Bioinformātika

Nils Rostoks
Latvijas Universitāte
Bioloģijas fakultāte

Lekcijas plāns

- Bioinformātika
- Bioloģiskā informācija - tās daudzveidība un apjoms
- Bioloģiskās informācijas datubāzes
- Genomu organizācija un evolūcija
- Salīdzinošā genomika

Mācību materiāli I

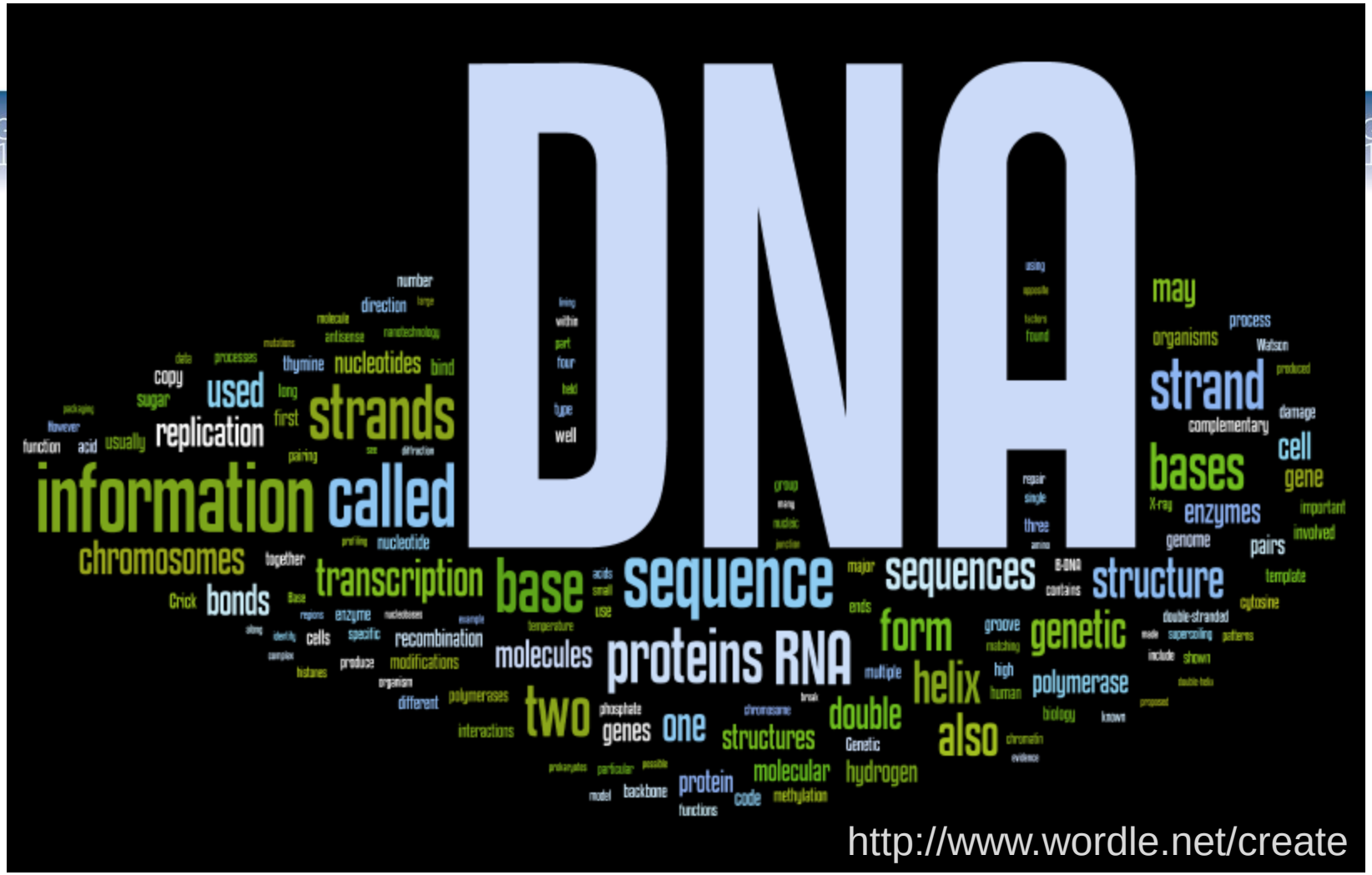
- Lekcijas
- Lesk AM (2008) Introduction to Bioinformatics. 3rd ed. Oxford University Press, New York, USA
- Lesk AM (2005) Introduction to Bioinformatics. 2nd ed. Oxford University Press, New York, USA
- Pevzner P, Shamir R (2011) Bioinformatics for biologists. Cambridge University Press, Cambridge, UK
- Higgs PG, Attwood TK (2006) Bioinformatics and molecular evolution. Blackwell Publishing, Malden, USA, Oxford, UK
- Claverie, Jean-Michel Bioinformatics for dummies 2003 LUB:Biologijas-zin.-bibl., LUB:Centr.bibl.-krājums
- Higgs, Paul G. Bioinformatics and molecular evolution 2005 LUB:Biologijas-zin.-bibl., LUB:Juridisko-zin.-bibl.
- Mount, David W. Bioinformatics 2001 LUB:Centr.bibl.-krājums

Mācību materiāli II

- Interneta resursi:

  http://www.ebi.ac.uk/training/online/
  http://www.ncbi.nlm.nih.gov/education/
  http://bioinformatics.oxfordjournals.org/

GAACATTCCATGCTCATGGGTAGGAAGAATCAATATCGTGAAAATGGCCATACTGCCCAAGGTAA
TTTATAGATTCAATGCCATCCCCATCAAGCTACCAATGACTTTCTTCACAGAATTGGAAAAAACT
ACTTTAAAGTTCATATGGAACCAAAAAAGAGCCCGCATTGCCAAGTCAATCCTAAGCCAAAAGAA
CAAAGCTAGAGGCATCACACTACCTGACTTCAAACTATACTACAAGGCTACAGTAACCAAAACAG
CATGGTACTGGTACCAAAGCAGAGATATAGACCAATGGAACAAAACAGTGCCCTCAGAAATAATA
CTGCATATCTACAACCATCTGATCTTTGACAAACCTGACAAAAACAAGCAATGGGGAAAGGATTC
CCTATTTAATAAATGGTGCTGGGAAAACTGGCTAGCCATATGTAGAAAGCTGAAATTGGATCCCT
TCCTTACACCTTGTACAAAAATTAATTCAAGATGGATTACAGACTTAAATGTTAGACCTAAAACC
ATAAAAACCCTAGAAGAAAACCTAGGCAATACCATTCAGGACATAGGCATGGGCAAGAACTTCAT
GTCTAGAACACCAAAAGTAATGGCAACAAAAGCCAAAATTGACAAATGGGTCTAATTAAACTAAA
GAGCTTCTGCACAGCAAAAGAAACTACCATCAGAGTGAAGAGGCAACCTACAGAATGGGAGAAAA
TTTTTGCAATCTGACAAAAGGGCTAATTTTTGCATCTGACAAAGGGCTAATATCCAGAATCTACA
ATGAACTCAAACAAATTTACAAGAAAAAAACAAATTTACAAGAAAAAAACAAAATTTACAAGAAA
AAAACAAATTTACAAGAAAAAAACAAACAACCCCATCAAAAAGTGGGCAAAGGATAAGAACAGTC
ACTTCTCAAAAGAAGACATTTATGCAGCCAAAAGACACATGAAAAAATTCTCATCATCACTGGCC
ATCAGAGAAATGCAAATCAAAACCACAATGAGATACCATCTCACACCAGTTAGAATGGCGATCAT
TAAAAAGTCAAGAAACAACAGGTGCTGGAGAGGATGTGGAGAAATAGGAACACTTTTACACTGTT
AGTGGGACTGTAAACTAGTTCAACATTGTGGAAGTCAGTATGGCGATTCCTCAGGGATCTAGAAC
TAGAAATACCATTTGACCCAGCCATCCCATTACTGGGTATATACCCAAAGGATTATAAATCATGC
TGCTATAAAGACACTTGCACACATATGTTTATTGTGGTACTATTCACAATAGCAAAGACTTGGAA
CCAACCCAAATGTCCAACAATGATAGACTAGATTAAGAAAATGTGGCACATATACACTATGGAAT
ACTTTGCAGCCATAAAAAAGGATGAGTTCATGTCCTTTGTAGGGACATGGATGAAGCTGGAAACC
ATCATTCTCAGCAAACTATCGCAAGGATAAAAAACCAAACACCGCATGTTCTCACTCATAGGTGG
TAACTGAACAATGAGAACACATGGTCACAGGAAGGGGAACATCACGCACTGGGGCCTGTTGTGGG
GTGGGGGGAGTGGGGAGGGATAGCATTAGGAGACATACCTAATGTTAAATGACGATTTAATGGGT



https://en.wikipedia.org/wiki/DNA

Kas ir bioinformātika?

Bioinformatics derives knowledge from computer analysis of biological data. These can consist of the information stored in the genetic code, but also experimental results from various sources, patient statistics, and scientific literature. Research in bioinformatics includes method development for storage, retrieval, and analysis of the data. Bioinformatics is a rapidly developing branch of biology and is highly interdisciplinary, using techniques and concepts from informatics, statistics, mathematics, chemistry, biochemistry, physics, and linguistics. It has many practical applications in different areas of biology and medicine

M. Nilges, J.P. Linge, Institut Pasteur

Kas ir bioinformātika?

Bioinformatics: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.

Computational Biology: The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems

NIH Biomedical Information Science and Technology Initiative Consortium

Kas ir bioinformātika?

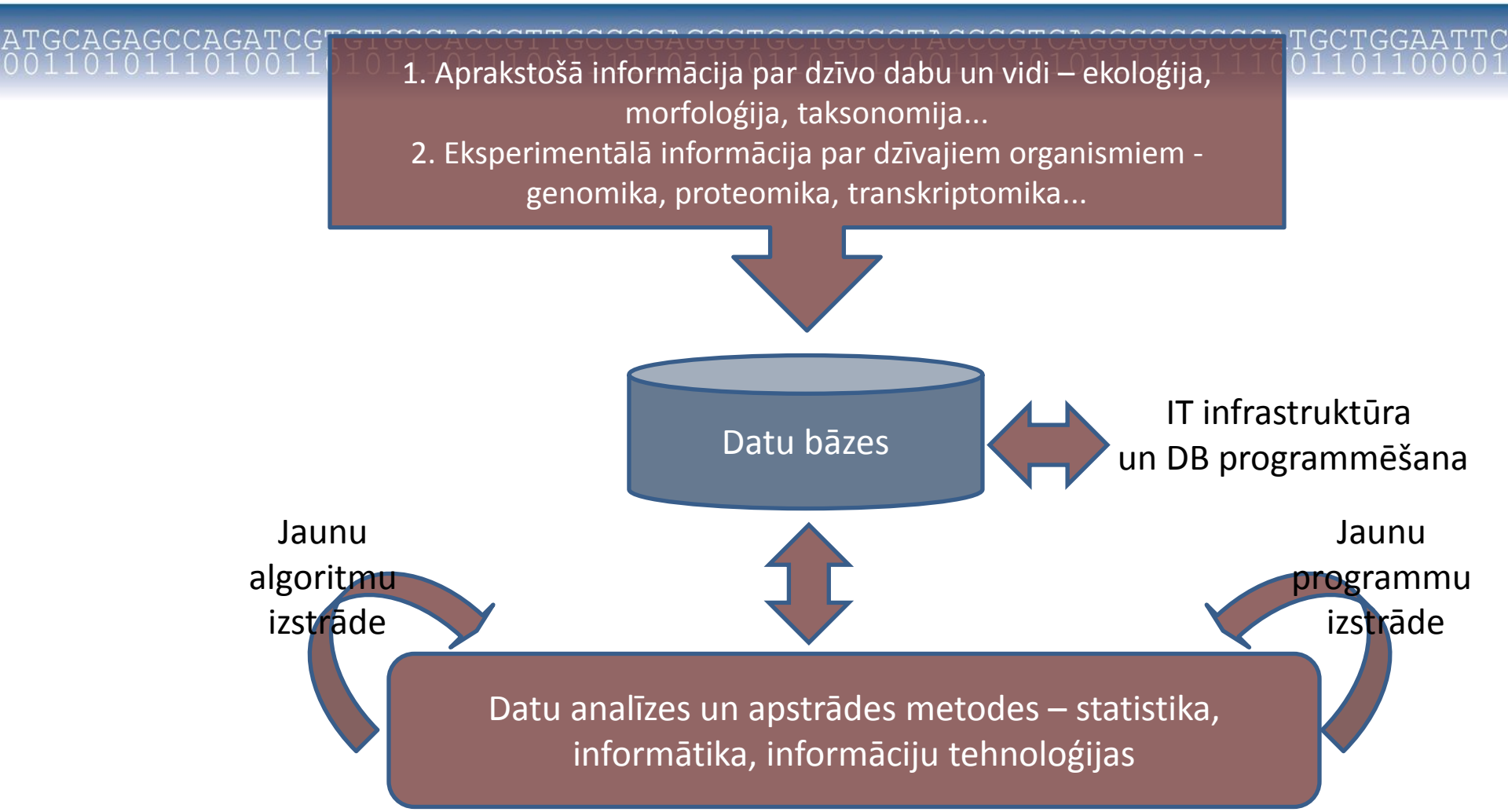
Bioinformatics is the field of science in which biology, computer science, and information technology merge to form a single discipline. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned

http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html

Bioinformātikas pētījumu objekts

- Genoma nukleotīdu secības (DNS/RNS) (genomika)
- Genoma ekspresija (dažādu RNS veidu analīze) (transkriptomika)
- Genoma kodētie proteīni (proteomika)
- Šūnas veidotie metabolīti (metabolomika)
- Mijiedarbības dažādu objektu starpā (sistēmbioloģija)

Bioloģija, statistika, programmēšana, informāciju tehnoloģijas



Bioloģiskā informācija - tās daudzveidība un apjoms

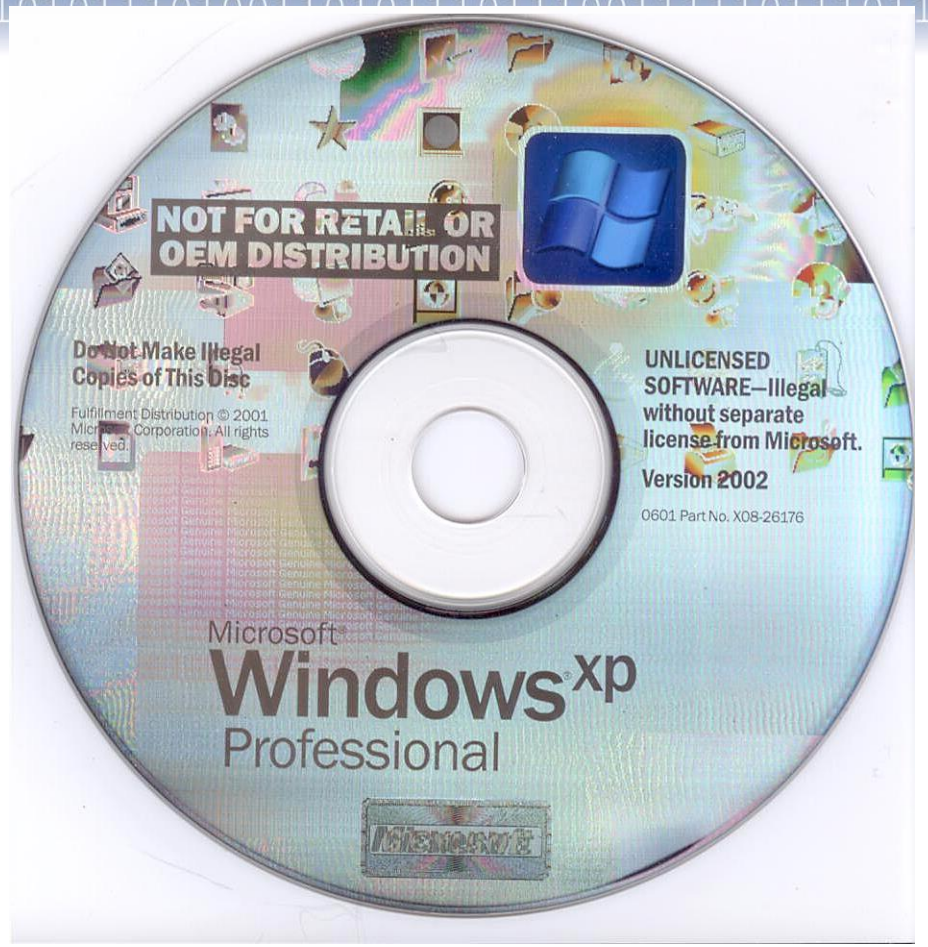
Digitālais cilvēka genoms

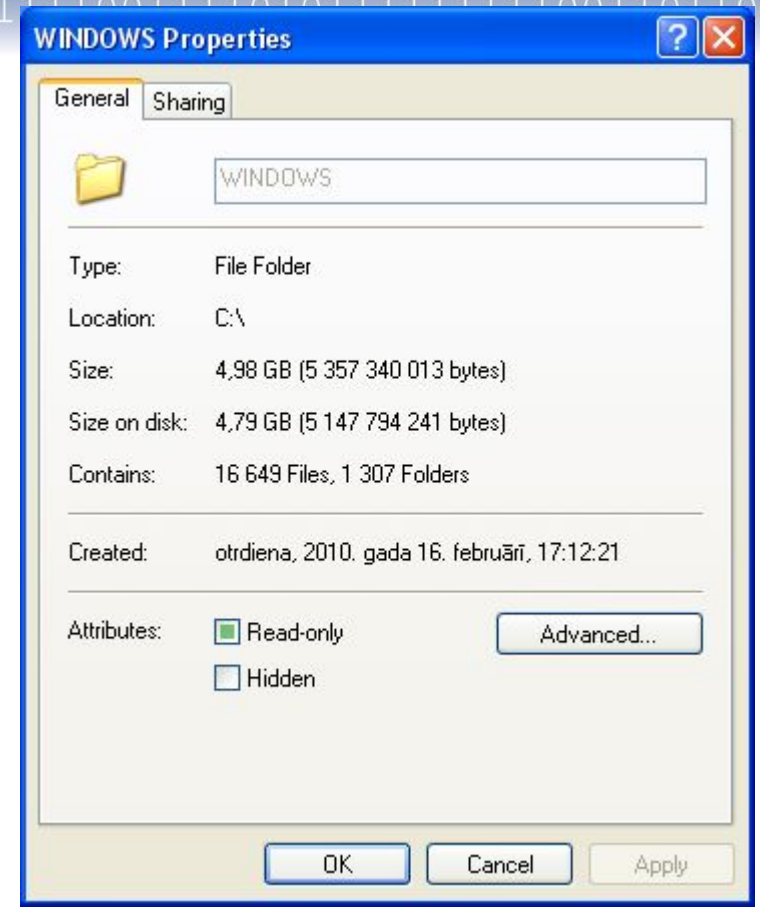
- Diploīds genoms – apmēram 6 x 10^9 nukleotīdu
- Viena nukleotīda kodēšanai binārā formā nepieciešami vismaz 2 biti:

  A = 00, C = 01, G = 10, T = 11
- 8 biti = 1 baits
- 1 baits var kodēt 4 nukleotīdus
- Diploīds cilvēka genoms binārā formā ir 1.5 x 10^9 baiti (datora darbībai vajag vairāk koda nekā dzīvības nodrošināšanai)

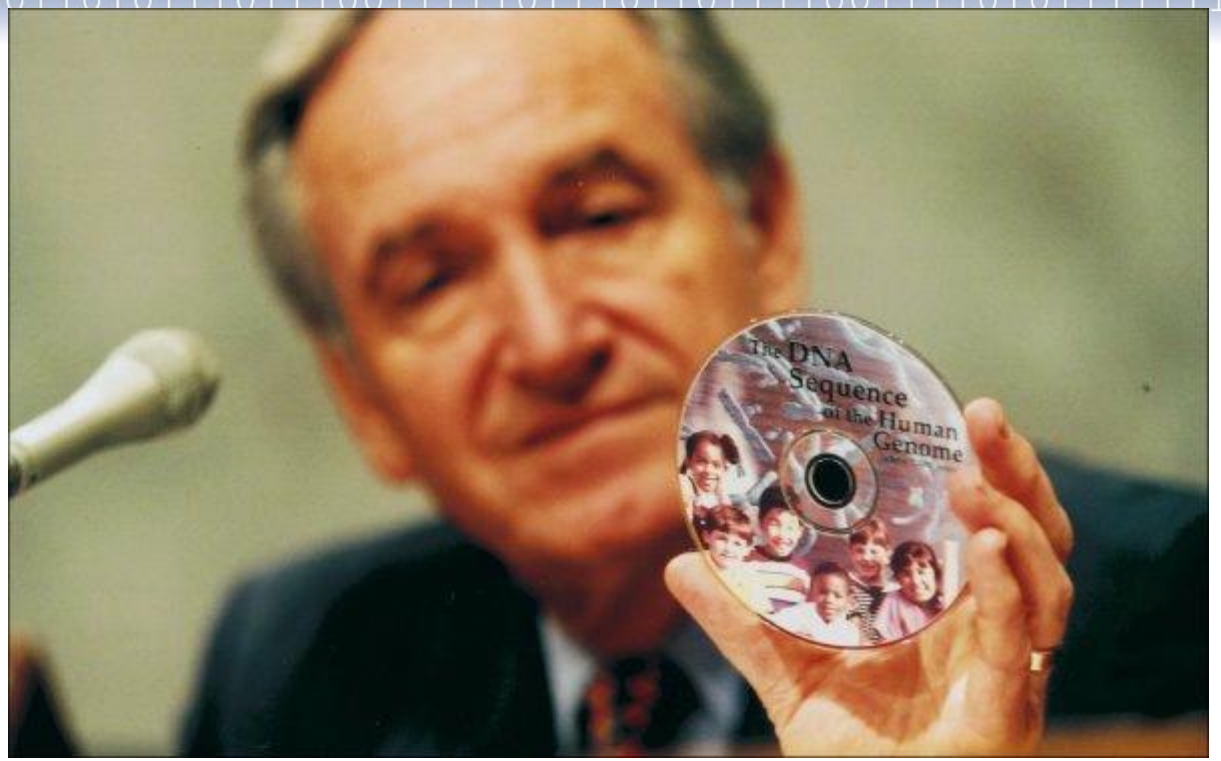
http://www.tmsoft.com/article-genome.html

Cilvēka genoms un MS Windows operētājsistēma?

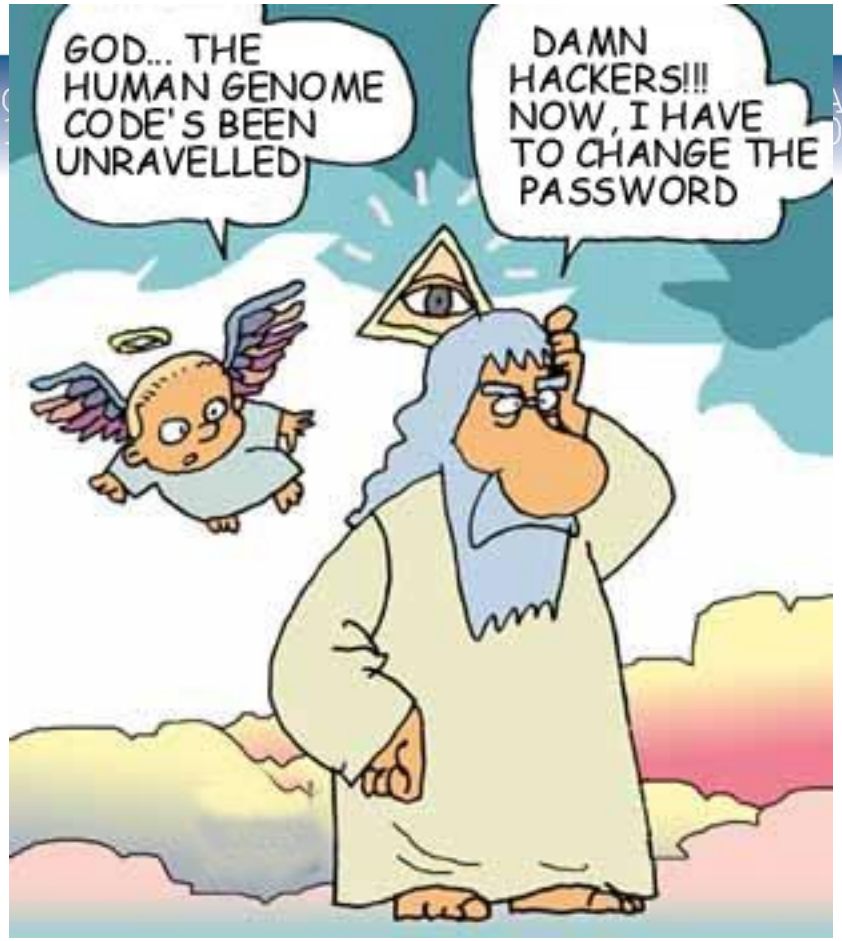




Cilvēka genoma sekvence



http://www.genome.gov/dmd/img.cfm?node=Photos/People/People%20Outside%20NHGRI&id=79269


GOD... THE HUMAN GENOME CODE'S BEEN UNRAVELLED
DAMN HACKERS!!! NOW, I HAVE TO CHANGE THE PASSWORD

Datorprogrammas DNS sekvences analīzei

Volume 4 Number 11 November 1977 Nucleic Acids Research

Sequence data handling by computer

R.Staden

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

Received 10 October 1977

ABSTRACT

The speed of the new DNA sequencing techniques has created a need for computer programs to handle the data produced. This paper describes simple programs designed specifically for use by people with little or no computer experience. The programs are for use on small computers and provide facilities for storage, editing and analysis of both DNA and amino acid sequences. A magnetic tape containing these programs is available on request.

Genomā ietvertā informācija nosaka organisma identitāti

... vismaz baktērijās ...

Genoma transplantēšana baktērijās

Originally published in *Science* Express on 28 June 2007
Science 3 August 2007:
Vol. 317. no. 5838, pp. 632 - 638
DOI: 10.1126/science.1144622

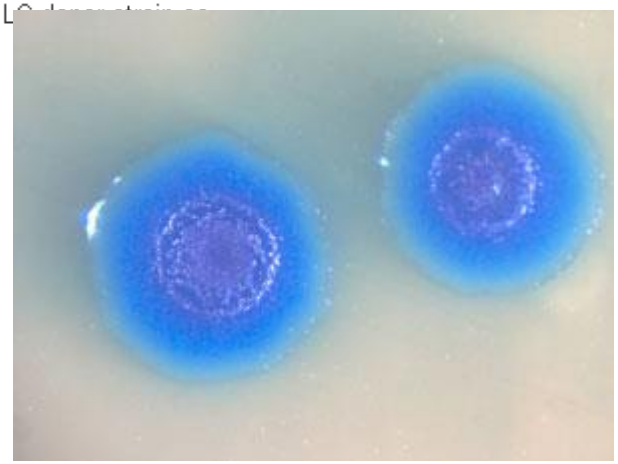
RESEARCH ARTICLES

Genome Transplantation in Bacteria: Changing One Species to Another

Carole Lartigue, John I. Glass,* Nina Alperovich, Rembert Pieper, Prashanth P. Parmar, Clyde A. Hutchison, III, Hamilton O. Smith, J. Craig Venter

As a step toward propagation of synthetic genomes, we completely replaced the genome of a bacterial cell with one from another species by transplanting a whole genome as naked DNA. Intact genomic DNA from *Mycoplasma mycoides* large colony (LC), virtually free of protein, was transplanted into *Mycoplasma capricolum* cells by polyethylene glycol–mediated transformation. Cells selected for tetracycline resistance, carried by the *M. mycoides* LC chromosome, contain the complete donor genome and are free of detectable recipient genomic sequences. These cells that result from genome transplantation are phenotypically identical to the *M. mycoides* L[illegible] judged by several criteria.

The J. Craig Venter Institute, Rockville, MD 20850, USA.



Pilna baktērijas genoma ķīmiskā sintēze

Originally published in *Science* Express on 24 January 2008
Science 29 February 2008:
Vol. 319. no. 5867, pp. 1215 - 1220
DOI: 10.1126/science.1151721

RESEARCH ARTICLES

Complete Chemical Synthesis, Assembly, and Cloning of a *Mycoplasma genitalium* Genome

Daniel G. Gibson, Gwynedd A. Benders, Cynthia Andrews-Pfannkoch, Evgeniya A. Denisova, Holly Baden-Tillson, Jayshree Zaveri, Timothy B. Stockwell, Anushka Brownley, David W. Thomas, Mikkel A. Algire, Chuck Merryman, Lei Young, Vladimir N. Noskov, John I. Glass, J. Craig Venter, Clyde A. Hutchison, III, Hamilton O. Smith*

We have synthesized a 582,970–base pair *Mycoplasma genitalium* genome. This synthetic genome, named *M. genitalium* JCVI-1.0, contains all the genes of wild-type *M. genitalium* G37 except MG408, which was disrupted by an antibiotic marker to block pathogenicity and to allow for selection. To identify the genome as synthetic, we inserted "watermarks" at intergenic sites known to tolerate transposon insertions. Overlapping "cassettes" of 5 to 7 kilobases (kb), assembled from chemically synthesized oligonucleotides, were joined by in vitro recombination to produce intermediate assemblies of approximately 24 kb, 72 kb ("1/8 genome"), and 144 kb ("1/4 genome"), which were all cloned as bacterial artificial chromosomes in *Escherichia coli*. Most of these intermediate clones were sequenced, and clones of all four 1/4 genomes with the correct sequence were identified. The complete synthetic genome was assembled by transformation-associated recombination cloning in the yeast *Saccharomyces cerevisiae*, then isolated and sequenced. A clone with the correct sequence was identified. The methods described here will be generally useful for constructing large DNA molecules from chemically synthesized pieces and also from combinations of natural and synthetic DNA segments.

Informācijas uzglabāšana DNS sekvencē

Sciencexpress

Brevia

Next-Generation Digital Information Storage in DNA

George M. Church,[1,2] Yuan Gao,[3] Sriram Kosuri[1,2]*

[1]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA. [2]Wyss Institute for Biolo Inspired Engineering, Boston, MA 02115, USA. [3]Department of Biomedical Engineering, Johns Hopki University, Baltimore, MD 21205, USA.

*To whom correspondence should be addressed. E-mail: sri.kosuri@wyss.harvard.edu

As digital information continues to accumulate, higher density and long-er-term storage solutions are necessary (*1*). DNA has many potential advantages as a medium for immutable, high latency information storage needs (*2*). For example, DNA storage is very dense. At theoretical max-imum, DNA can encode two bits per nucleotide (nt) or 455 exabytes per gram of ssDNA (*3*). Unlike most digital storage media, DNA storage is not restricted to a planar layer, and is often readable despite degradation in non-ideal conditions over millennia (*4*, *5*). Finally, DNA's essential biological role provides access to natural reading and writing enzymes and ensures that DNA will remain a readable standard for the foreseea-ble future.

Storing messages in DNA was first demonstrated in 1988 (*6*) and the

for all but century-sca
sis and sequencing ha
fold per year, respect
fold per year (*11*). I
becoming available,
information (*12*). Our
combined with librar
compatible with futu
Reciprocally, large-sc
could accelerate deve
(*13*). Future work co

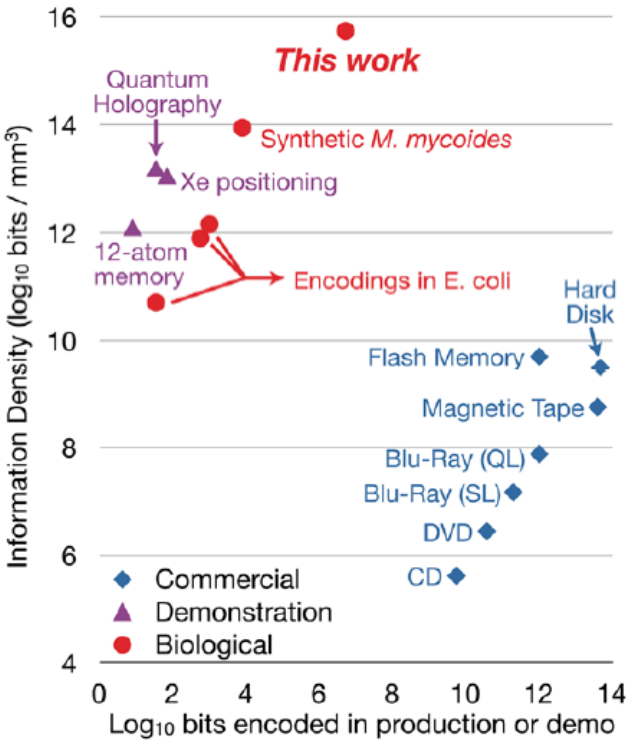


Fig. 1. Comparison to other measured by the $\log_{10}$ of bits encoded in the report or commercial technologies. We plotted information density ($\log_{10}$ of bits/mm^3) versus current scalability as unit (*3*).

29.05.2014.

wnloaded from www.sciencemag.org on August 20, 2012

DNS sekvenēšana

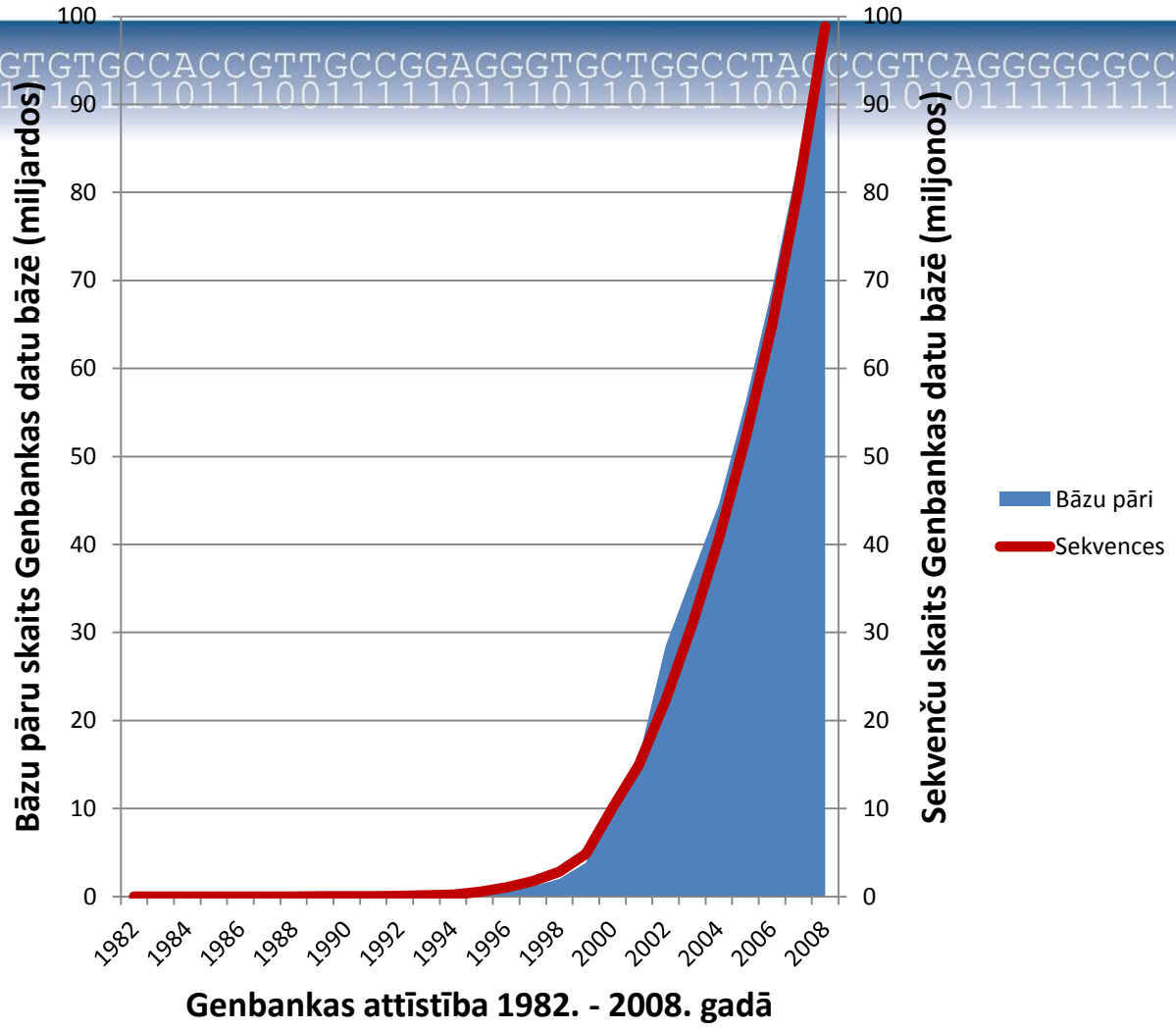
- Divas metodes 1975. - 1977. gadā (*Allan Maxam un Walter Gilbert, Frederick Sanger*)

"These chemical procedures ... soon allowed the entire sequence of the plasmid cloning vector pBR322 (4362 bp) to be worked out by a single scientist in only one year."

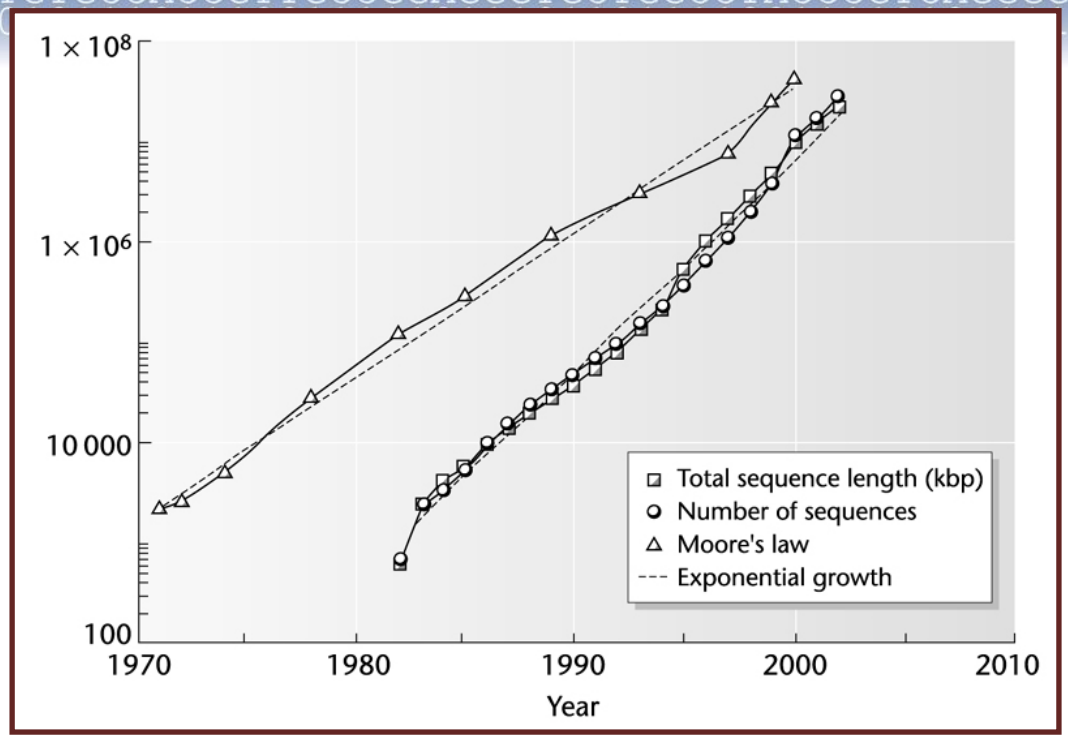
Molecular Biology of the Gene IVth ed. 1987

Genoma sekvenēšanas centri

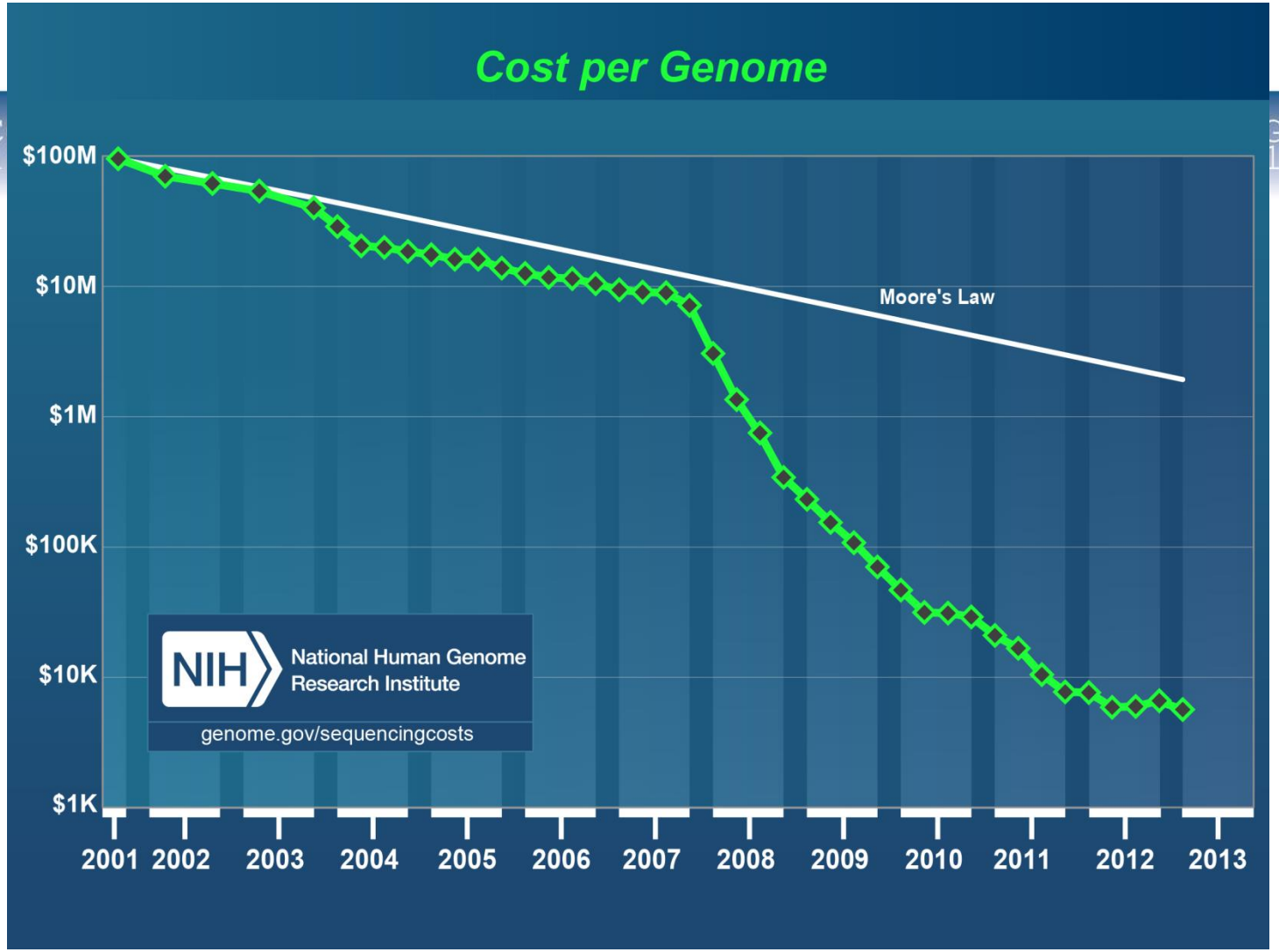
DNS sekvenču pieaugums GenBank



DNS sekvences un Mūra likums



Mūra likums – tranzistoru skaits datoru procesoros katru gadu dubultojas
Gan DNS sekvenču, gan tranzistoru skaita pieaugums uz datoru mikroshēmām ir eksponenciāls


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013

Indivīda genoma sekvence

- Cilvēka genoma projekts (1990. – 2003.) noteica pilnu cilvēka genoma sekvenci, taču tā ir dažādu indivīdu genomu hibrīds
- Levy et al. (2007) The diploid genome sequence of an individual human. PLoS Biol 5: e254

Type	Number of Variants	bp Length	Min	Max	Mean
heterozygous SNP	1,762,541	1,762,541	1	1	1
homozygous SNP	1,450,860	1,450,860	1	1	1
heterozygous MNP	38,985	227,531	2	206	5.8
homozygous MNP	14,838	31,590	2	22	2.1
heterozygous indel	263,923	635,314	1	321	2.4
Complex	28,179	330,803	2	571	11.7
homozygous insertion	275,512	3,117,039	1	82,711	11.3
homozygous deletion	283,961	2,820,823	1	18,484	9.9
inversion	90	1,914,477	7	670,345	21,272
Total	4,118,889	12,290,978			

1000 genomu sekvences

- The Genomes Project (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073

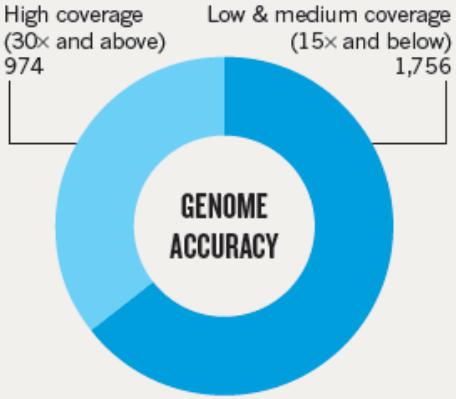
GRAPHICS BY NIK SPENCER; DATA COMPILED BY ALLA KATSNELSON

Why scientists want tens of thousands of genomes — and more

To understand populations

Comparing lots of genomes lets researchers identify points at which one genome differs from the next. Costs may be falling, but sequencing and data analysis are still pricey. So most researchers face a trade-off between the number of subjects and the accuracy in the sequences they can afford. For projects examining how populations commonly differ, sequencing a large number of individuals at relatively low accuracy or 'depth of coverage' is enough. About 900 genomes sequenced so far by the 1000 Genomes Project have been read three times on average.

Cost still limits the accuracy with which genomes are sequenced.



To understand disease

Researchers trying to uncover rare disease-linked mutations — perhaps limited to just one family or an individual — need precision, typically sequencing each genome 30 times on average. Cancer genomes, many sequenced under the auspices of large collaborations, account for a sizeable chunk of high-coverage genome sequences completed to date. Projects scrutinizing people with diabetes, Crohn's disease and other disorders are starting to emerge. Analysing all the genome data is a huge challenge, as is turning genetic discoveries into clinical benefits.

1026 | NATURE | VOL 467 | 28 OCTOBER 2010

Genomu struktūra un modernās metodes tās analīzei

Genoms

- Organisma genoms ir tā kopējā DNS molekulas nukleotīdu secība
- Prokariotu genomu veido hromosoma(s) un plazmīdas
- Eikariotu genoms sastāv no kodola genoma, kā arī mitohondriju un plastīdu (augiem) genoma

Genomika

- Bioloģijas zinātnes apakšnozare, kas nodarbojas ar sistemātisku un pilnīgu iedzimtības elementu pētīšanu molekulārā līmenī
- Genomika apraksta kā genomā kodētā informācija nosaka šūnu, organismu, populāciju un ekosistēmu attīstību

Bioinformātika un genomu struktūras anotācija


```
1981 tatgtacgta cggaacgaaa cagtagtgat caatgcgcga ggccacctag ctagctgcgc
2041 catgtatatg taagctatcc gtccacctca cgcgccaccg cgagctgatg cccgcggccg
2101 cgcgatcgac tcgacccacg gcctcgccgg cgcagccgta gccagctagc cgcgcgcgca
2161 ccgccgccat gttcgggcgag ctctccgcgc gctcgtcgcc ctcgttcgcc gcctcgccct
2221 ccggcgccga cgacgcgcgg gcaggagagt cgacgccgct gcgggccgag gccgagagca
2281 gcaggcccgg cagcgccggc gtgaagctcc ggcggcggtg gcagcggcgg ctggggcggt
2341 ggcgcgtcgg ggacacgtgg gcgctggacc cgcgcgcgag gtgggtccgg gactggaacc
2401 gcgcctacct gctggcctgc gcggcggggc tgatggtgga cccgctcttc ctgtacgcgg
2461 tgtcggtgag cggcccgctg atgtgcgtct tcctcgacgg ctggctcgcc gccgcggtca
2521 ccgcgctccg ctgcatggtg gacgccatgc acgcctggaa cctcctcacg cagctccgcg
2581 tcgcgcgcgc ggcggccgcg gcggtggcgc gcggccggcg gggggtcgcg gacgaggagc
2641 aggccgaggc cgacgccgcg gccgcgcgaa gcctgccggc gtacgccagg tccaggagag
2701 ggatggcgct cgacttcttc gtcatcctcc ccgtgatgca ggtgcacata tgccccctt
```

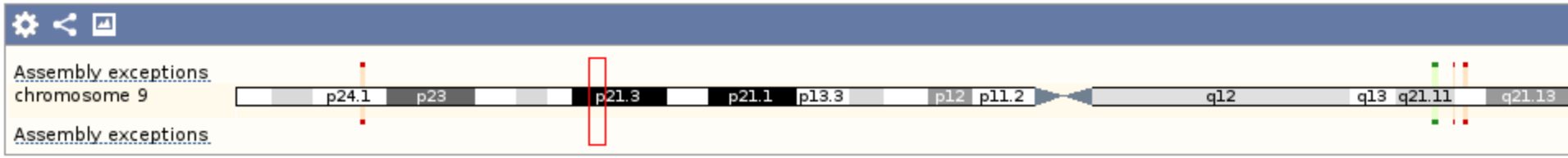
Eikariotu gēnu paredzēšana

- Eikariotu genomi ir lieli un kompleksi
- Gēni ir lieli, sadalīti intronos un eksonos, pastāv alternatīvais splaisings
- Intronu sekvences evolucionē ātrāk par eksonu sekvencēm, tādēļ tās var stipri atšķirties pat starp evolucionāri tuviem organismiem
- Homoloģija ar citiem gēniem ļauj atrast tikai jau zināmus gēnus
- Eksperimentāla gēnu struktūras paredzēšana salīdzinot genomisko DNS ar cDNS
- Gēnu paredzēšana izmantojot bioinformātikas metodes

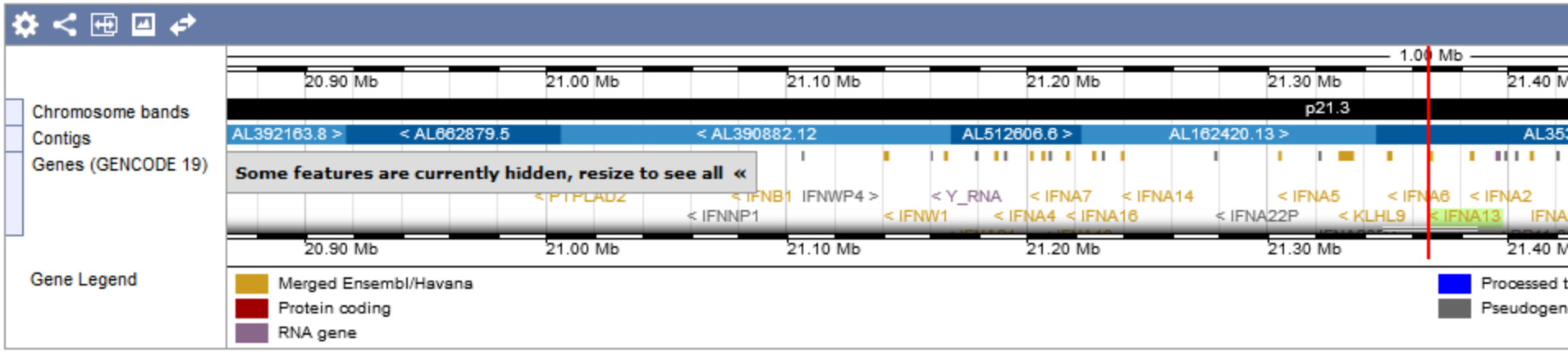
Eikariotu genomu komponenti

- **Unikālas sekvences** - vienas kopijas gēni, unikālas regulatorās secības
- **Dažādo atkārtojumu anotācija ir ļoti svarīga, lai saprastu eikariotu genoma struktūru**
- **Bieži atkārtojumi** – minisatelīti, mikrosatelīti, telomēru atkārtojumi

Chromosome 9: 21,367,371-21,368,075



Region in detail



http://www.ensembl.org/index.html

Bioloģiskās informācijas datubāzes. Informācijas meklēšanas un iegūšanas sistēmas

Bioloģiskā informācija

- Molekulārās sekvences (nukleotīdu un aminoskābju)
- Proteīnu struktūras
- Gēnu ekspresijas dati
- Literatūra, kas saistīta ar bioloģisko informāciju

Bioinformātikas datu bāzes

- Datubāze – organizēta informācijas kolekcija datoram nolasāmā formātā
- Dažādi datu bāzu veidi

  Nukleotīdu, aminoskābju

  Sekvenču, struktūras, literatūras, gēnu ekspresijas, molekulāro mijiedarbību, noteiktu organismu datu bāzes
- Informācijas meklēšana un iegūšana no dažādām datu bāzēm
- Internets un pieeja datu bāzēm

Arhīvs?

- Informācijas uzglabāšana ir tikai viena no datu bāzu funkcijām
- Datu bāze ir informācijas organizēšanas līdzeklis, informācijas meklēšanas efektivitāte datu bāzē ir atkarīga no tā, cik labi informācija ir organizēta
- Datu bāzes ir saistītas savā starpā un ļauj iegūt vispusīgu informāciju par meklēto jautājumu
- Datu bāzes piedāvā integrētus instrumentus, kas lietotājam ļauj patstāvīgi veikt datu analīzi

Kas kopīgs datu bāzēm?

- Dati ir sakārtoti tabulās, sadalīti pa grupām
- Katram datu bāzu ierakstam ir savs unikāls identifikators (ieraksta numurs – *Accession number* un GI)
- Dažādiem datu laukiem ir noteikts identifikators
- Datu bāzu ieraksti ir noteiktā standartizētā formātā, kas atvieglo to uzglabāšanu, meklēšanu un atrastās informācijas tālāku apstrādi

H.sapiens gene for interferon-alpha 13

GenBank: X75934.1

FASTA Graphics

Go to:

```
LOCUS       X75934                  1539 bp    DNA     linear   PRI 18-APR-2005
DEFINITION  H.sapiens gene for interferon-alpha 13.
ACCESSION   X75934
VERSION     X75934.1  GI:439666
KEYWORDS    interferon alpha.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordat
            Mammalia; Eutheria; Euarcho
            Catarrhini; Hominidae; Homo
REFERENCE   1
  AUTHORS   Henco,K., Brosius,J., Fujis
            Hochstadt,J., Kovacic,T., P
            Todokoro,K., Waelchli,M., N
  TITLE     Structural relationship of
            pseudogenes
  JOURNAL   J. Mol. Biol. 185 (2), 227-
   PUBMED   4057246
  REMARK    (sites)
REFERENCE   2  (bases 1 to 1539)
  AUTHORS   Rostoks,N.
  JOURNAL   Unpublished
REFERENCE   3  (bases 1 to 1539)
  AUTHORS   Rostoks,N.
  TITLE     Direct Submission
  JOURNAL   Submitted (21-DEC-1993) Nil
            of Genetics of Plants and M
            Latvia LV-1842
FEATURES             Location/Qualifier
     source          1..1539
                     /organism="Homo sa
                     /mol_type="genomic
                     /db_xref="taxon:96
     TATA_signal     855..860
                     /citation=[1]
     mRNA            887..>1539
                     /citation=[1]
     CDS             956..1525
                     /citation=[1]
                     /codon_start=1
                     /product="interferon-alpha 13"
                     /protein_id="CAA53538.1"
                     /db_xref="GI:439667"
                     /db_xref="GDB:136353"
                     /db_xref="GDB:136355"
                     /db_xref="GOA:P01562"
                     /db_xref="HGNC:5417"
                     /db_xref="HGNC:5419"
                     /db_xref="InterPro:IPR000471"
                     /db_xref="InterPro:IPR009079"
                     /db_xref="InterPro:IPR012351"
                     /db_xref="InterPro:IPR015589"
                     /db_xref="UniProtKB/Swiss-Prot:P01562"
                     /translation="MASPFALLMALVVLSCKSSCSLGCDLPETHSLDNRRTLMLLAQM
                     SRISPSSCLMDRHDFGFPQEEFDGNQFQKAPAISVLHELIQQIFNLFTTKDSSAAWDE
                     DLLDKFCTELYQQLNDLEACVMQEERVGETPLMNADSILAVKKYFRRITLYLTEKKYS
                     PCAWEVVRAEIMRSLSLSTNLQERLRRKE"
     sig_peptide     956..1024
                     /citation=[1]
     mat_peptide     1025..1522
                     /product="unnamed"
                     /citation=[1]
     variation       984
ORIGIN
        1 agatctcaaa gtcatatcat gagaggtgtc ctctgcatac atatggtttg tcactggcca
       61 tcttatagat attgcttatg tttgatcctt agcatttctg tctgtgttgt gggctttgaa
      121 atgaaatata aataatttat attttaacaa ttctactgaa gttgttcaac acatctatat
      181 ttaggtcaag aattgaagaa aaaattcttc acactctcta ggtgagtatc tagcagctag
```

Integrēta pieeja datu bāzēm

- Zinot cik daudz un dažādu datu tipu pastāv (nukleotīdu un aminoskābju sekvences, literatūra, gēnu ekspresijas dati u.t.t.), būtu pamats bažām, kā atrast sakarības dažādu datu veidu starpā
- NCBI un EMBL piedāvā integrētas meklēšanas sistēmas – NCBI Entrez un EMBL SRS
- Abas meklēšanas sistēmas piedāvā gan teksta meklēšanu, gan molekulāro sekvenču meklēšanu izmantojot BLAST vai FASTA programmas

Būla operatori un ierobežotāji

- ***Boolean operators*** – AND, OR un NOT

  http://www.ncbi.nlm.nih.gov/books/NBK3837/#EntrezHelp.Entrez_Searching_Options

- ***Field*** – [Author], [Organism], [Journal], [Gene Name], bet noteikti *Field* attiecas uz noteiktām datu bāzēm

Meklēšanas sistēma paveic lielu darbu mūsu vietā, bet meklēšanas teksts vienmēr jāformulē pēc iespējas precīzi

Piemēram: «human BRCA1» (3846 rezultāti nukleotīdu datu bāzē, bet

«Homo sapiens[Organism] AND BRCA1[Gene Name]» (147 rezultāti nukleotīdu datu bāzē)

Primārās datu bāzes

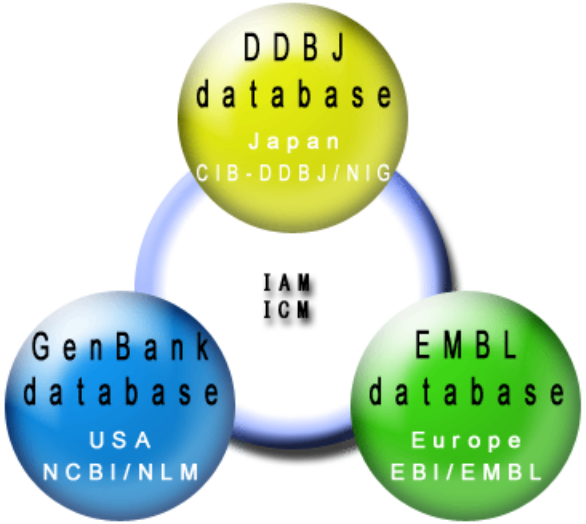
- National Center for Biotechnology Information (NCBI) GenBank

  http://www.ncbi.nlm.nih.gov
- DNA Data Bank of Japan

  http://www.ddbj.nig.ac.jp/
- The European Molecular Biology Laboratories (EMBL) Nucleotide Archive http://www.ebi.ac.uk/ena/
- Primārās informācijas saturs visās 3 DB ir viens un tas pats, jo datu apmaiņa notiek katru dienu
- Sekvences identifikators (*accession number*) ir viens un tas pats visās DB

International Nucleotide Sequence Database Collaboration

- http://www.insdc.org/
- INSDC ir sadarbības rezultāts starp GenBank, DDBJ, EMBL



DDBJ/EMBL/GenBank
International
Nucleotide Sequence Database

DDBJ: DNA Data Bank of Japan
CIB-DDBJ: Center for Information Biology and DNA Data Bank of Japan
NIG: National Institute of Genetics

EMBL: European Molecular Biology Laboratory
EBI: European Bioinformatics Institute

NCBI: National Center for Biotechnology Information
NLM: National Library of Medicine

IAM: International Advisory Meeting
ICM: International Collaborative Meeting

GenBank

- GenBank ir DNS sekvenču datu bāze
- GenBank datu bāze izveidota vēl pirms NCBI, kas ir struktūrvienība, kas šobrīd nodrošina šīs datu bāzes uzturēšanu
- Benson et al. (2013) GenBank. Nucleic Acids Res 41: D36-42

GenBank is a comprehensive public database of nucleotide sequences and supporting bibliographic and biological annotation

Genomu evolūcija un salīdzinošā genomika

Sugu un genomu evolūcija

- Visas uz Zemes sastopamās organismu grupas ir cēlušās no viena kopēja senča, bet pēc tam evolucionējušas atsevišķi (izņēmums – horizontālā gēnu pārnese un organellu simbioze)
- DNS un proteīnu līdzības pamatā ir "*Identity by descent*", vai arī konverģentā evolūcija
- Ja gēni dažādās sugās ir līdzīgi pēc DNS vai aminoskābju secības, tad iespējams tie veic līdzīgu funkciju
- Līdzīgas DNS un aminoskābju secības sauc par **homologām** (ar kopīgu izcelsmi)

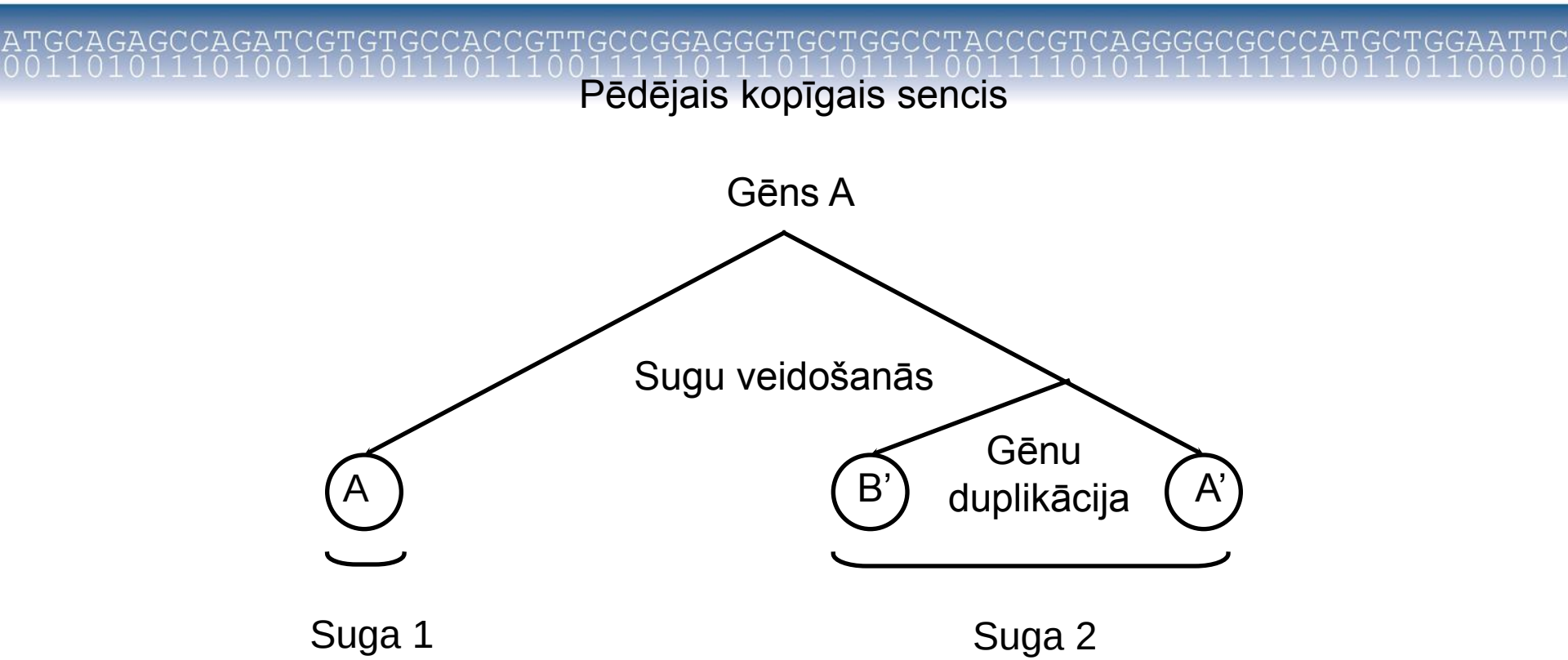
Salīdzinošā genomika funkcionāli nozīmīgu elementu identifikācijai

- Pēc būtības visspēcīgākā metode genomu anotācijai – ja DNS secības ir saglabājušas homoloģiju pēc miljoniem gadu ilgas neatkarīgas evolūcijas, tad tām droši vien ir funkcionāla nozīme

Dzīvības koks

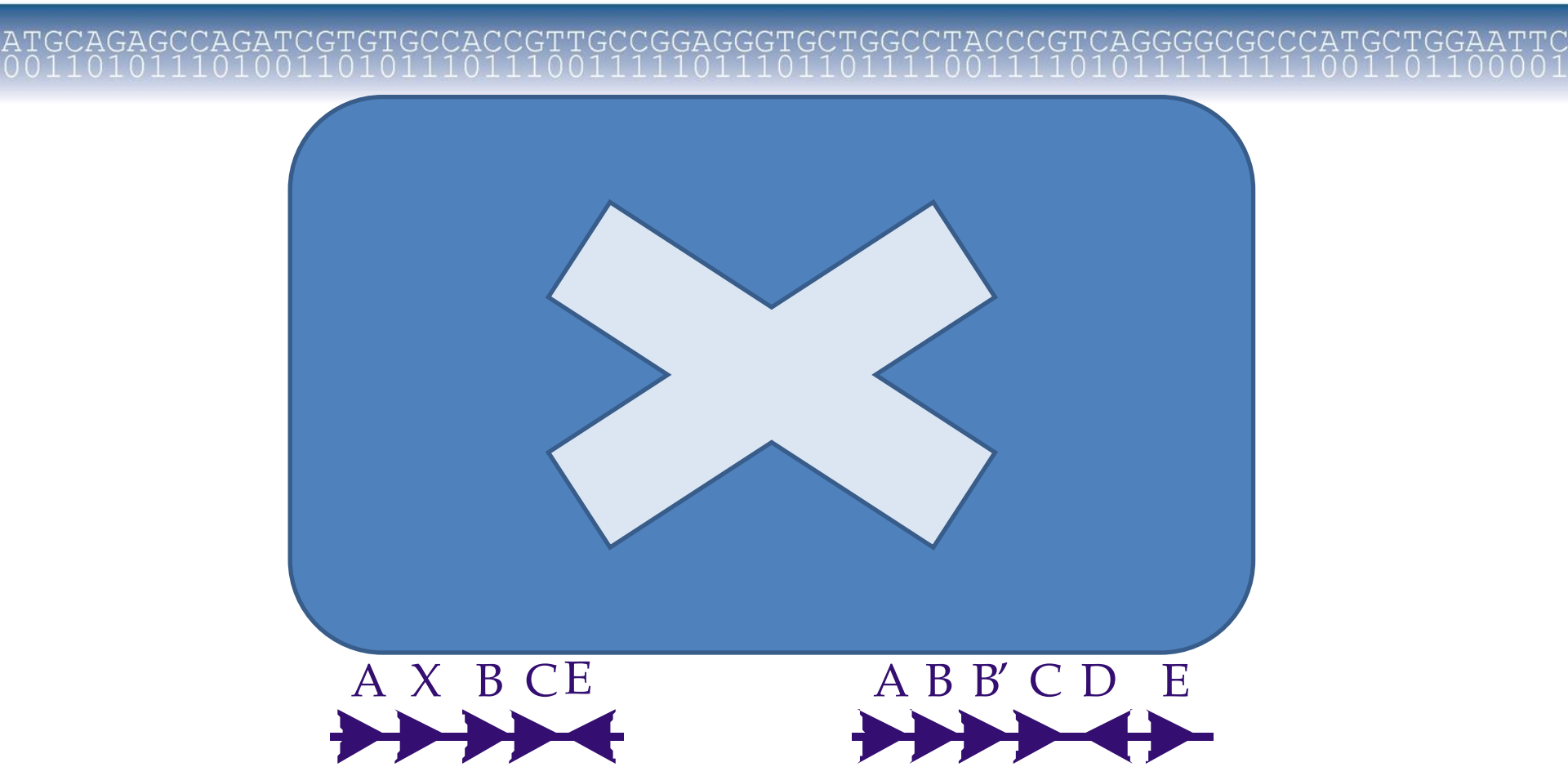
http://www.tolweb.org/tree/

Ortologi un paralogi

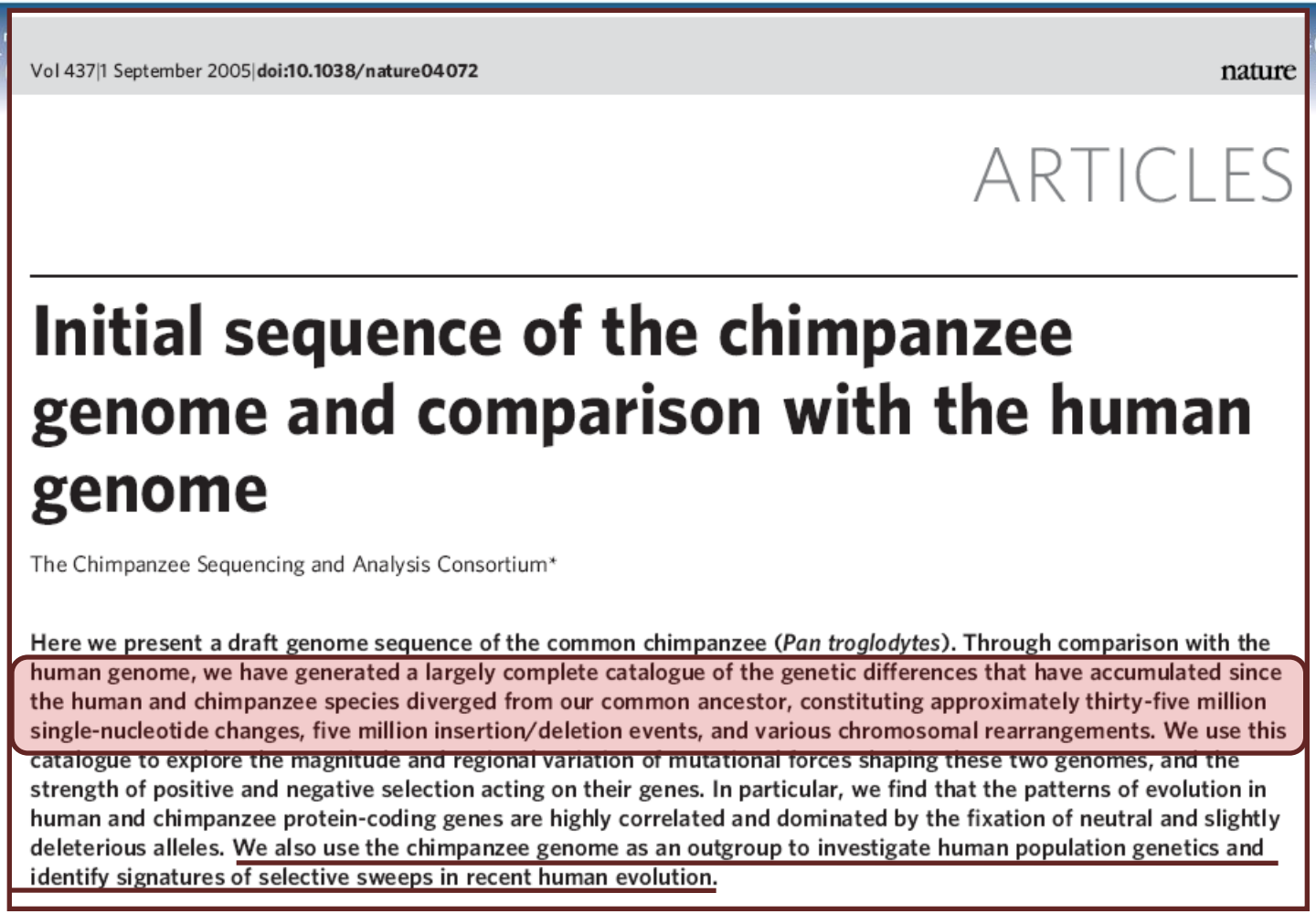


A un A’ ir ortologi. A’ un B’ ir paralogi. Pēc gēnu duplikācijas viena no gēna kopijām var turpināt pildīt iepriekšējo funkciju, kamēr otra gēna kopija mutāciju rezultātā var izmainīties un iegūt jaunu funkciju.

Gēnu kārtības saglabāšanās evolūcijas gaitā (sintēnija)



Cilvēka un šimpanzes genomu salīdzinājums



Vol 437|1 September 2005|doi:10.1038/nature04072 nature

ARTICLES

Initial sequence of the chimpanzee genome and comparison with the human genome

The Chimpanzee Sequencing and Analysis Consortium*

Here we present a draft genome sequence of the common chimpanzee (*Pan troglodytes*). Through comparison with the human genome, we have generated a largely complete catalogue of the genetic differences that have accumulated since the human and chimpanzee species diverged from our common ancestor, constituting approximately thirty-five million single-nucleotide changes, five million insertion/deletion events, and various chromosomal rearrangements. We use this catalogue to explore the magnitude and regional variation of mutational forces shaping these two genomes, and the strength of positive and negative selection acting on their genes. In particular, we find that the patterns of evolution in human and chimpanzee protein-coding genes are highly correlated and dominated by the fixation of neutral and slightly deleterious alleles. We also use the chimpanzee genome as an outgroup to investigate human population genetics and identify signatures of selective sweeps in recent human evolution.

cilvēks – šimpanze
cilvēks – pele

- Genomu salīdzinājumi ļauj atbildēt uz fundamentāli atšķirīgiem jautājumiem
- **Cilvēks – pele**. Iespējams identificēt gēnus un kontroles elementus, kas saglabājušies nemainīgi evolūcijas gaitā. Var pētīt proteīnu funkcionālos domēnus, kas saglabājušies nemainīgi, vai arī gēnus, kas atšķir abus organismus.
- **Cilvēks – šimpanze**. Genomi, tai skaitā nekodējošā DNS, ļoti līdzīgi, kas ļauj identificēt mutācijas, kas radušās tieši cilvēka evolūcijas gaitā. Iespējams identificēt mutācijas gēnos un kodējošās daļās, kas potenciāli atšķir cilvēku no šimpanzes.

ENCODE project

- ENCyclopedia of DNA Elements (2003)
- Mērķis – identificēt visus funkcionālos elementus cilvēka genoma sekvencē
- Vairāk nekā 80% genoma sekvences noteikta sekvence, atklāti vairāk nekā 4 miljoni regulācijas rajonu
- Cilvēka genoma funkcionālo elementu analīze ar salīdzinošās genomikas palīdzību
- http://www.genome.gov/10005107 un http://www.nature.com/encode

Genomu salīdzinājums no dažādām cilvēka populācijām

ATGCAGAGCCAGATCGTGTGCCACCGTTGCCGGAGGGTGCTGGCCTACCCGTCAGGGGCGCCCATGCTGGAATTC

Sciencexpress Report

An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia

Morten Rasmussen,[1,2]* Xiaosen Guo,[2,3]* Yong Wang,[4]* Kirk E. Lohmueller,[4]* Simon Rasmussen,[5] Anders Albrechtsen,[6] Line Skotte,[6] Stinus Lindgreen,[1,6] Mait Metspalu,[7] Thibaut Jombart,[8] Toomas Kivisild,[9] Weiwei Zhai,[10] Anders Eriksson,[11] Andrea Manica,[11] Ludovic Orlando,[1] Francisco De La Vega,[12] Silvana Tridico,[13] Ene Metspalu,[7] Kasper Nielsen,[5] María C. Ávila-Arcos,[1] J. Víctor Moreno-Mayar,[1,14] Craig Muller,[15] Joe Dortch,[16] M. Thomas P. Gilbert,[1,2] Ole Lund,[5] Agata Wesolowska,[5] Monika Karmin,[7] Lucy A. Weinert,[8] Bo Wang,[3] Jun Li,[3] Shuaishuai Tai,[3] Fei Xiao,[3] Tsunehiko Hanihara,[17] George van Driem,[18] Aashish R. Jha,[19] François-Xavier Ricaut,[20] Peter de Knijff,[21] Andrea B Mig[illegible] Kristiansen,[2,3,6] David M. Lambert,[23] Søren Brunak,[5,24] Peter Forster,[25,26] Be[illegible] Michael Richards,[27,28] Ramneek Gupta,[5] Carlos D. Bustamante,[12] Anders K[illegible] Balloux,[8] Thomas Sicheritz-Pontén,[5,29] Richard Villems,[7,30] Rasmus Nielse[illegible]

[1]Centre for GeoGenetics, Natural History Museum of Denmark, and Depart[illegible] Voldgade 5-7, 1350 Copenhagen, Denmark. [2]Sino-Danish Genomics Cente[illegible] University of Copenhagen, Denmark. [3]Shenzhen Key Laboratory of Transo[illegible]

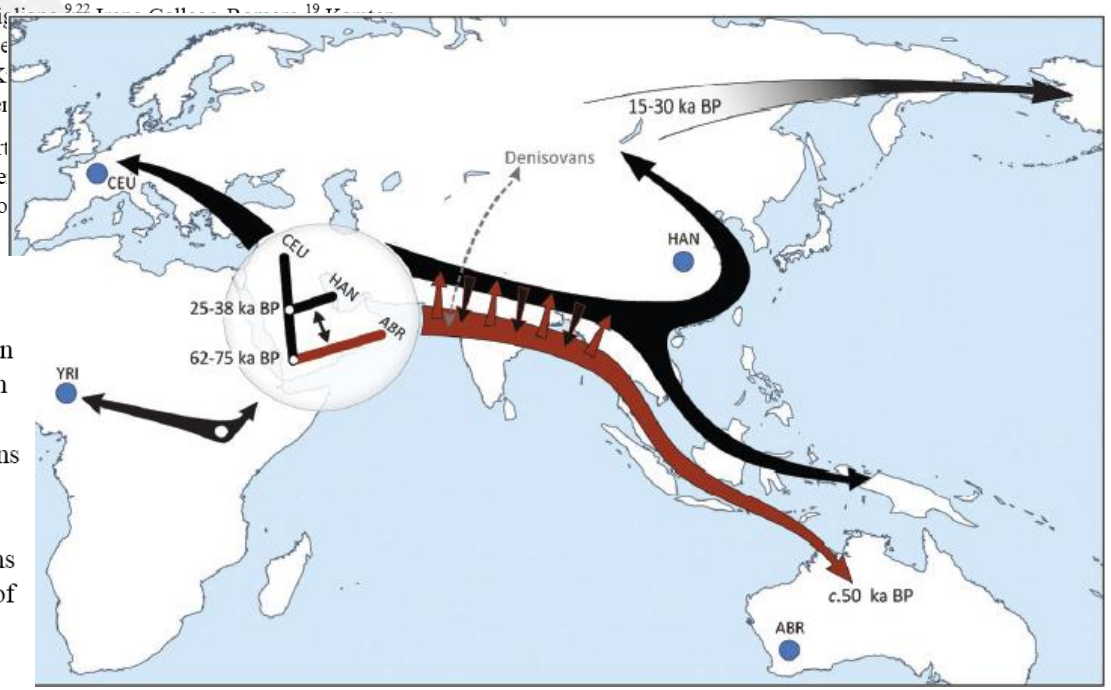


Fig. 2. Reconstruction of early spread of modern humans outside Africa. The tree shows the divergence of the Aboriginal Australian (ABR) relative to the CEPH European (CEU) and the Han Chinese (HAN) with gene flow between aboriginal Australasians and Asian ancestors. Purple arrow shows early spread of the ancestors of Aboriginal Australians into eastern Asia ~62,000 to 75,000 years B.P. (ka BP), exchanging genes with Denisovans, and reaching Australia ~50,000 years B.P. Black arrow shows spread of East Asians ~25,000 to 38,000 years B.P. and admixing with remnants of the early dispersal (red arrow) some time before the split between Asians and Native American ancestors ~15,000 to 30,000 years B.P. YRI, Yoruba.

Rasmussen et al. (2011) Science 10.1126/science.1211177

Nukleīnskābju un proteīnu sekvenču līdzības pamatprincipi

Dažādas salīdzināšanas metodes, to priekšrocības un pielietošanas nosacījumi

Sekvenču salīdzināšana

- Angliski – *sequence alignment*
- Divi galvenie salīdzinājumu veidi – *pairwise sequence alignment* un *multiple sequence alignment*
- Latviskais tulkojums varētu būt “sekvenču pāru salīdzinājums” un “daudzkārtējs sekvenču salīdzinājums”

Sekvenču salīdzināšana

- Ja mums ir divas vai vairākas DNS vai aminoskābju secības, tās analizējot vajadzētu:

  - izmērīt to vispārējo līdzību;

  - noteikt to līdzību katrā sekvences pozīcijā;

  - novērot līdzīgo un atšķirīgo rajonu izvietojumu sekvencēs;

  - novērtēt sekvenču evolucionārās attiecības

Sekvenču salīdzinājumu pielietojumi

- Genomu anotācijas pamatojas uz salīdzinājumu ar jau raksturotām sekvencēm (gēnu un atkārtojumu identifikācija)
- Funkcionālo domēnu noteikšana proteīnu sekvencēs
- DNS un proteīnu sekvenču radniecības pakāpes noteikšana, lai izvērtētu to filoģenētiskās attiecības un rekonstruētu šo secību evolūciju

Salīdzinājuma pamatprincipi

- Sekvences salīdzina, nosakot to līdzību katrā sekvences pozīcijā
- Lai veidotu optimālu salīdzinājumu, iespējams sekvencēs ieviest pārtraukumus

```
abcdef
acdefg
```

→

```
abcdef-
a-cdefg
```

- Nepieciešams izveidot kritērijus pēc kuriem spriest par salīdzinājuma kvalitāti

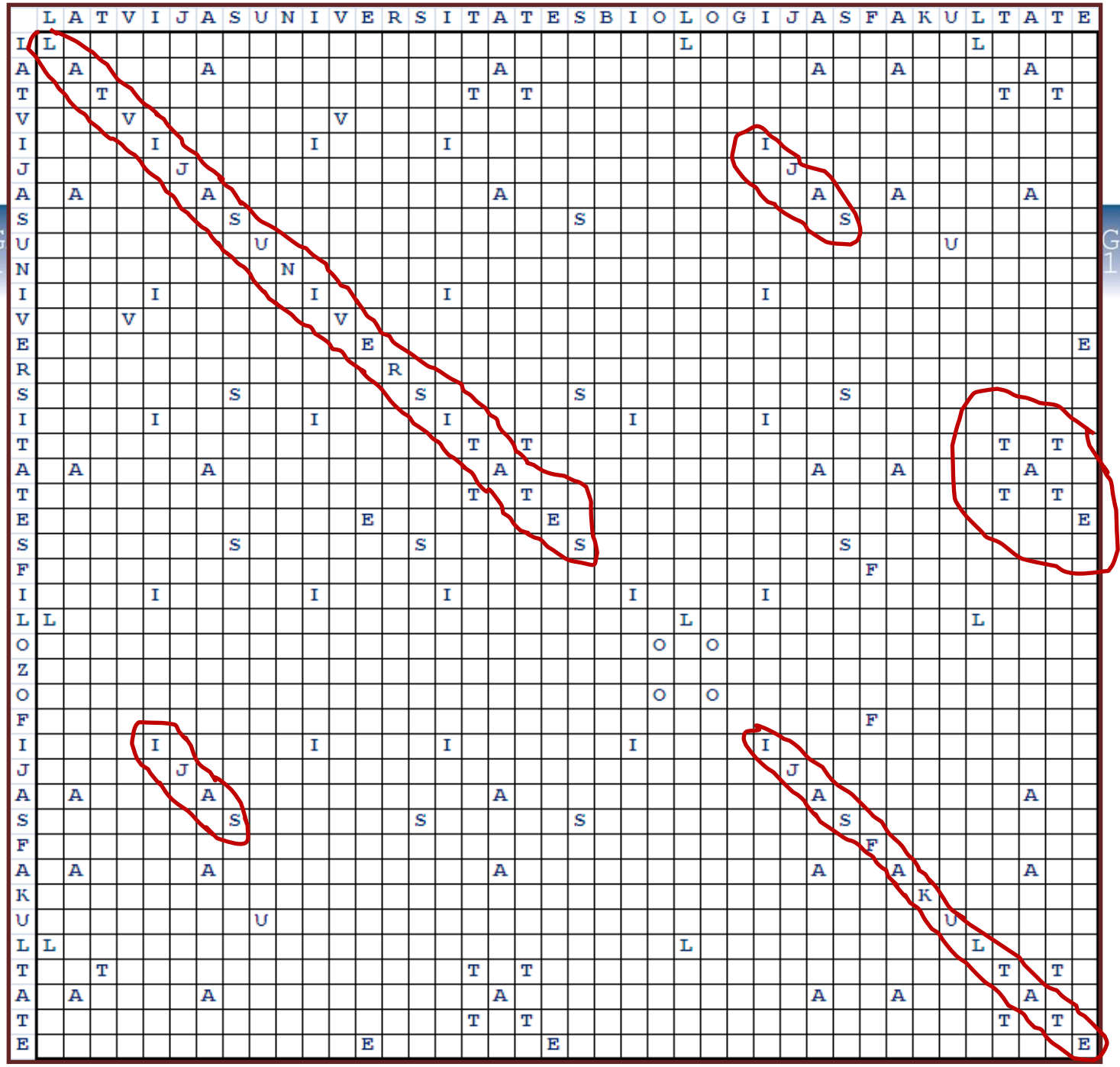
```
acgtctga
agtttgat
```

```
acgtctga-
agttt-gat
```

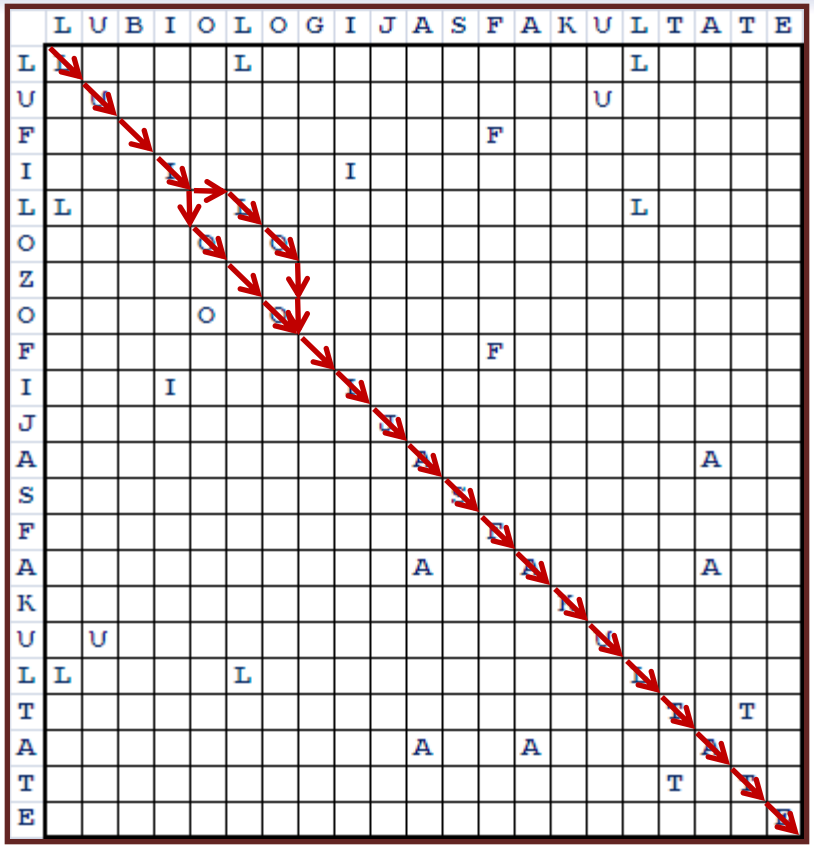
```
acgtctga-
a-gtttgat
```

Dot plots

- Visvienkāršākais divu sekvenču salīdzināšanas veids
- Dot plots ir kā tabula vai matrica, kurā rindas atbilst vienai sekvencei, bet kolonnas atbilst otrai sekvencei
- Vienkāršākajā veidā tiek atzīmēti tikai burti (nukleotīdi vai aminoskābes), kas katrā pozīcijā starp sekvencēm ir identiski


LATVIJASUNIVERSITATESBIOLOGIJASFAKULTATE

Dot plots un sekvenču salīdzinājums (*alignment*)



Salīdzinājumu kvalitāte

L	U	B	I	-	O	L	O	G	I	J	A	S	F	A	K	U	L	T	A	T	E
L	U	F	I	L	O	Z	O	F	I	J	A	S	F	A	K	U	L	T	A	T	E
*	*		*		*		*		*	*	*	*	*	*	*	*	*	*	*	*	*

L	U	B	I	O	L	O	-	-	G	I	J	A	S	F	A	K	U	L	T	A	T	E
L	U	F	I	-	L	O	Z	O	F	I	J	A	S	F	A	K	U	L	T	A	T	E
*	*		*		*	*				*	*	*	*	*	*	*	*	*	*	*	*	*

Ceļš, kas tiek noiets dot plotā, atbilst sekvenču salīdzinājumam. Ja sekvences ir samērā līdzīgas, dot plots ļauj rekonstruēt sekvenču salīdzinājumu. Pastāv iespēja, ka vairāki ceļi dot plotā būs vienādi varbūtīgi. Nedrīkst pieņemt, ka ceļs dot plotā atbilst sekvenču molekulārajai evolūcijai

Divu sekvenču salīdzinājums

- Kad izveidota salīdzinājumu novērtēšanas sistēma, iespējams meklēt **optimālu salīdzinājumu** – tādu, kas dod vislielāko punktu skaitu
- Ir situācijas, kad pastāv vairāki optimālie salīdzinājumi
- Globāls salīdzinājums – tiek salīdzinātas divas pilna garuma sekvences
- Lokāls salīdzinājums – salīdzina vienas sekvences segmentu ar otras sekvences segmentu
- Globāls salīdzinājums var būt ārkārtīgi laikietilpīgs (ja sekvenču garums ir *n* un *m*, tad matricas izmērs ir *n* x *m*)

NCBI BLAST

- ***Basic Local Alignment Search Tool***
- Līdzīgi kā globālajā sekvenču salīdzinājumā tiek izmantots dotplots, taču salīdzināšana notiek nevis ar pilna garuma sekvenci, bet gan tās noteikta garuma fragmentiem
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. J. Mol. Biol. 215:403-410
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402

BLAST mehānisms

- Visam pamatā – dotplots
- BLAST sadala meklēšanā izmantoto sekvenci noteikta garuma vārdos (piemēram, $k = 4$), tad atrod visas sekvences datu bāzē, kuras satur precīzi tādus vārdus
- BLAST cenšas pagarināt katru no sakrītošajiem vārdiem uz abām pusēm, nepieļaujot atšķirības un pārtraukumus (*mismatch*, *gap*)
- Pēc tam pagarinātie rajoni tiek apvienoti, pieļaujot atšķirības un pārtraukumus

BLAST

- BLASTN – nukleotīdu secība pret nukleotīdu datu bāzi
- BLASTP – proteīnu aminoskābju secība pret proteīnu datu bāzi
- BLASTX – translēta nukleotīdu secība pret proteīnu datu bāzi
- TBLASTN – aminoskābju secība pret translētu nukleotīdu datu bāzi
- TBLASTX – translēta nukleotīdu secība pret translētu nukleotīdu datu bāzi

Daudzkārtēja sekvenču salīdzināšana

- *Multiple sequence alignment* (MSA) – vairāku nukleotīdu vai aminoskābju secību salīdzinājums
- Viena no biežāk lietotajām bioinformātikas metodēm
- Plaši izmanto dažādu proteīnu domēnu meklēšanā, proteīnu struktūras modelēšanā un filoģenētiskajā analīzē

>ref|XP_003428362.1| PREDICTE
[Ornithorhynchus anatinus]
Length=704

GENE ID: 100074825 LOC100074825 | pa
[Ornithorhynchus anatinus]

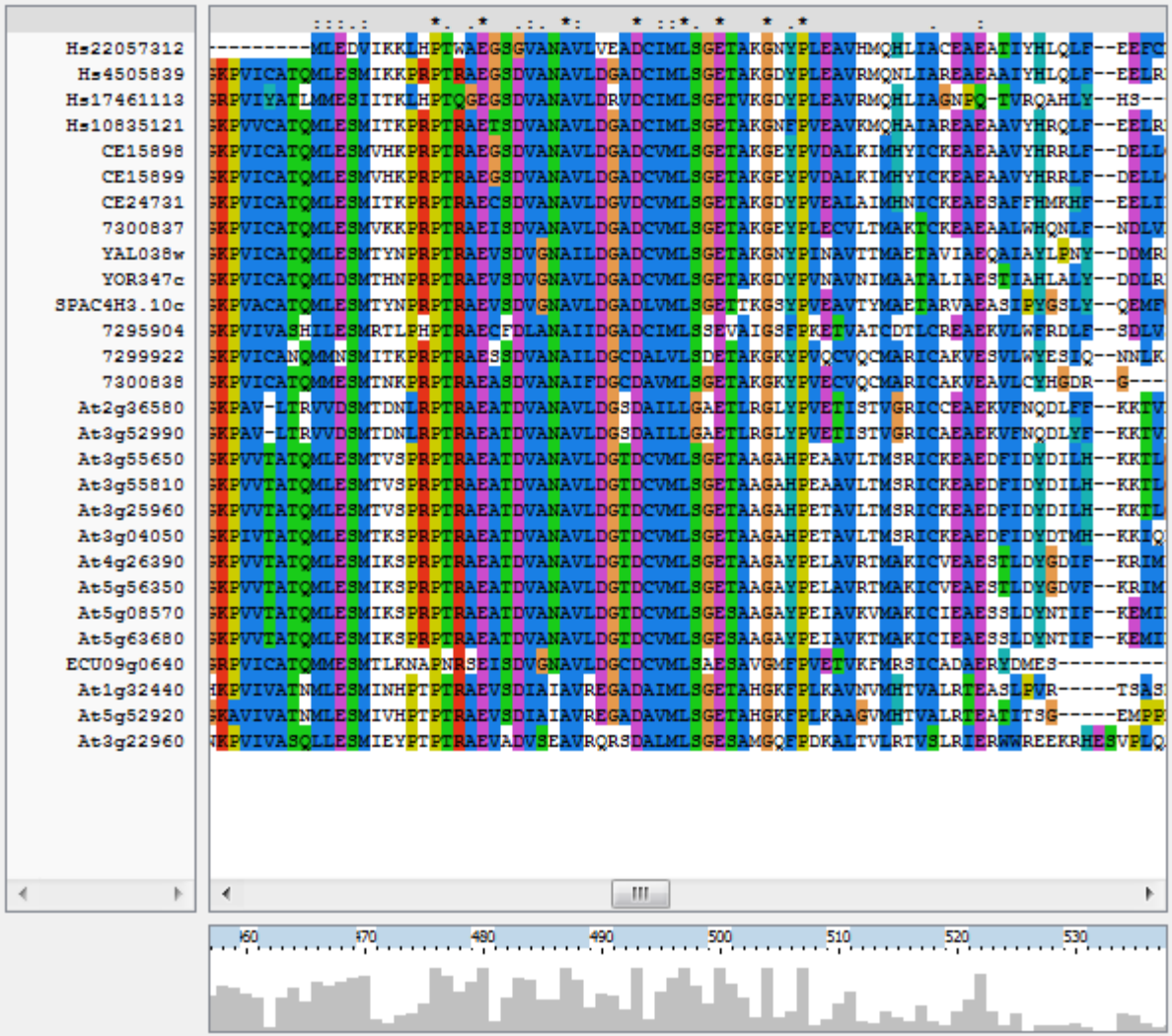
Score = 644 bits (1662), Expect =
Identities = 334/394 (85%), Positive

```
Query  48   VSNGCVSKILGRYYETGSIRPRAIG
            VSNGCVSKILGRYYETGSIRPRAIG
Sbjct  315  VSNGCVSKILGRYYETGSIRPRAIG

Query  108  SEGVCTNDNIPSVSSINRVLRNLAS
            SEGVCTNDNIPSVSSINR  R
Sbjct  375  SEGVCTNDNIPSVSSINRGTR--GE

Query  150  N-GQTGSWGTRPGWYPGTSVPGQPT
               G  G  G R   +     P  P
Sbjct  433  ERGMEGCPGPRSKVWIDCPAPIIP-

Query  209  KLQRNRTSFTQEQIEALEKEFERTH
            KLQRNRTSFTQEQIEALEKEFERTH
Sbjct  491  KLQRNRTSFTQEQIEALEKEFERTH

Query  269  EKLRNQRRQASNTPSHIPISSSFST
            EKLRNQRRQAS TPSHIPISSSFST
Sbjct  551  EKLRNQRRQASTTPSHIPISSSFST

Query  329  LPPMPSFTMANNLPMQPPVPSQTSS
            LPPMPSFTMANNLPMQPPVPSQTSS
Sbjct  611  LPPMPSFTMANNLPMQPPVPSQTSS

Query  389  SGTTSTGLISPGVSVPVQVPGSEPD
            SGTTSTGLISPGVSVPVQVPGSEPD
Sbjct  671  SGTTSTGLISPGVSVPVQVPGSEPD
```

ClustalX 2.1

File Edit Alignment Trees Colors Quality Help

Mode: Multiple Alignment Mode Font: 10



CLUSTAL-Alignment file created [C:/Users/Nils/Documents/LU_Dokumenti/Kursi/Bioinformatika/leva

Daudzkārtēju sekvenču salīdzinājumu veidošana

- Tāpat kā divu sekvenču salīdzināšanas gadījumā tiek meklēta līdzība starp nukleotīdiem vai aminoskābēm katrā salīdzinājuma pozīcijā
- Daudzkārtējā sekvenču salīdzinājumā var izdalīt sekojošus etapus:

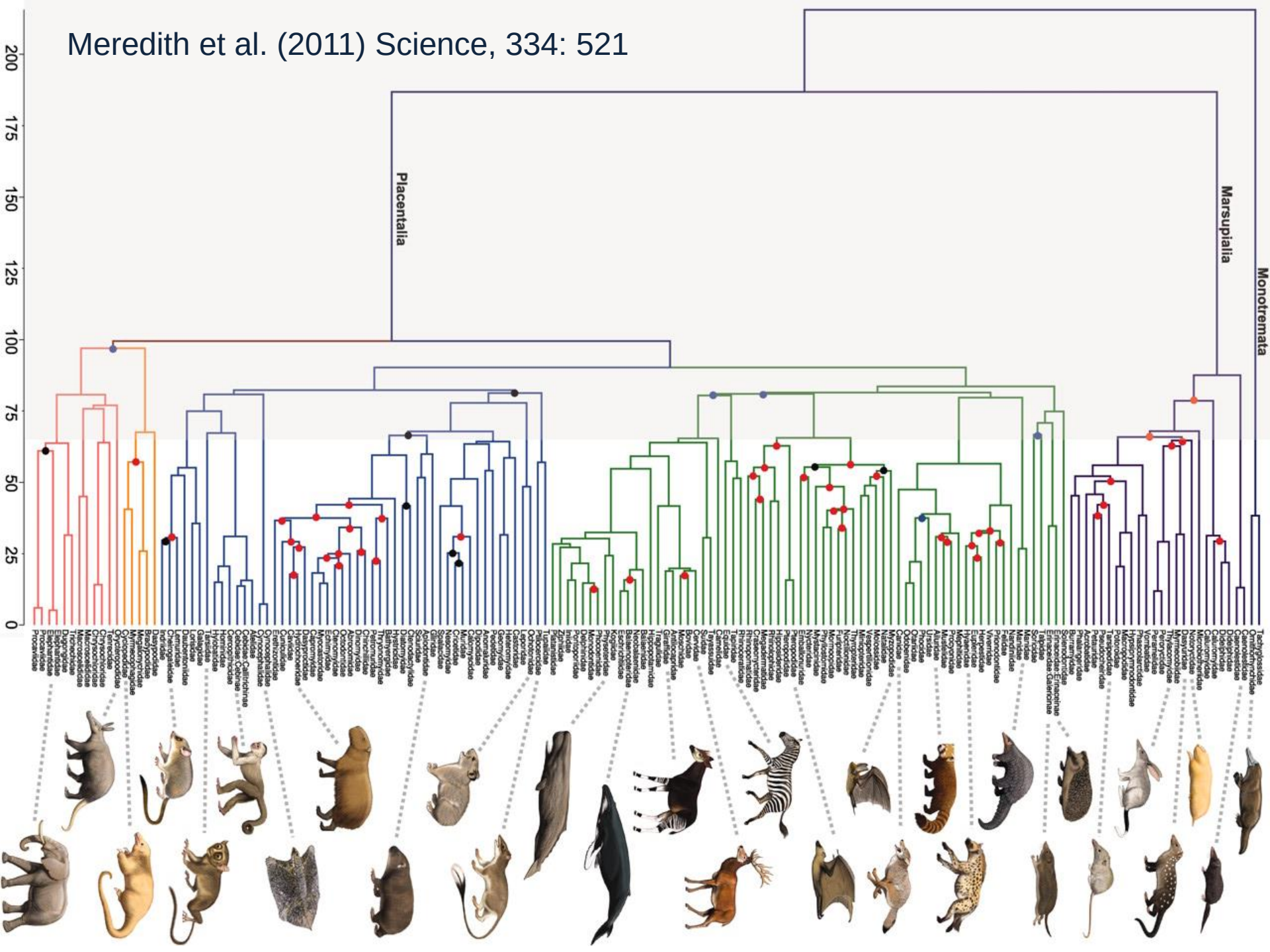
  1. salīdzina savā starpā visus sekvenču pārus un izveido attālumu matricu;

  2. aprēķina dendrogrammu, kas balstīta uz attālumu matricas;

  3. Vienu pēc otras salīdzinājumam pievieno sekvences atbilstoši to secībai dendrogrammā sākot ar līdzīgākajām sekvencēm

Clustal X un Clustal W

- Viena no populārākajām salīdzināšanas metodēm
- Oriģinālā publikācija - Higgins and Sharp (1988) **CLUSTAL**: a package for performing multiple sequence alignment on a microcomputer. Gene 73,237-244
- Thompson et al. (1994) **CLUSTAL W**: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Research, 22:4673-4680
- Thompson et al. (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research, 24:4876-4882
- http://www.clustal.org/clustal2/


Meredith et al. (2011) Science, 334: 521
Monotremata
Marsupialia
Placentalia
200
175
150
125
100
75
50
25
0

Evolūcija

"Nothing in Biology Makes Sense Except in the Light of Evolution“

Theodosius Dobzhansky (1973)

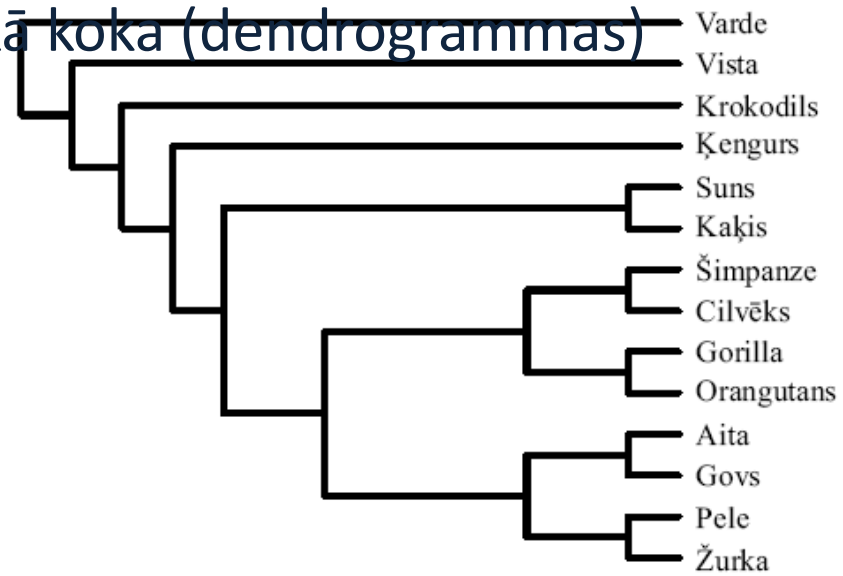
http://www.2think.org/dobzhansky.shtml

Filoģenēze

- Filoģenēze (*Phylogeny*) – sugas (vai citas taksonomiskas vienības) evolucionārā vēsture

  Filoģenētiskās analīzes pamatā ir ideja, ka visi dzīvie organismi ir cēlušies no kopīga senča

  Evolucionārās attiecības starp taksonomiskajām vienībām parasti tiek attēlotas filoģenētiskā koka (dendrogrammas) veidā



Molekulārā filoģenēze

- Evolūcijas kokus var veidot balstoties uz dažādām pazīmēm – morfoloģiskām, fizioloģiskām u.t.t.
- Molekulārā filoģenēze tātad ir sugu (taksonomisko vienību) evolucionāro attiecību pētīšana izmantojot molekulārās sekvences

Terminoloģija

- Homoloģija (*Homology*)– līdzība, ko nosaka kopīga izcelšanās. Apgalvojums par kādu objektu homoloģiju nozīmē, ka tiek pieņemta to evolucionāra saistība
- Līdzība (*Similarity*) – objektu līdzības vai atšķirības mērs, neiedziļinoties to varbūtējā saistībā
- Klasteri – līdzīgu objektu grupas, kas tiek sakārtotas pēc līdzības pakāpes.
- Hierarhiskie klasteri – klasteri, kas veido klasterus, kas veido klasterus...

Sugu un gēnu dendrogrammas

- Parastākās dendrogrammas attēlo dažādu taksonomisko vienību evolūciju – sugu dendrogrammas
- Molekulārie dati (DNS un proteīnu sekvences) ļauj noteikt atsevišķu gēnu evolūciju
- Gēnu un sugu dendrogrammas ne vienmēr sakrīt

Dendrogrammas

- Dendrogrammas topoloģija – norāda taksonomisko vienību radniecību
- Zaru garumi – norāda evolucionāro attālumu vai laiku kopš taksonomiskās vienības nodalījās.
- Dendrogrammas mezgli (*node*) – norāda kopējo priekšteci
- Molekulārais pulkstenis – atkarīgs no mutāciju ātruma. To kalibrē izmantojot fosīliju datus