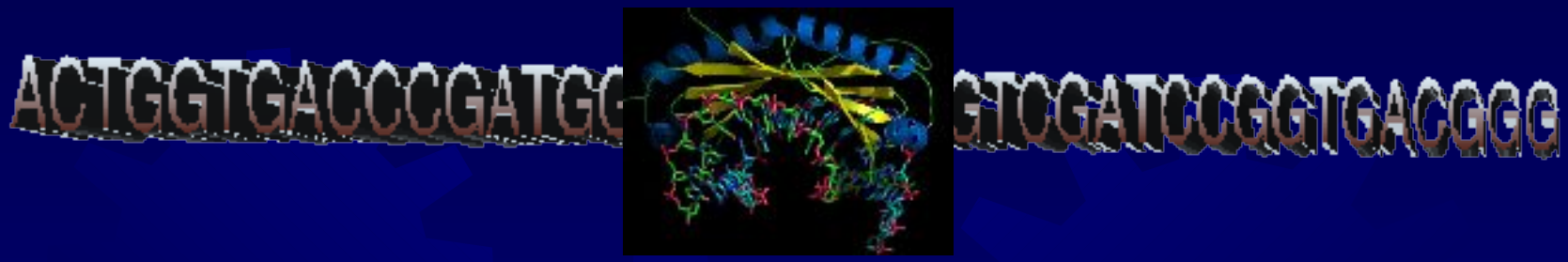
matematické přístupy k reprezentaci a zkoumání biologických procesů, často simulace

Lékařská informatika

práce s medicínskými daty, převážně záznamy pacientů

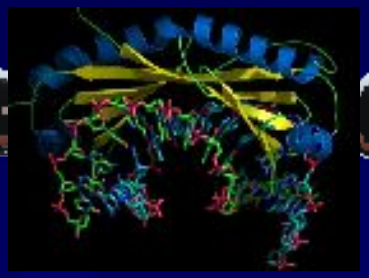
ACTGGTGACCCGATGG  GTCGATCCGGTGACGGG



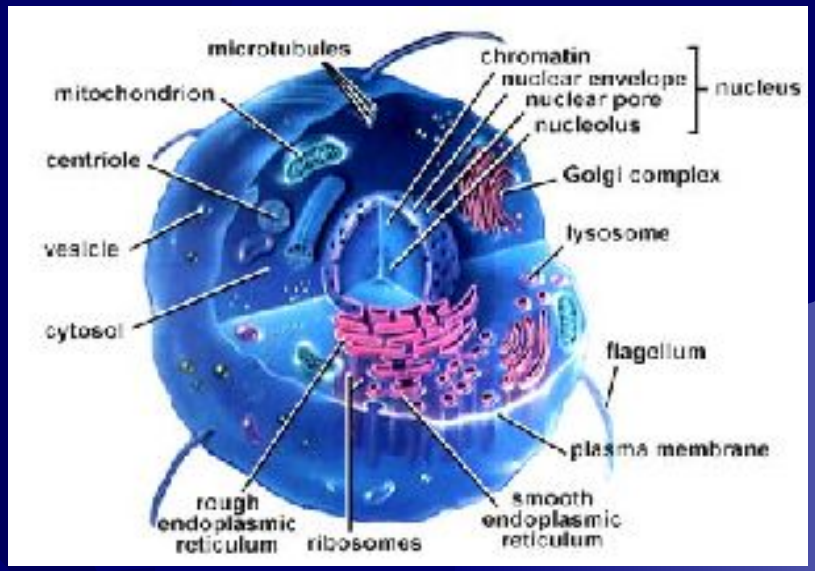


Bioinformatická data

- Člověk se skládá z asi 10**14 buněk. Každá obsahuje 3x10**9 vesměs stejných párů bazí DNA, které vytvářejí svými kombinacemi kolem 30 000 genů. Každá buňka aktivuje v každé chvíli určitou podmnožinu této sady.
- Výsledkem je obrovské množství možných stavů buňek, asi tak 2**30000 jenom za předpokladu, že geny můžou být pouze aktivovány nebo deaktivovány.
- Samotné geny u jednotlivých organizmů jsou vybrané sady ze zhruba 4**1000 možných sekvencí DNA

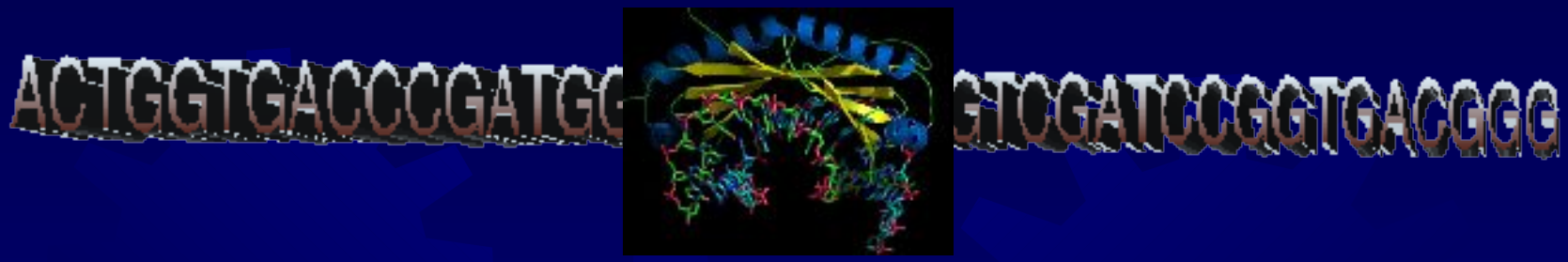


Buňky



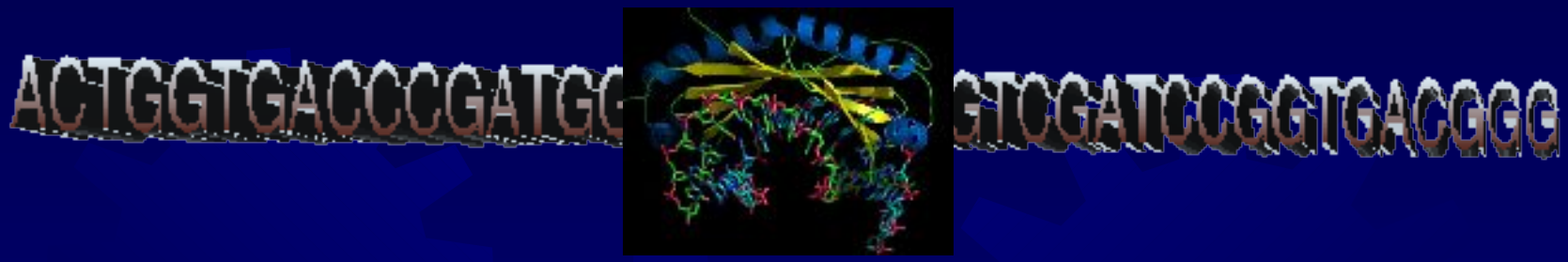
- Základní forma organizace živé hmoty

- Molekuly/geny/proteiny
- Proteinové komplexy/membrány
- Organely a jiné substruktury
- **Buňka**
- Tkáň/pletivo
- Organismy



Bioinformatická data

- Sekvence DNA a RNA
- Sekvence proteinů
- Struktura proteinů
- Údaje o aktivitě genů – DNA čip, „microarray“
- Údaje o expresi proteinů – 2-D gely + MS
- Mapy interakcí mezi proteiny a DNA
- Mapy interakcí mezi proteiny navzájem
- Literatura

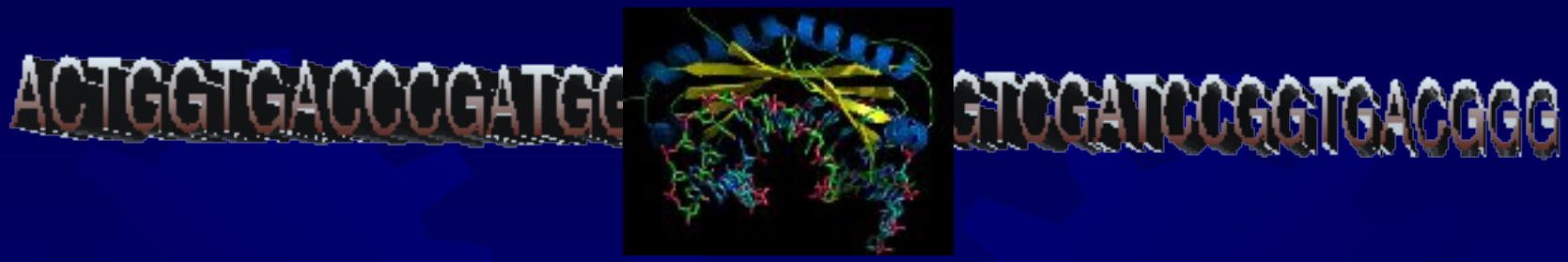


Bioinformatik

- Biolog – uživatel - návrh a interpretace
- Informatik – tvůrce

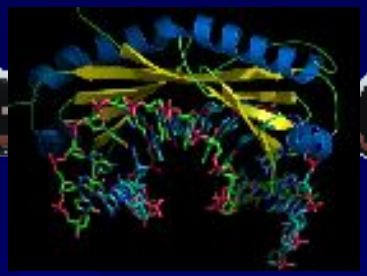
Odhad: 80% rozšířeného softwaru bylo vytvořeno biology, kteří se naučili programovat

Výsledek: Pro informatiky, kteří rozumí biologii zůstáva hodně práce



Co dělá bioinformatik?

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

ACTGGTGACCCGATGG  GTCGATCCGGTGACGGG

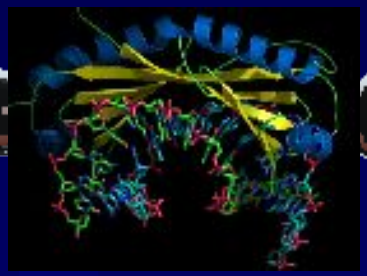
Biochemists then recognized that a given type of protein (such as insulin or myoglobin) always contains precisely the same number of total amino acids (generically called *residues*) in the same proportion. Thus, a better formula for a protein looks like:

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:

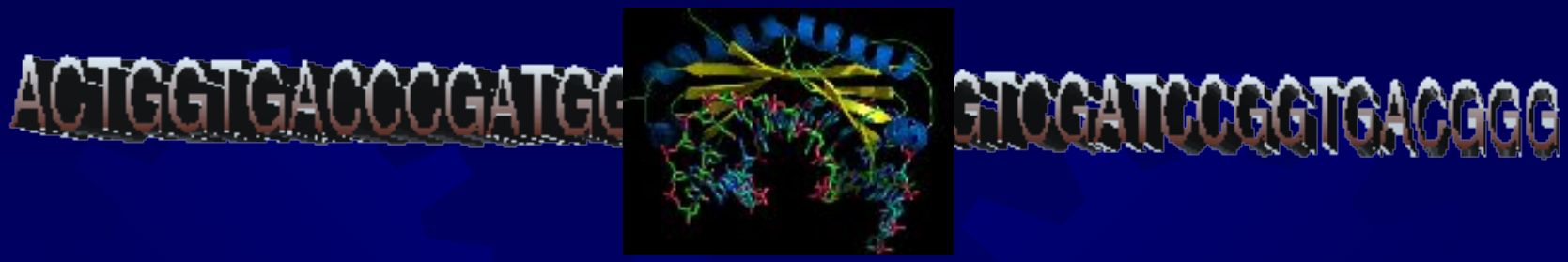
insulin = MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSH-
LVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKR-
GIVEQCCTSICSLYQLENYCN

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



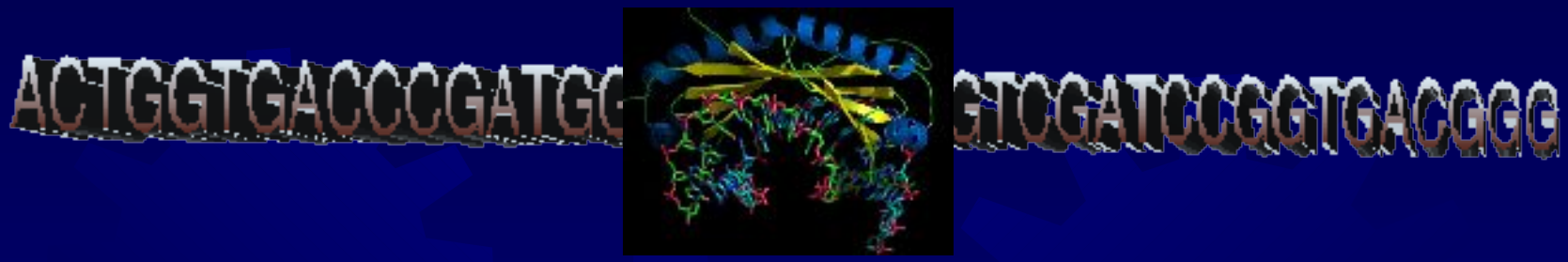
Co dělá bioinformatik

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 dělá bioinformatik

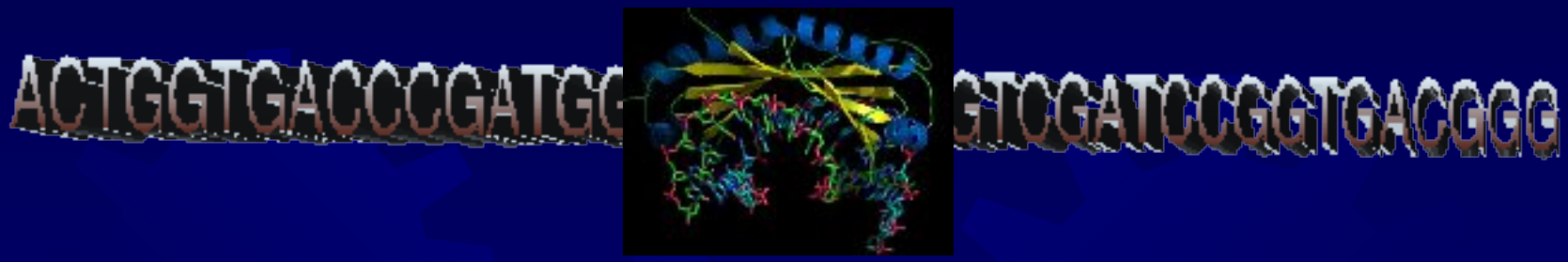
- Umí pracovat s velkými datovými soubory
- Moudrými triky ovláda výkonné počítače
- V datech hledá zajímavé subsekvence
- Srovnává podobné sekvence
- Předpovídá strukturu a funkci genů a proteinů
- Studuje vývoj sekvencí a organizmů
- Data a výsledky analýz zobrazuje graficky



Co dělá bioinformatik

- biologie
- informatika

- analýza sekvencí
- strukturní bioinformatika
- dynamické modelování
- analýza obrazu
- databázy a vyhledávání znalostí
- lingvistika
- neurologie



Způsoby nahlížení na data

KLASICKÝ

směs biologie, chemie, fyziky atd.

MECHANISTICKÝ

živé buňky jsou stroje, které chceme pochopit a ovládat

EVOLUCE A ŽIVOT JAKO HRA

sekvence jsou definiční soubory hráčů

GENETICKÉ INFORMACE JAKO JAZYKY

sekvence se skládají z frází a slov s určitou

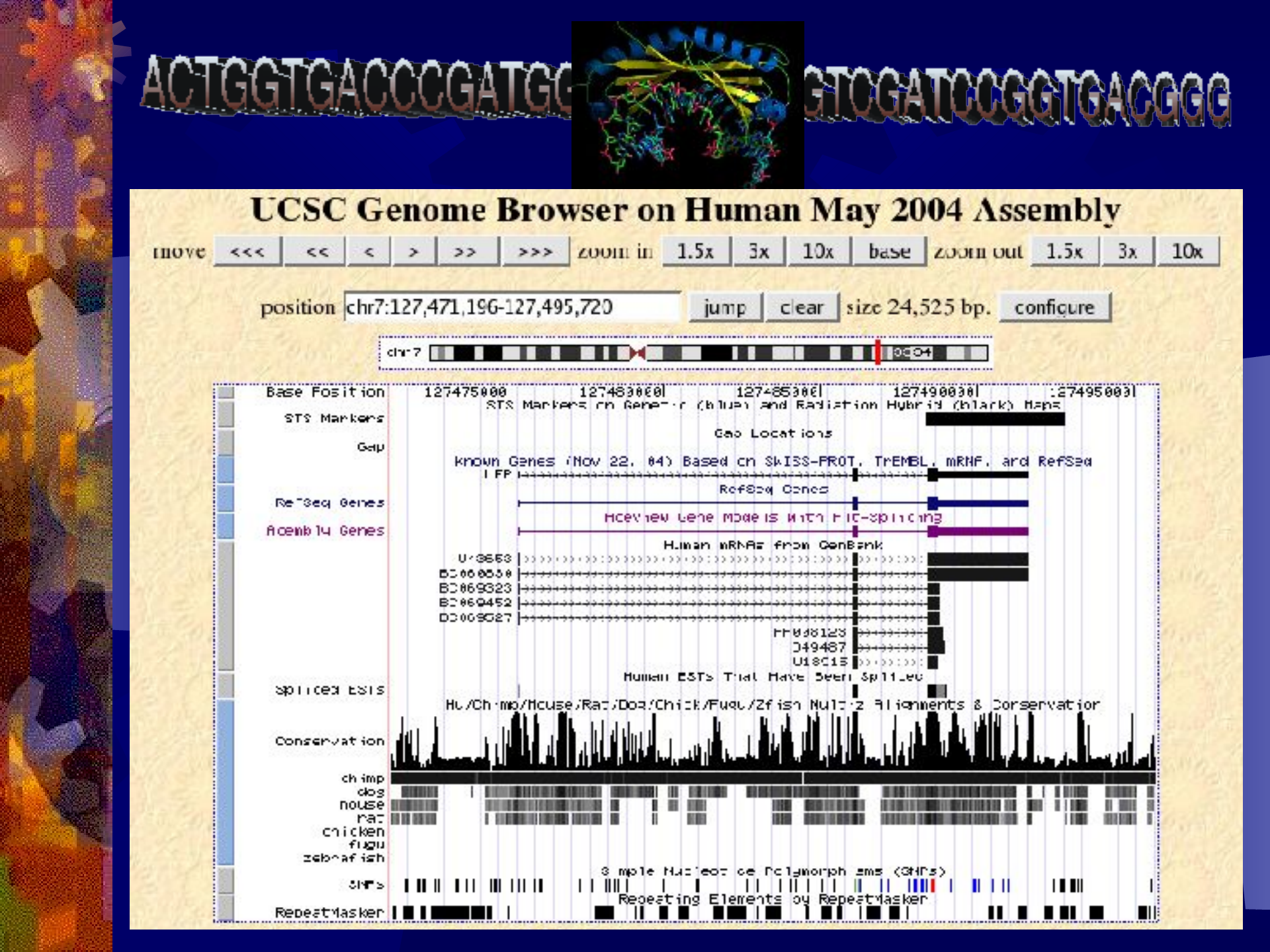
Jim Kent

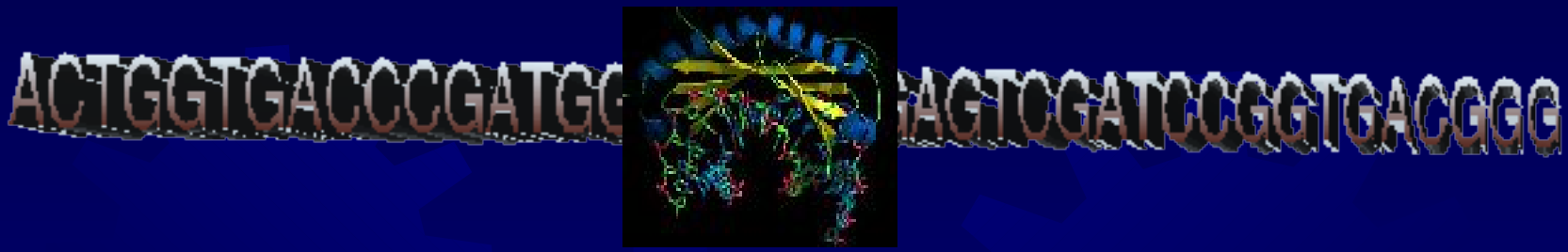
- autor Aegis Animator, Cyber Paint a Autodesk Animator

- po shlédnutí 12 CD-ROM vývojového prostředí pro Windows 95 přesedlává na bioinformatiku s odůvodněním, že lidský genom se vejde na jedno CD

- autor Genome Browser

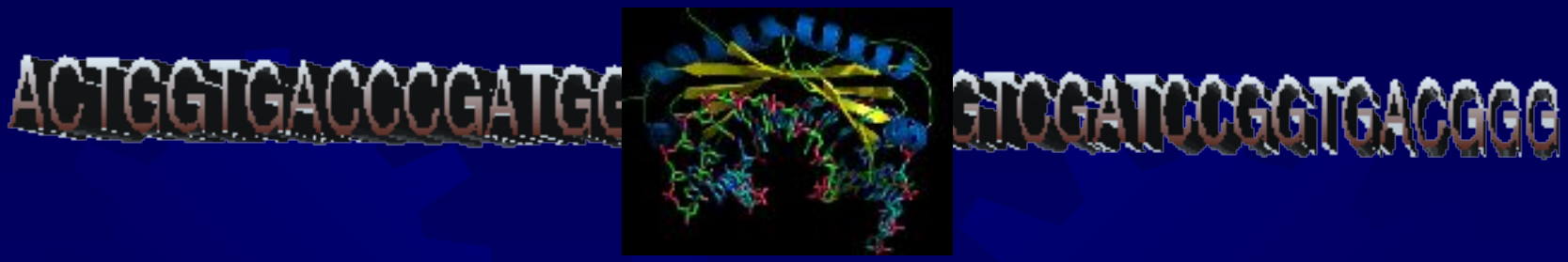
- sehrává důležitou roli v honičce o přečtení a skompletování lidského genomu (GigAssembler)


ACTGGTGACCCGATGG
GTCGATCCGGTGACGGG
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
Gap
Gap Locations
Known Genes (Nov 22, 04) Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
RefSeq Genes
RefSeq Genes
Assembly Genes
Acembly Gene Models With Alt-Splicing
Human mRNAs from GenBank
Spliced ESTs
Human ESTs That Have Been Spliced
Conservation
Multiz Alignments & Conservation
chimp
dog
mouse
rat
chicken
fugu
zebrafish
SNPs
Simple Nucleotide Polymorphisms (SNPs)
RepeatMasker
Repeating Elements by RepeatMasker



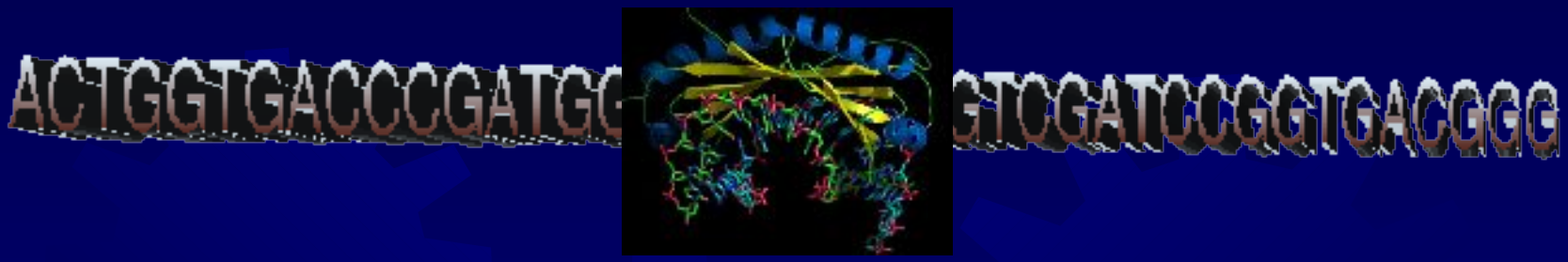
Human vs. Human

- 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.



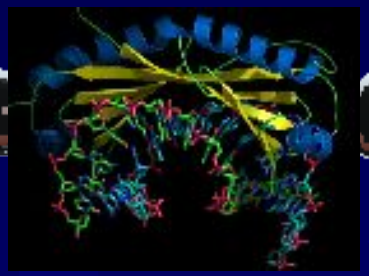
Human vs. Chimpanzee

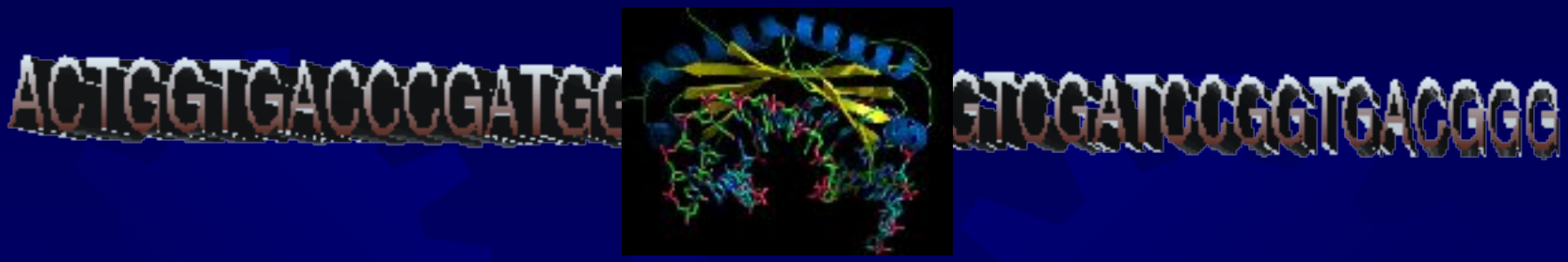
- A difference every 100 bases.
- A new transposon every 50000 bases
- Two chromosome in one species fused compared to the other.



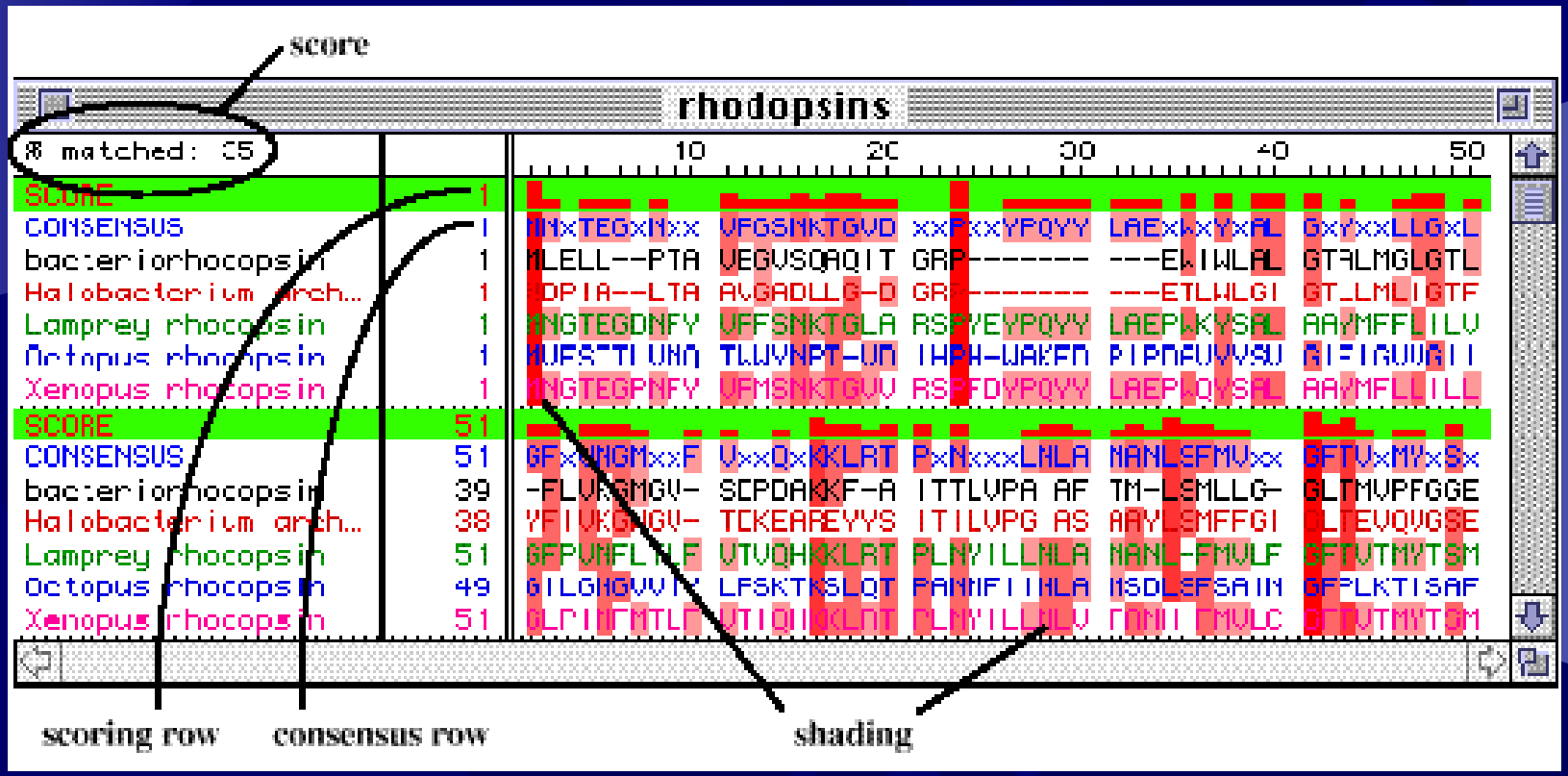
Human vs. Mouse

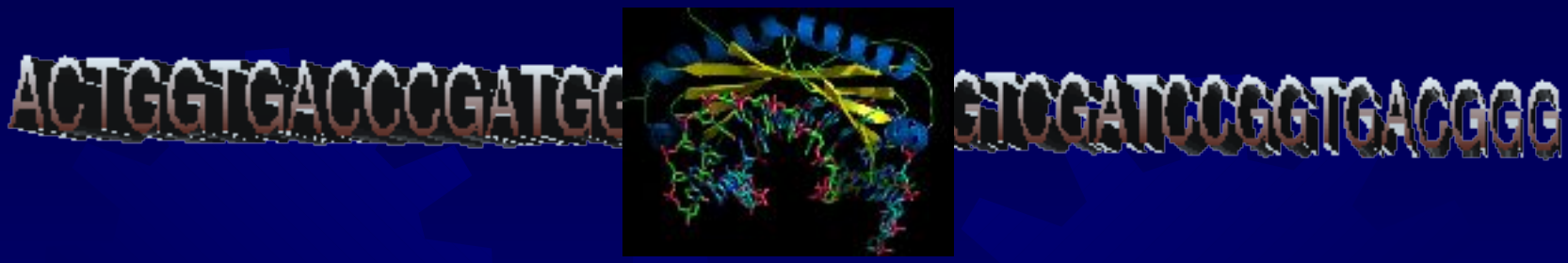
- In general 40% of bases have changed.
- In functional regions only 15% of bases have changed.
- Looking for conserved regions between human and mouse helps identify functional parts of human genome.

ACTGGTGACCCGATGG  GTCGATCCGGTGACGGG

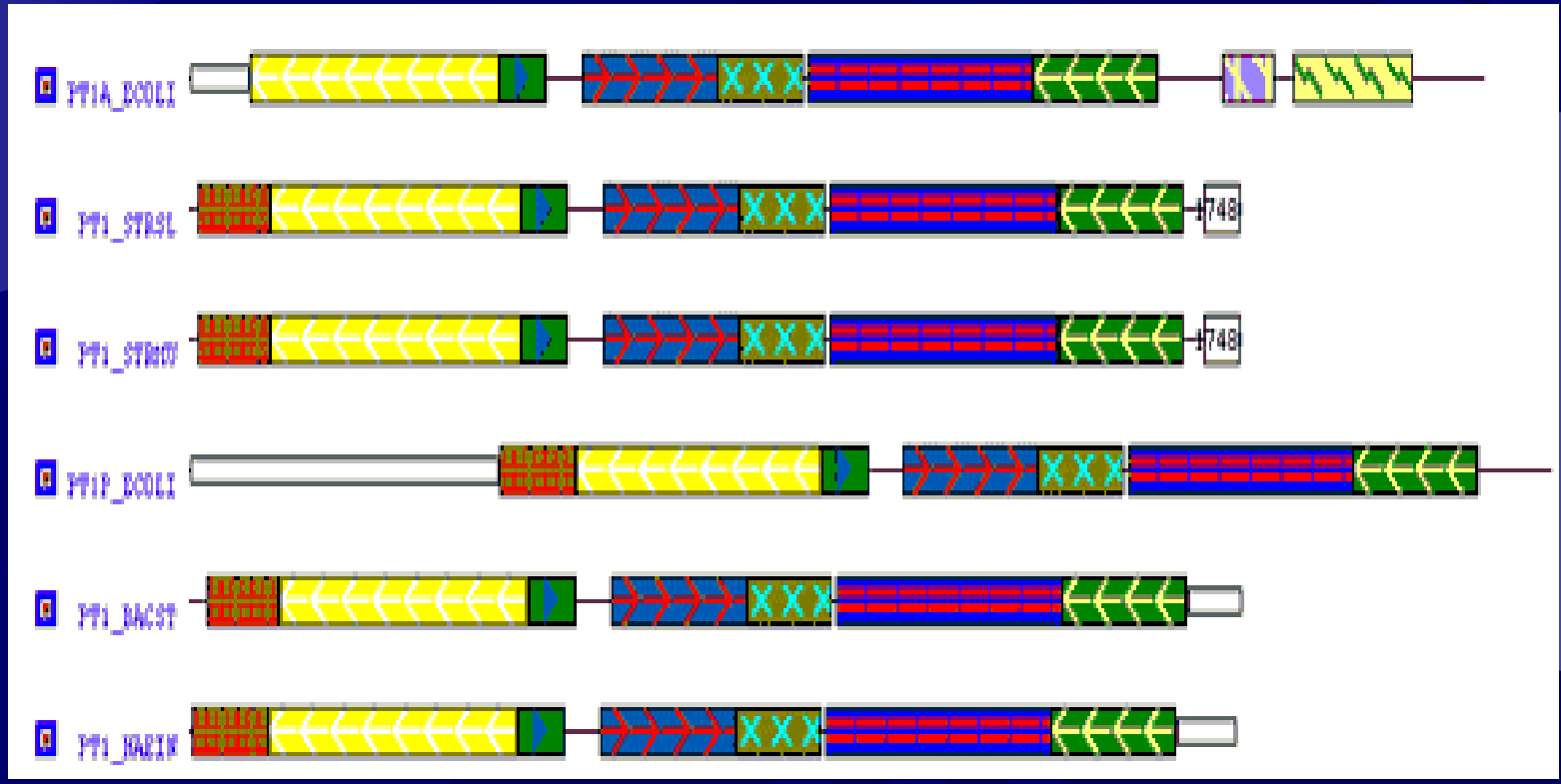


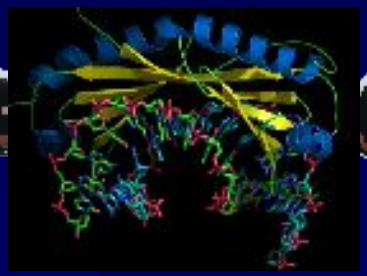
Co dělá bioinformatik



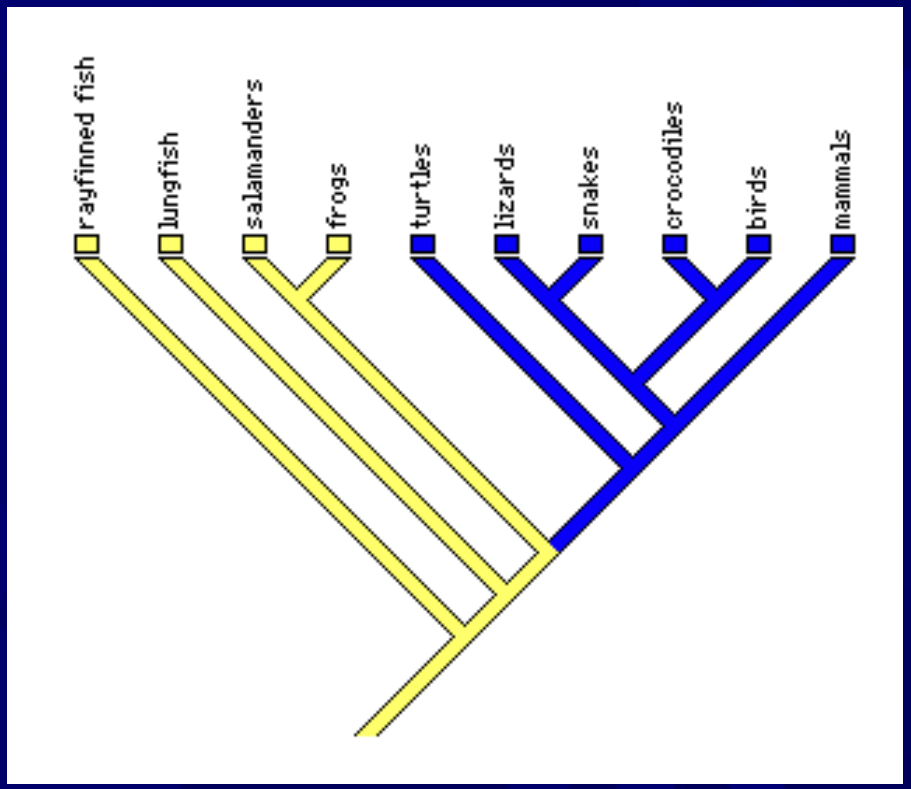


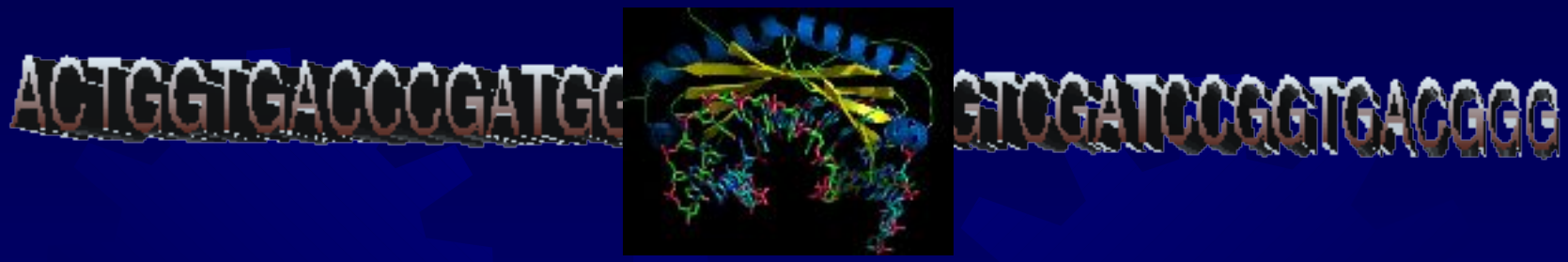
Co dělá bioinformatik



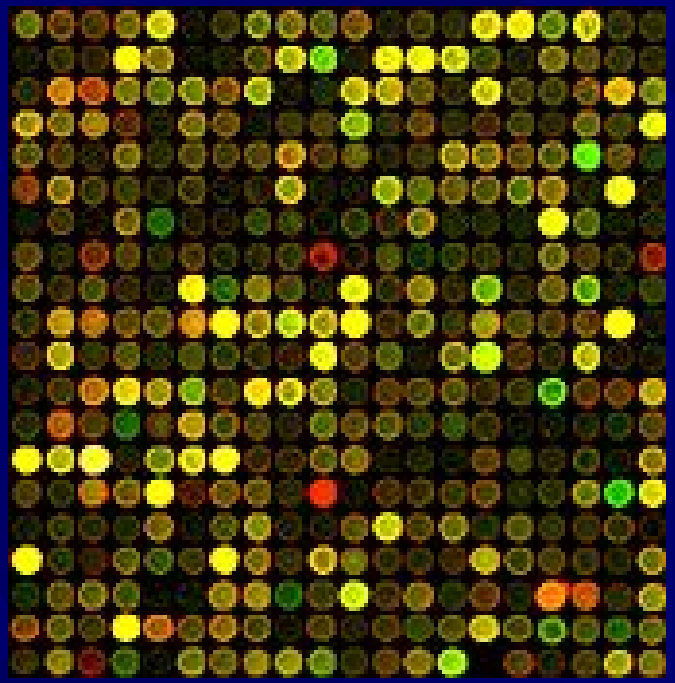


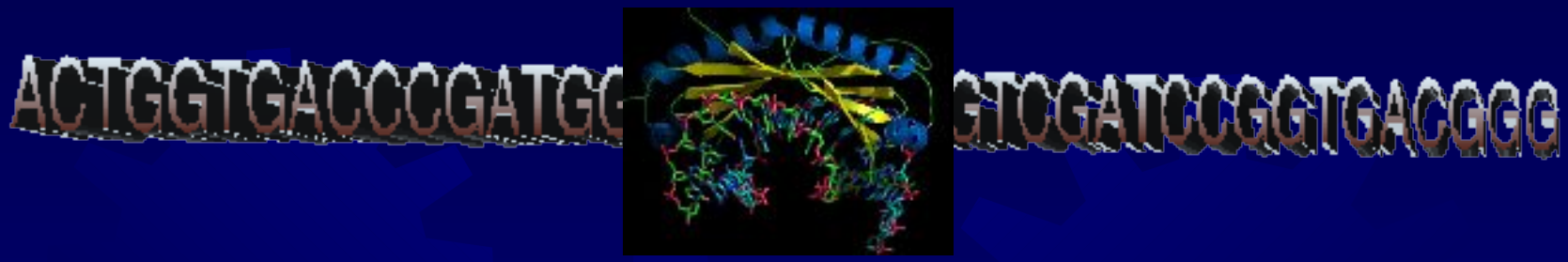
Co dělá bioinformatik



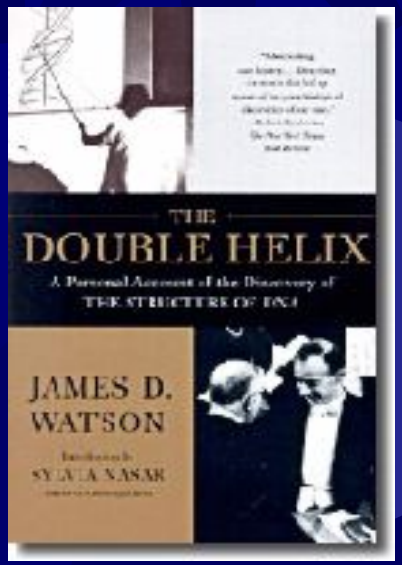


Co dělá bioinformatik?

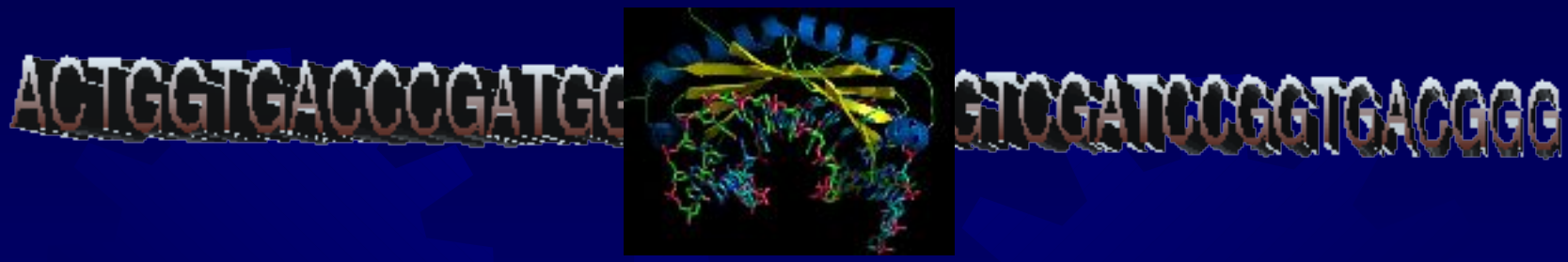




1953 – Watson, Crick, Franklin



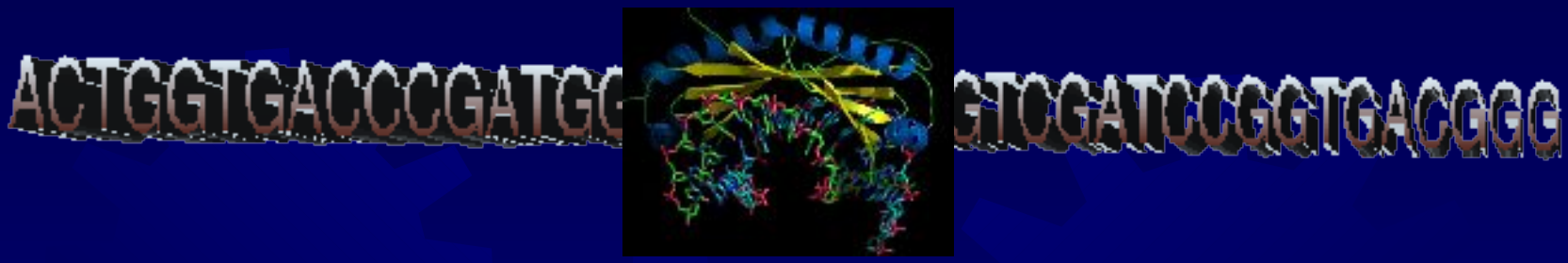
We wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.



1951 – Pauling	struktura proteinů
1952 – Turing	chemické základy vývoje
1953 – Watson and Crick	struktura DNA
1956 – Gamow et al.	genetický kód
1969 – Britten and Davidson	génová regulace

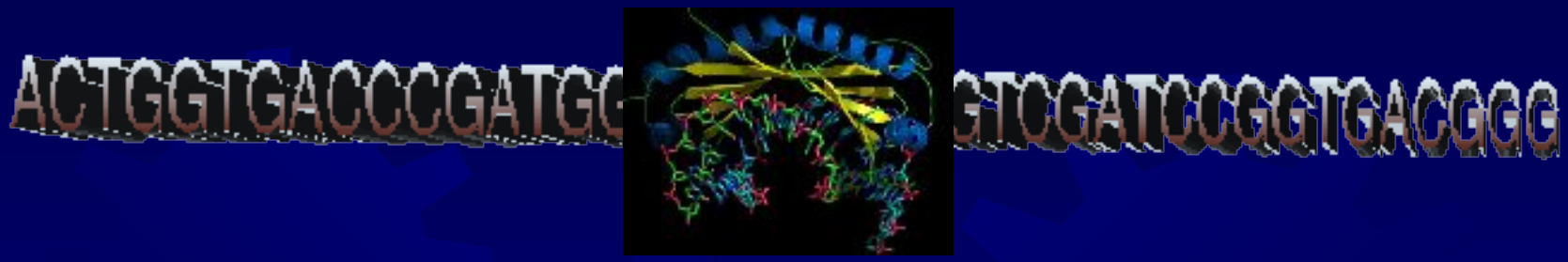
1959 – Chomsky	gramatiky
1962 – Shannon and Weaver	informační teorie
1966 – Martin-Lof	náhodné řetězce
1966 – Neumann	automata

Kořeny BIOINFORMATIKY sahají do 60. let

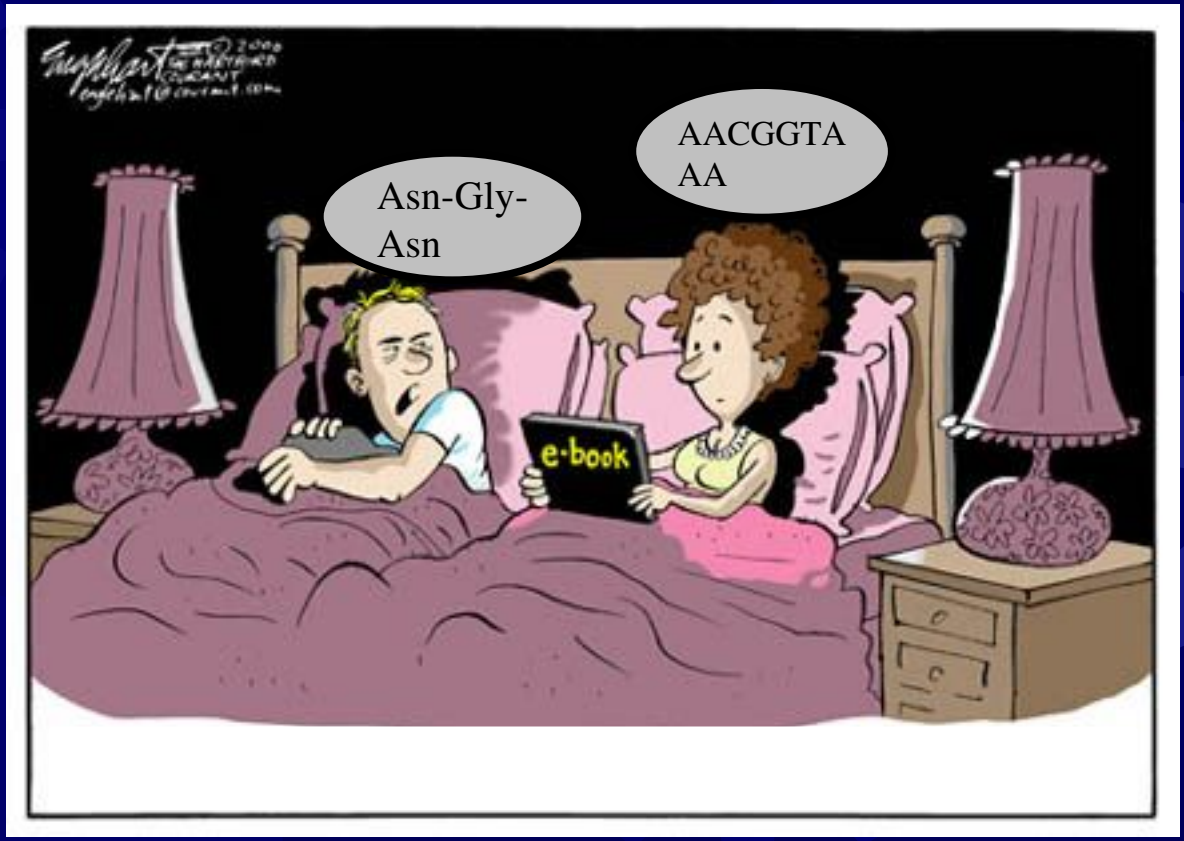


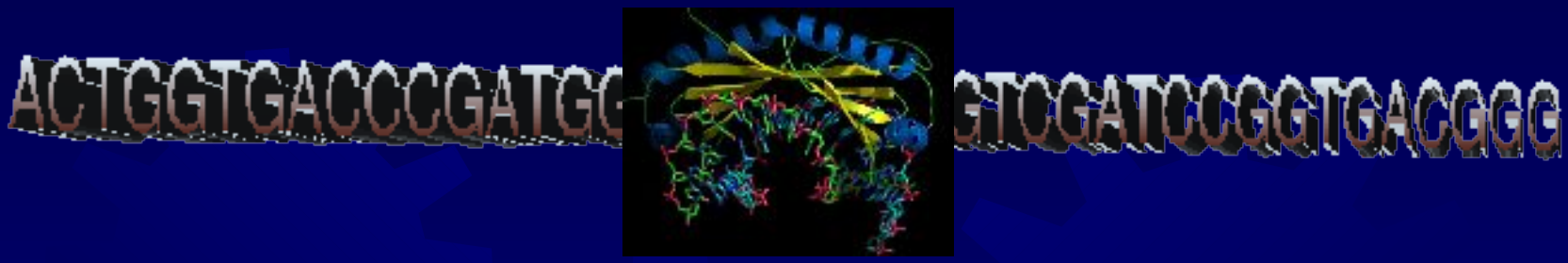
1965 – Zuckerkandl and Pauling	první použití sekvence v evoluční studii
1967 – Fitch and Margoliash	sestrojení prvních fylogenetických stromů
1970 – Needleman and Wunsh	užití dyn. programování k zarovnávání
1974 – Chou and Fasman	predikce sekundární struktury proteinů
1975 – Tanaka and Sheraga	simulace skládání proteinů
1978 – Dayhoff	první sbírka sekvencí proteinů
1981 – Smith and Waterman	modifikace algoritmu pro zarovnávání
1984 – Kabsch and Sander	modelování struktury proteinů
1986 – Bilofsky et al.	GenBank
1986 – Hamm and Cameron	EMBL Data Library
1987 – Feng and Doolitle	mnohonásobné zarovnání sekvencí
1987 – Gribskov	analýza sekvenčních profilů
1990 – Altschul et al.	efektivní hledání lokálních podobností

1998 – The journal Comp Appl Biosci becomes Bioinformatics

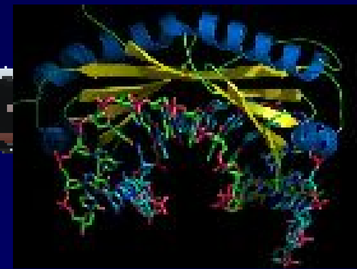


CENTRÁLNÍ DOGMA DNA – RNA – PROTEIN



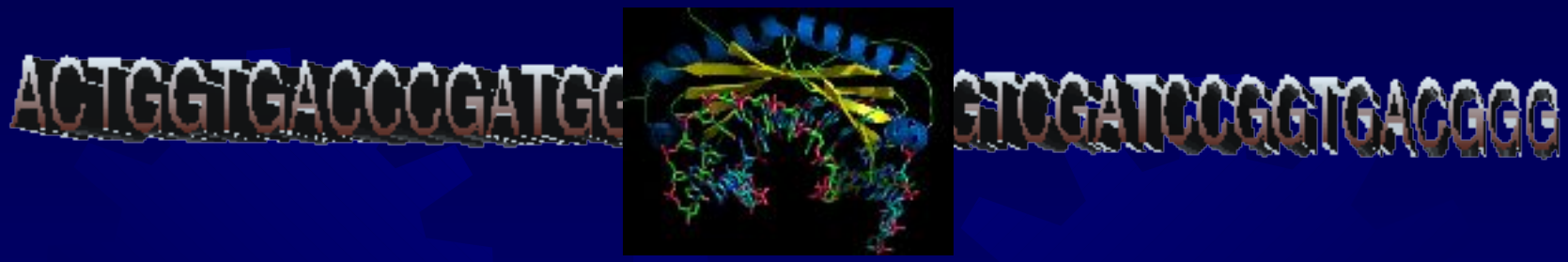


CENTRÁLNÍ DOGMA 2? PROTEIN/GEN – STRUKTURA - FUNKCE



Aktuální problémy

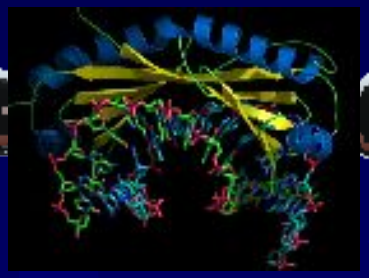
```
AAC  GGT  AAA                    Assembler?
 |    |    |
Asn-Gly-Asn


             MASAQSF

       /         |         \           C++?/English
     _   _     ___    ___     _
   _/ \_/   \_/    _/ \/\/
```



Aktuální problémy

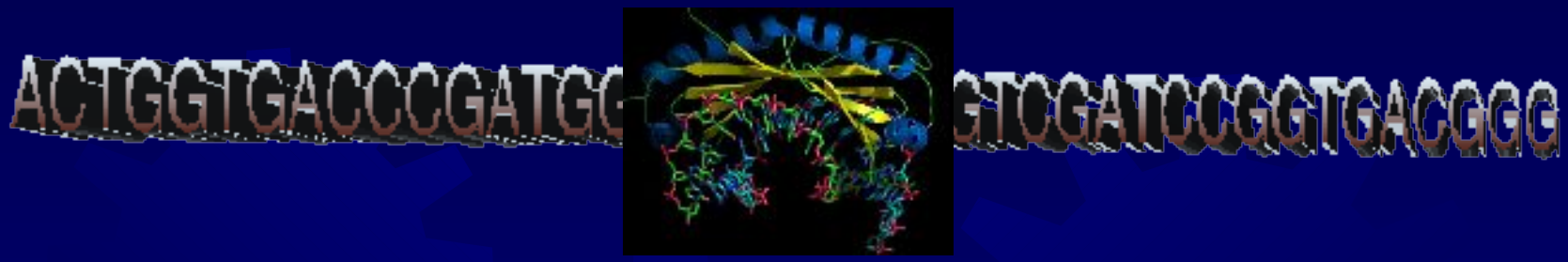
BIOLOGICKÉ SEKVENCE JAKO JAZYK

PROTEIN/GEN	STRUKTURA	FUNKCE
VĚTA	SYNTAX	VÝZNAM



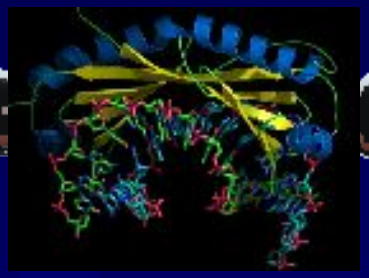
Aktuální problémy

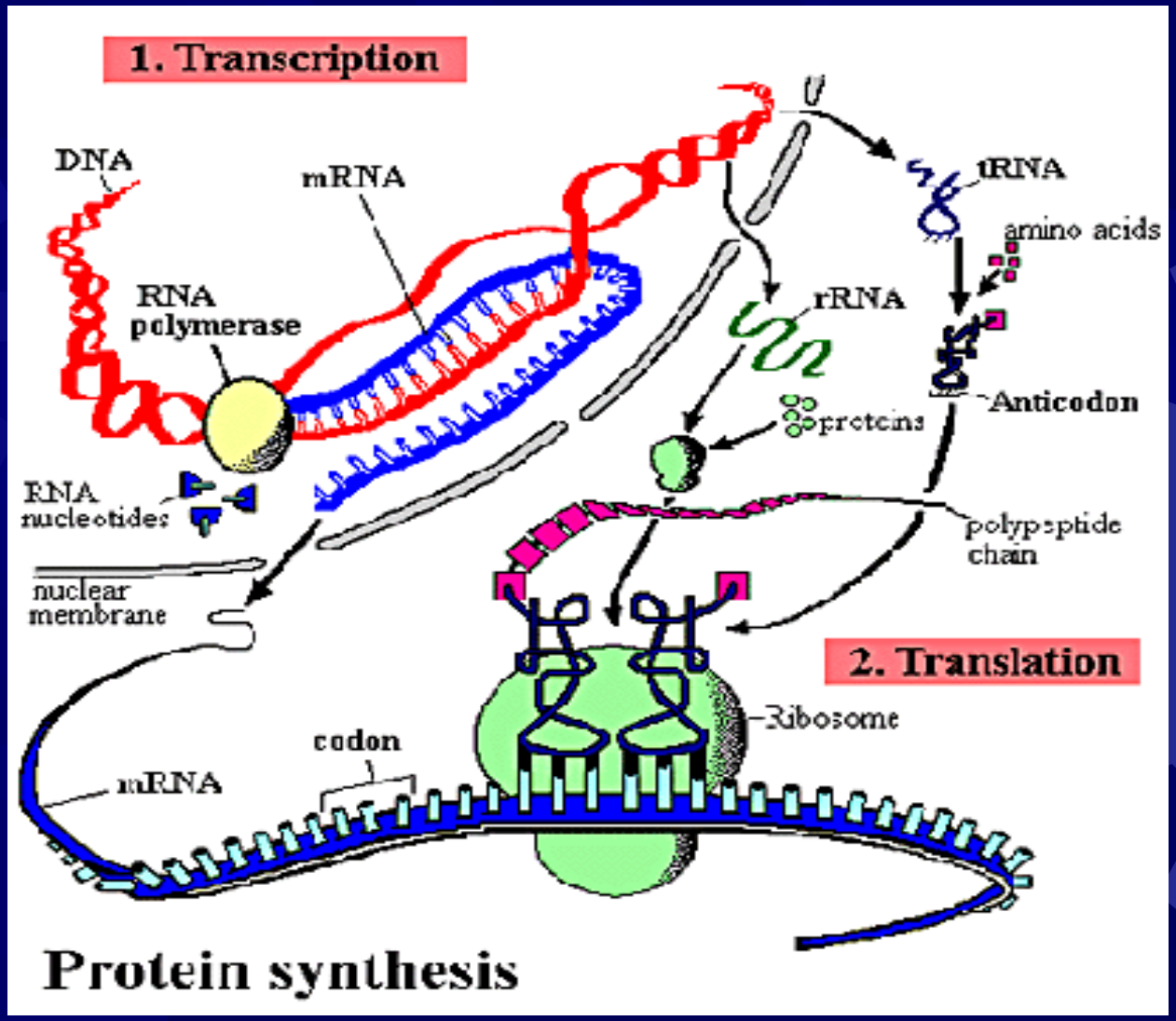
```
Mám z toho velkou radost.
Mám toho kocoura dost.

Mámztohovelk__ouradost.
::: ::::   :  ::::::::
Mám_toho___kocouradost.
```

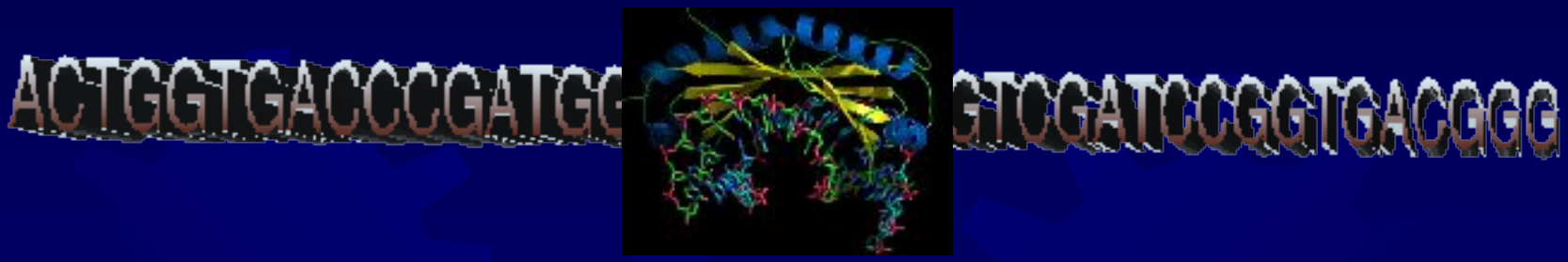


Aktuální problémy

01000101001000001111
11010100100101010010
10101001010010010100
10100101010100010010
10101001010101001010
10101010010101010010

ACTGGTGACCCGATGG  GTCGATCCGGTGACGGG





Centrální dogma

- DNA -> RNA -> PROTEIN