Al-Farabi Kazakh National University
Higher School of Medicine
Department of Fundamental Medicine

Bioinformatics

Lecturer and creator: PhD Pinsky Ilya Vladimirovich

LEARNING OUTCOMES
As a result of the lesson you will be able to:

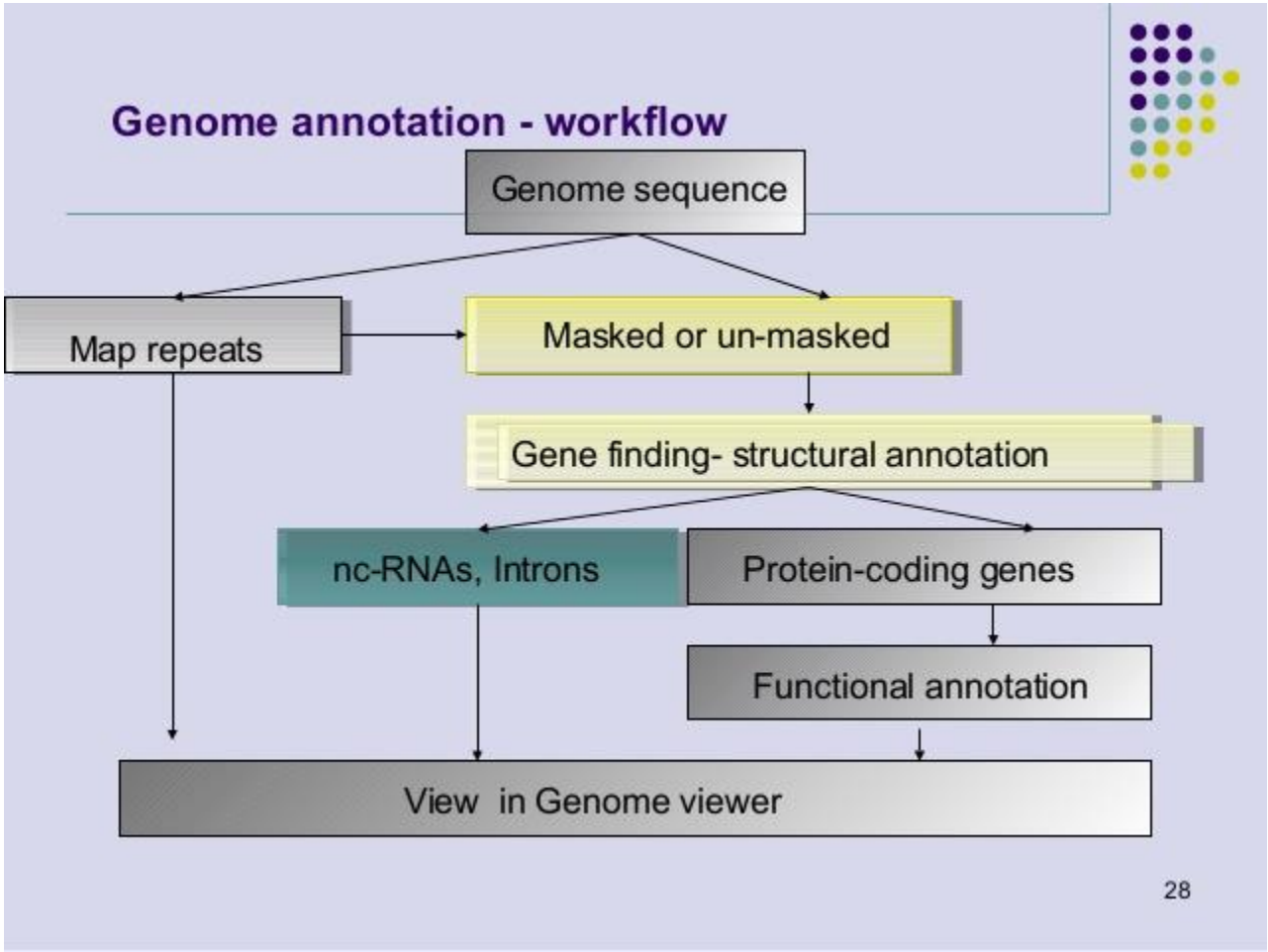
- 1. Explain the terms “bioinformatics”, “computational biology” and “system biology”.
- 2. Analyze the bioinformatical methods used in different “Omics” technologies, give the specific examples.
- 3. Explain the differences between structural, functional and evolutional bioinformatics.
- 4. Give and describe the examples of bioinformatical computer programs used for different tasks.
- 5. Classify and describe the main bioinformatical databases, give the specific examples.

Definitions

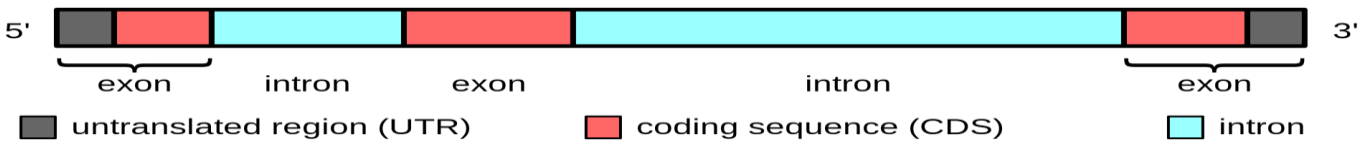
Bioinformatics (/ˌbaɪ.oʊˌɪnfərˈmætɪks/) is an interdisciplinary field that develops methods and software tools for understanding biological data, in particular when the data sets are large and complex. As an interdisciplinary field of science, bioinformatics combines **biology, computer science, information engineering, mathematics and statistics** to analyze and interpret the biological data. Bioinformatics has been used for *in silico* analyses of biological queries using mathematical and statistical techniques.

Bioinformatics includes biological studies that use computer programming as part of their methodology, as well as a specific analysis "pipelines" that are repeatedly used, particularly in the field of genomics. Common uses of bioinformatics include the identification of **candidates genes and single nucleotide polymorphisms (SNPs)**. Often, such identification is made with the aim of better understanding the genetic basis of disease, unique adaptations, desirable properties (esp. in agricultural species), or differences between populations. In a less formal way, bioinformatics also tries to understand the organizational principles within nucleic acid and protein sequences, called **proteomics**.[1]

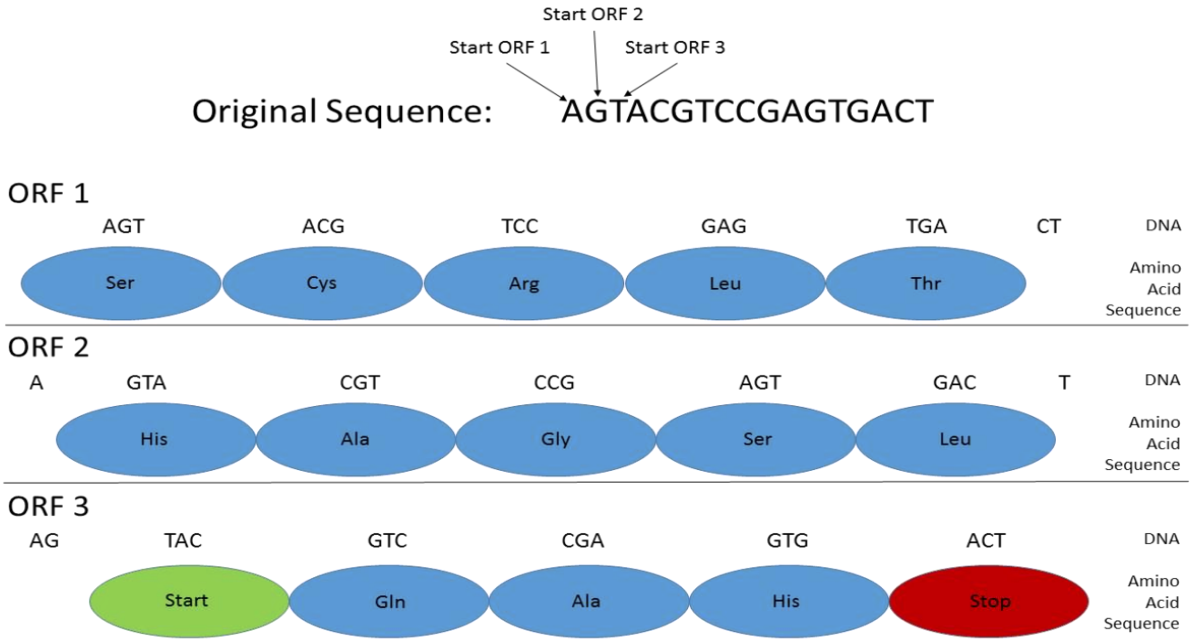
- **Systems biology** is the **computational and mathematical analysis and modeling of complex biological systems**. It is a biology-based interdisciplinary field of study that focuses on complex interactions within biological systems, using a **holistic approach** (holism instead of the more traditional **reductionism**) to biological research.
- **Computational biology** involves the development and application of data-analytical and theoretical methods, mathematical modelling and computational simulation techniques to the study of biological, ecological, behavioural, and social systems. The field is broadly defined and includes foundations in biology, applied mathematics, statistics, biochemistry, chemistry, biophysics, molecular biology, genetics, genomics, computer science, and evolution.
- Computational biology is different from **biological computing**, which is a subfield of computer engineering, using bioengineering and biology to build computers. [2-7]



Genome Annotation | CityU-Bioinformatics
cityu-bioinformatics.netlify.app



https://en.wikipedia.org/wiki/Gene_prediction#/media/File:Gene_structure.svg



https://en.wikipedia.org/wiki/Gene_prediction#/media/File:Gene_Prediction.png

GeneMark

A family of gene prediction programs developed at
Georgia Institute of Technology , Atlanta, Georgia, USA.

What's New:
Publication on GeneMark-EP+

Gene Prediction in Bacteria, Archaea, Metagenomes and Metatranscriptomes

Novel genomic sequences can be analyzed either by the self-training program **GeneMarkS** (sequences longer than 50 kb) or by **GeneMark.hmm with Heuristic models**. For many species pre-trained model parameters are ready and available through the **GeneMark.hmm** page. Metagenomic sequences can be analyzed by **MetaGeneMark** , the program optimized for speed.

Gene Prediction in Eukaryotes

Novel genomes can be analyzed by **GeneMark-ES**, an algorithm utilizing models parameterized by unsupervised training. Notably, GeneMark-ES has a special option for fungal genomes to account for fungal-specific intron organization. To integrate into GeneMark-ES information on mapped RNA-Seq reads, we made semi-supervised GeneMark-ET. Recently, we have developed **GeneMark-EP+** that uses homologous protein sequences of any evolutionary distance in both training and predictions.

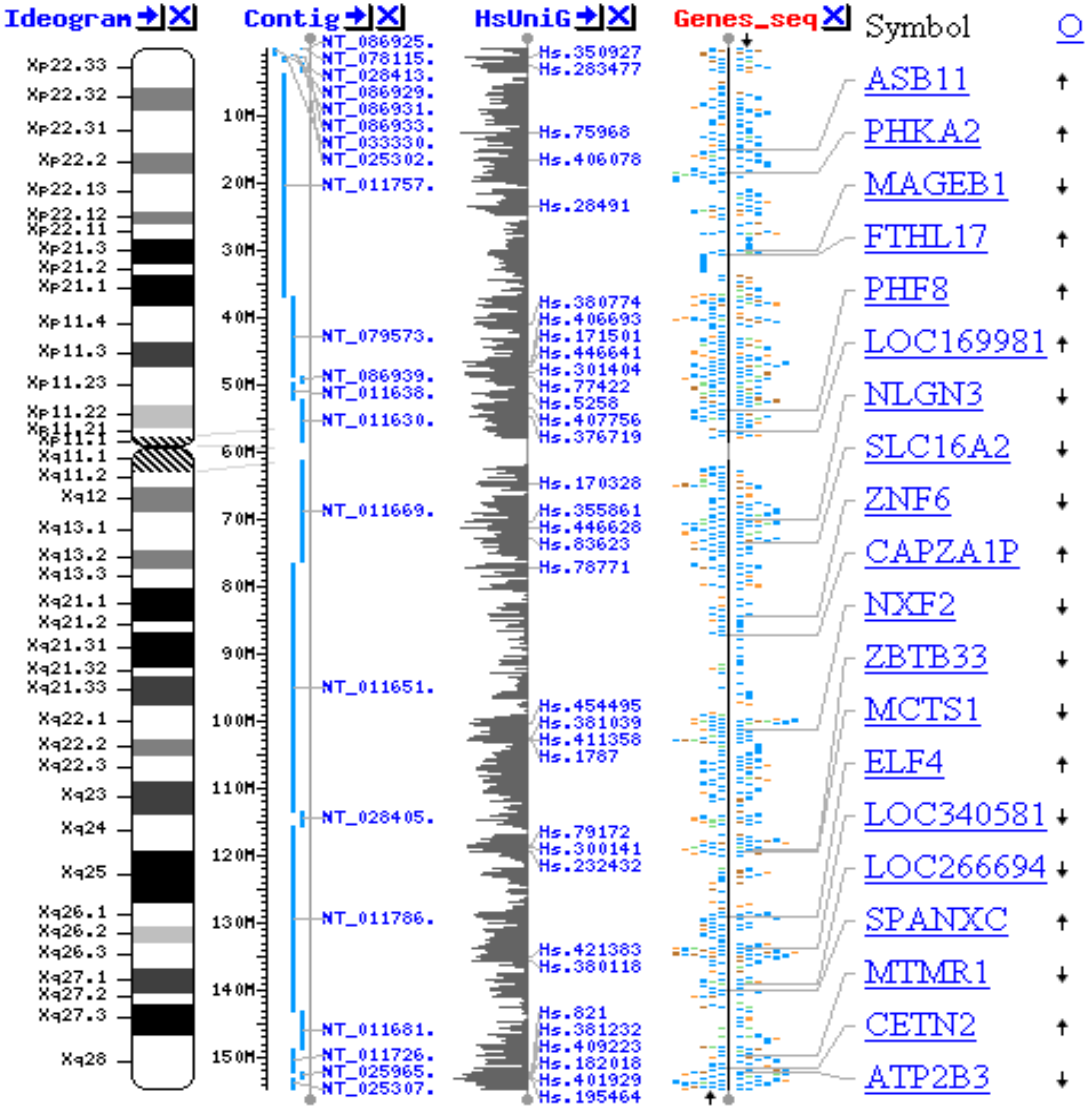
Gene Prediction in Transcripts

Sets of assembled eukaryotic transcripts can be analyzed by the modified **GeneMarkS** algorithm (the set should be large enough to permit self-training). A single transcript can be analyzed by a special version of **GeneMark.hmm with Heuristic models**. A new advanced algorithm GeneMarkS-T was developed recently (manuscript sent to publisher); The GeneMarkS-T software (beta version) is available for download .

Gene Prediction in Viruses, Phages and Plasmids

Sequences of viruses, phages or plasmids can be analyzed either by the **GeneMark.hmm with Heuristic models** (if the sequence is shorter than 50 kb) or by the self-training program **GeneMarkS**.

All the software programs mentioned here are available for download and local installation



Map of the human X chromosome (from the National Center for Biotechnology Information website).
https://en.wikipedia.org/wiki/Bioinformatics#/media/File:Genome_viewer_screenshot_small.png

```
A5ASC3.1   14 SIKLWPPSQTTRLLLVERMANNLST..PSIFTRK..YGSLSKEEARENAKQIEEVACSTANQ.....HYEKEPDGDGGSAVQLYAKECSKLILEVLK 101
B4F917.1   13 SIKLWPPSESTRIMLVDRMTNNLST..ESIFSRK..YRLLGKQEAHENAKTIEELCFALADE.....HFREEPDGDGSSAVQLYAKETSKMMLEVLK 100
A9S1V2.1   23 VFKLWPPSQGTREAVRQKMALKLSS..ACFESQS..FARIELADAQEHARAIEEVAFGAAQE......ADSGGDKTGSAVVMVYAKHASKLMLETLR 109
B9GSN7.1   13 SVKLWPPGQSTRLMLVERMTKNFIT..PSFISRK..YGLLSKEEAEEDAKKIEEVAFAAANQ.....HYEKQPDGDGSSAVQIYAKESSRLMLEVLK 100
Q8H056.1   30 SFSIWPPTQRTRDAVVRRLVDTLGG..DTILCKR..YGAVPAADAEPAARGIEAEAFDAAAA..SGEAAATASVEEGIKALQLYSKEVSRRLLDFVK 120
Q0D4Z3.2   44 SLSIWPPSQRTRDAVVRRLVQTLVA..PSILSQR..YGAVPEAEAGRAAAAVEAEAYAAVTES.SSAAAAPASVEDGIEVLQAYSKEVSRRLLELAK 135
B9MVW8.1   56 SFSIWPPTQRTRDAIISRLIETLST..TSVLSKR..YGTIPKEEASEASRRIEEEAFSGAST.......VASSEKDGLEVLQLYSKEISKRMLETVK 141
Q0IYC5.1   29 SFAVWPPTRRTRDAVVRRLVAVLSGDTTTALRKRYRYGAVPAADAERAARAVEAQAFDAASA....SSSSSSSVEDGIETLQLYSREVSNRLLAFVR 121
A9NW46.1   13 SIKLWPPSESTRLMLVERMTDNLSS..VSFFSRK..YGLLSKEEAAENAKRIEETAFLAAND.....HEAKEPNLDDSSVVQFYAREASKLMLEALK 100
Q9C500.1   57 SLRIWPPTQKTRDAVLNRLIETLST..ESILSKR..YGTLKSDDATTVAKLIEEEAYGVASN.......AVSSDDDGIKILELYSKEISKRMLESVK 142
Q2HRI7.1   25 NYSIWPPKQRTRDAVKNRLIETLST..PSVLTKR..YGTMSADEASAAAIQIEDEAFSVANA.......SSSTSNDNVTILEVYSKEISKRMIETVK 110
Q9M7N3.1   28 SFKIWPPTQRTREAVVRRLVETLTS..QSVLSKR..YGVIPEEDATSAARIIEEEAFSVASV.ASAASTGGRPEDEWIEVLHIYSQEIXQRVVESAK 119
Q9M7N6.1   25 SFSIWPPTQRTRDAVINRLIESLST..PSILSKR..YGTLPQDEASETARLIEEEAFAAAGS.......TASDADDGIEILQVYSKEISKRMIDTVK 110
Q9LE82.1   14 SVKMWPPSKSTRLMLVERMTKNITT..PSIFSRK..YGLLSVEEAEQDAKRIEDLAFATANK.....HFQNEPDGDGTSAVHVYAKESSKLMLDVIK 101
Q9M651.2   13 SIKLWPPSLPTRKALIERITNNFSS..KTIFTEK..YGSLTKDQATENAKRIEDIAFSTANQ.....QFEREPDGDGGSAVQLYAKECSKLILEVLK 100
B9R748.1   48 SLSIWPPTQRTRDAVITRLIETLSS..PSVLSKR..YGTISHDEAESAARRIEDEAFGVANT.......ATSAEDDGLEILQLYSKEISRRMLDTVK 133
```

https://en.wikipedia.org/wiki/Bioinformatics#/media/File:WPP_domain_alignment.PNG

BLAST: Basic Local Alignment Se

blast.ncbi.nlm.nih.gov/Blast.cgi

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

A new feature was added to Primer-BLAST.

We now offer the ability for user to run primer-blast f
NCBI assembly page..

Tue, 23 Feb 2021 12:00:00 EST

Web BLAST

Nucleotide BLAST
nucleotide ▶ nucleotide

blastx
translated nucleotide ▶ protein

blastx

tblastn
protein ▶ translated nucleotide

Protein BL
protein ▶ protein

BLAST Genomes

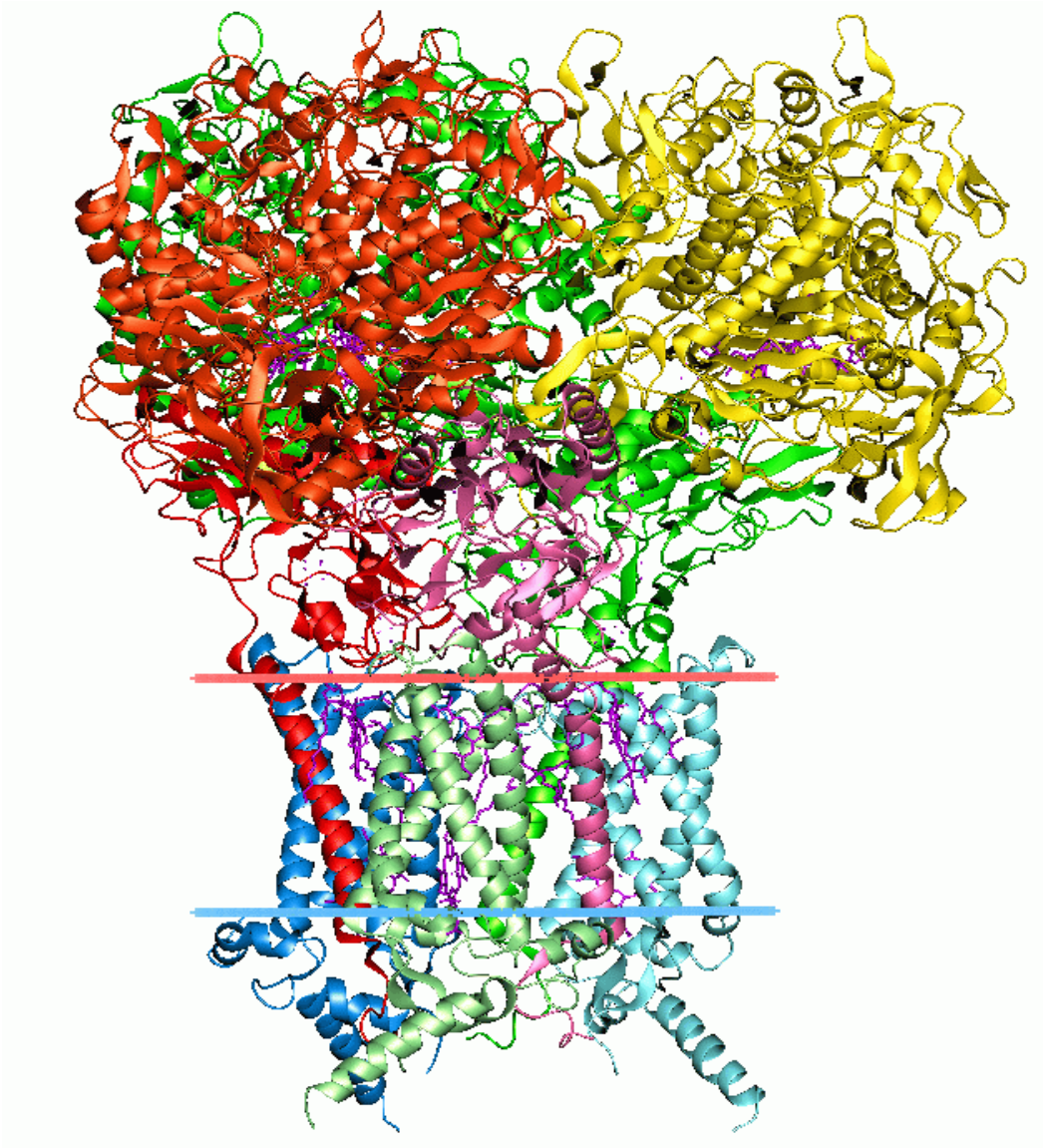
Enter organism common name, scientific name, or tax id **Search**

Human Mouse Rat Microbes

Standalone and API BLAST

https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastx&PAGE_TYPE=BlastSearch&LINK_LOC...

Введите здесь текст для поиска

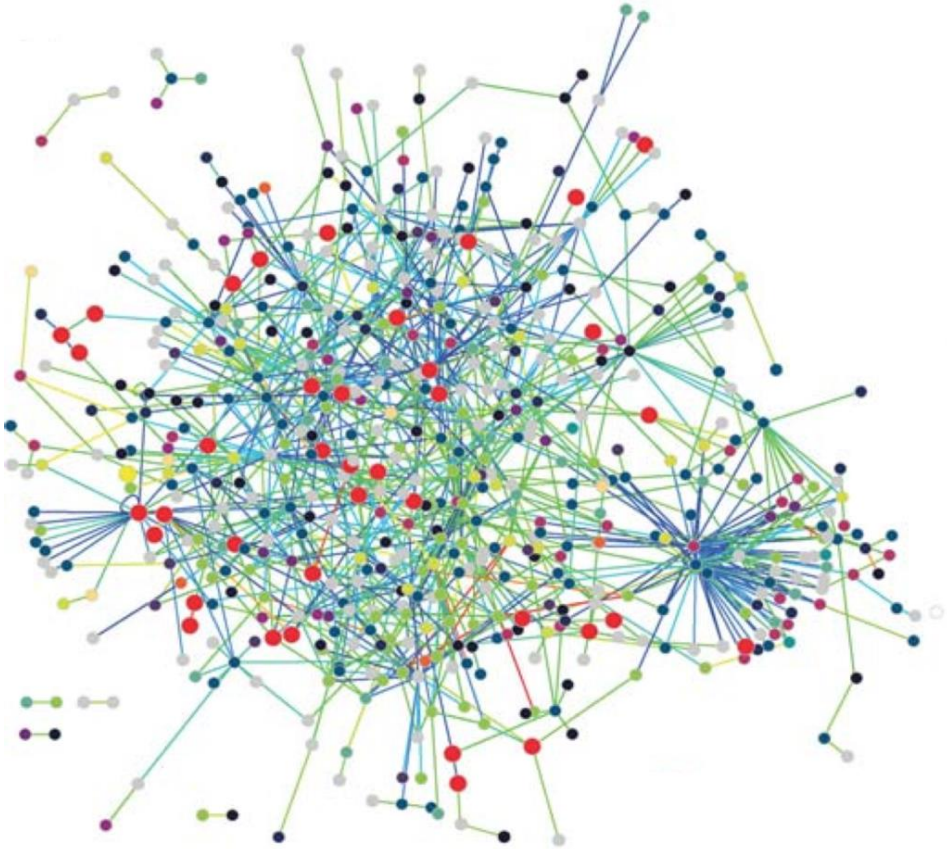


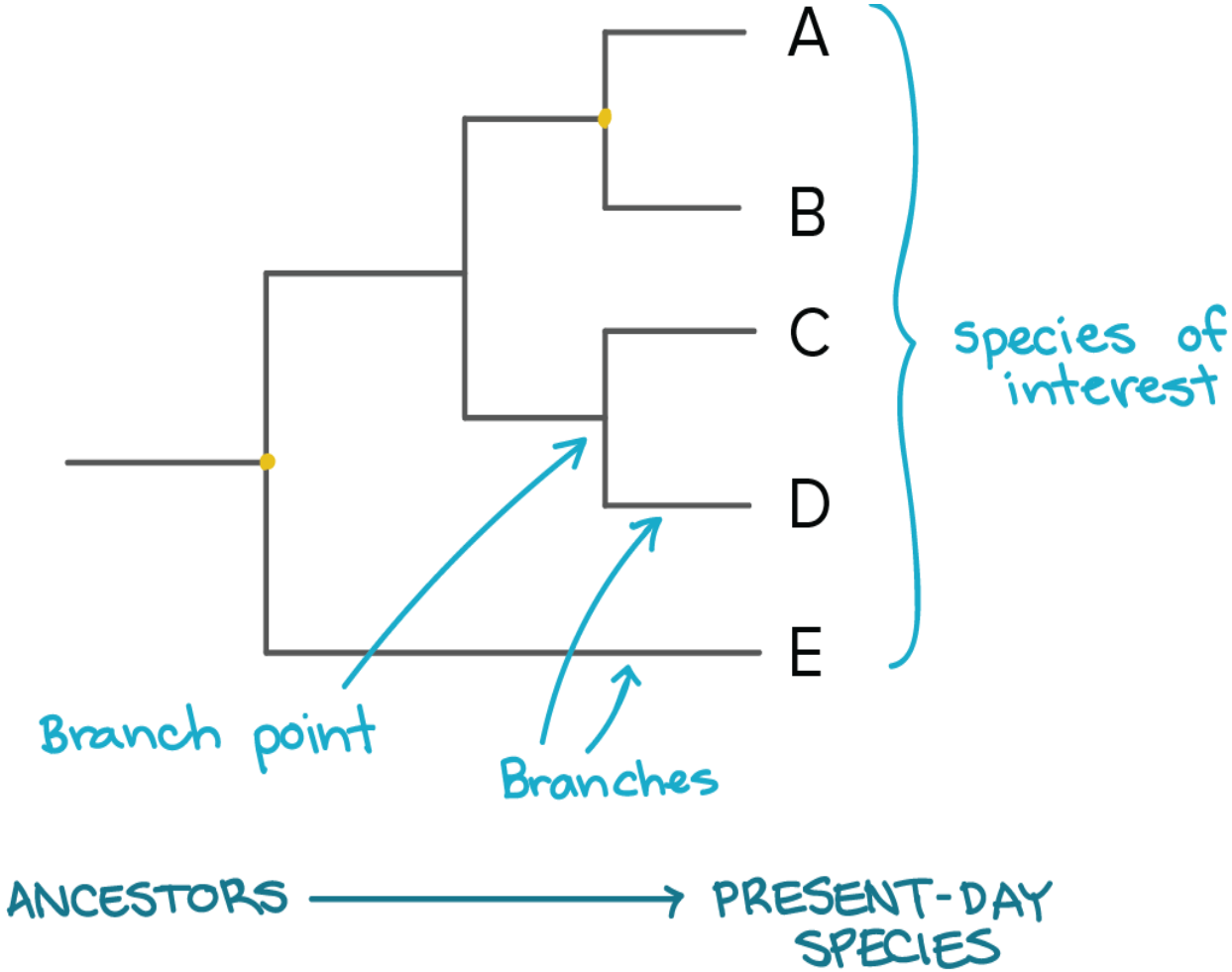
3-dimensional protein structures such as this one are common subjects in bioinformatic analyses.

https://en.wikipedia.org/wiki/Bioinformatics#/media/File:1kqf_opm.png

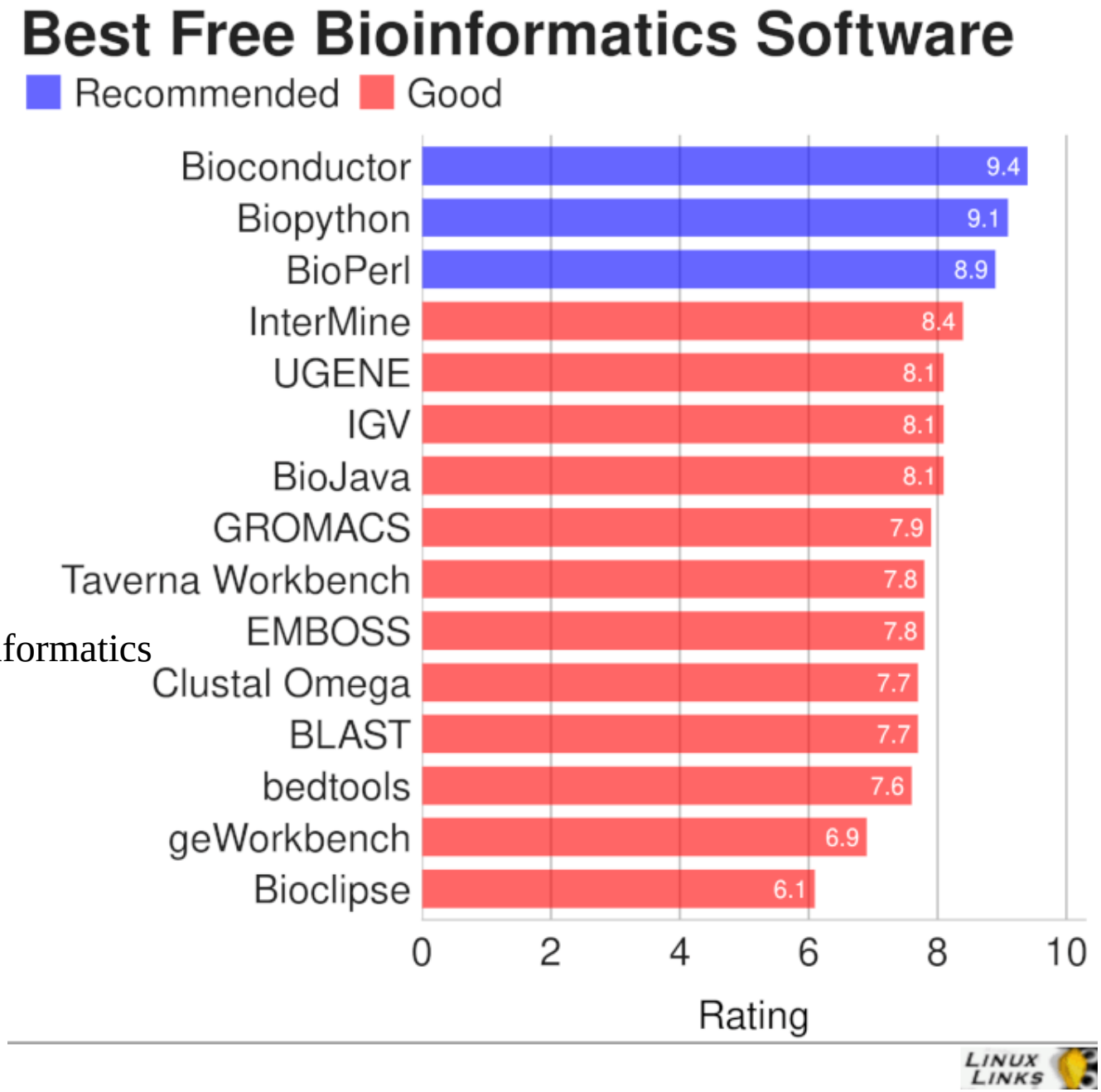
Häuser et al. - Titz B, Rajagopala SV, Goll J, Häuser R, McKevitt MT, et al. (2008) The Binary Protein Interactome of *Treponema pallidum* – The Syphilis Spirochete. PLoS ONE 3(5): e2292. doi:10.1371/journal.pone.0002292

The protein interaction network of *T. pallidum* including 576 proteins and 991 interactions. Nodes are color-coded according to TIGR main roles. Proteins involved in DNA metabolism are shown as enlarged red circles.





Phylogenetic trees | Evolutionary tree
khanacademy.org



15 Best Free Linux Bioinformatics Tools
linuxlinks.com

Apache Taverna - Bioinformatics
taverna.org.uk

Types of bioinformatic databases

- Archives (GeneBank & EMBL, PDB)
- Curated (Swiss-Prot, KEGG, FlyBase, COG)
- Derivatives (SCOP, PFAM , GO, ProDom, AsMamDB)
- Integrated (NCBI Entrez, Ecocyc)

Home - Gene - NCBI

ncbi.nlm.nih.gov/gene/

NCBI Resources How To

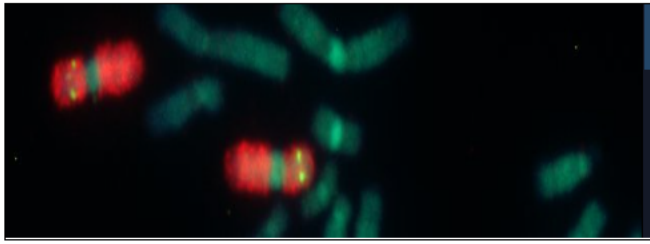
Gene

Gene

Advanced

COVID-19 is an emerging, rapidly evolving situation.

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS)



Gene

Gene integrates information from a wide range of species. A record may include nomenclatur
(RefSeqs), maps, pathways, variations, phenotypes, and links to genome-, phenotype-, and l
worldwide.

Using Gene

- Gene Quick Start
- FAQ
- Download/FTP
- RefSeq Mailing List
- Gene News
- Factsheet

Gene Tools

- Submit GeneRIFs
- Submit Correction
- Statistics
- BLAST
- Genome Workbench
- Splign

Other Resources

- OMIM
- RefSeq
- RefSeqGene
- Protein Clusters

Representative queries

Find genes by...	Search text
free text	human muscular dystrophy
chromosome and symbol	(II[chr] OR 2[chr]) AND adh*[sym]

Введите здесь текст для поиска

The European Bioinformatics Inst

ebi.ac.uk

EMBL-EBI | Services | Research | Training

EMBL-EBI

The home for big data in biology

We help scientists exploit complex information to make discoveries that benefit humankind.

Find tools and resources or deposit data.

Explore dozens of biological data resources with our Search servic

Find a gene, protein or chemical

Example searches: blast keratin bfl1 | Build query

We have been working behind the scenes to tackle the COVID-19 pandemic. Read more about our response and resources.

Featured topic

Latest news

Введите здесь текст для поиска

UniProt

uniprot.org

UniProt

UniProtKB

BLAST Align Retrieve/ID mapping Peptide search SPARQL

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional informatio

UniProtKB

UniProt Knowledgebase

Swiss-Prot (564,277)

Manually annotated and reviewed.

Records with information extracted from literature and curator-evaluated computational analysis.

TrEMBL (207,800,733)

Automatically annotated and not reviewed.

Records that await full manual annotation.

UniRef

The UniProt Reference Clusters (UniRef) provide clustered sets of sequences from the UniProt Knowledgebase (including isoforms) and selected UniParc records.

UniParc

UniParc is a comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world.

Proteomes

A proteome is the set of proteins thought to be expressed by an organism. UniProt provides proteomes for species with completely sequenced genomes.

Supporting data

Literature citations

Taxonomy

Subcellular locations

Cross-ref. databases

Diseases

Keywords

News

Forthcoming char
Planned changes

UniProt release 2
(Almost) all abou

UniProt release 2
Venoms, gold mir
cross-references

UniProt release 2

News archive

Getting started

Text search

Download latest release

Введите здесь текст для поиска

wwPDB: Worldwide Protein Data

Не защищено | wwpdb.org

WORLDWIDE PDB PROTEIN DATA BANK

VALIDATION DEPOSITION DICTIONARIES DOCUMENTATION TASK FORCES FTP STATISTICS

Since 1971, the Protein Data Bank archive (PDB) has served as the single repository of information about the 3D structures of proteins, nucleic acids, and complex assemblies.

The Worldwide PDB (wwPDB) organization manages the PDB archive and ensures that the PDB is freely and publicly available to the global community.

Celebrating 50 Years of the PDB

Validate Str

or View validat

Deposit Str

All Deposition

Download A

Instructions

Vision and Mission

Vision

Sustain freely accessible, interoperating Core Archives of structure data and metadata for biological macromolecules as an enduring public good to promote basic and applied research and education across the sciences.

Mission

- Manage the wwPDB Core Archives as a public good according to the FAIR Principles.
- Provide expert deposition, validation, biocuration, and remediation services at no charge to Data Depositors worldwide.
- Ensure universal open access to public domain structural biology data with no limitations on usage.
- Develop and promote community-endorsed data standards for archiving and exchange of global

wwPDB Resources

Data Dictionaries

- Macromolecular Dictionary (PDBx/mmCIF)
- Small Molecule Dictionary (CCD)
- Peptide-like antibiotic and inhibitor molecules (BIRD)

Biocuration

- Procedures and policies
- Improvements for consistency and accuracy

Community Input: Task Forces and Working Groups

- Validation Task Forces (X-ray, NMR, 3DEM)
- Small Angle Scattering Task Force
- PDBx/mmCIF Working Group
- Hybrid/Integrative Methods Task Force
- Ligand Validation Workshop

News & Ann

03/08/2021

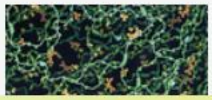
- Submit Abstracts for PD

Ea
Ma
PD
4-5

Read more

03/02/2021

- More than 1,000 SARS-C Protein Structures Availa



Op
pro

Введите здесь текст для поиска

References

1. Lesk, A. M. (26 July 2013). "Bioinformatics". Encyclopaedia Britannica.
2. Tavassoly, Iman; Goldfarb, Joseph; Iyengar, Ravi (2018-10-04). "Systems biology primer: the basic methods and approaches". Essays in Biochemistry. 62 (4): 487–500. doi:10.1042/EBC20180003. ISSN 0071-1365. PMID 30287586.
3. Zewail, Ahmed (2008). Physical Biology: From Atoms to Medicine. Imperial College Press. p. 339.
4. Longo, Giuseppe; Montévil, Maël (2014). Perspectives on Organisms - Springer. Lecture Notes in Morphogenesis. doi:10.1007/978-3-642-35938-5. ISBN 978-3-642-35937-8. S2CID 27653540.
5. Bu Z, Callaway DJ (2011). "Proteins MOVE! Protein dynamics and long-range allostery in cell signaling". Protein Structure and Diseases. Advances in Protein Chemistry and Structural Biology. 83. pp. 163–221. doi:10.1016/B978-0-12-381262-9.00005-7. ISBN 978-0-123-81262-9. PMID 21570668.
6. "NIH working definition of bioinformatics and computational biology" (PDF). Biomedical Information Science and Technology Initiative. 17 July 2000. Archived from the original (PDF) on 5 September 2012. Retrieved 18 August 2012.
7. "About the CCMB". Center for Computational Molecular Biology. Retrieved 18 August 2012.