Genome Informatics 2024

Lesson 1 - Introduction

Communication

github.com/vladimirkovacevic/gi-2023-etf
(All info about the course. Create an issue!)

vladimir.kovacevic@etf.bg.ac.rs
General info (not relevant to all)
Questions about lessons 3-8

marko.misic@etf.bg.ac.rs
Course professor
Questions about lessons 1-2

pedjao@etf.bg.ac.rs
Teaching assistant
Questions about lessons 9-11, technical details and exercises

Course staff

Vladimir Kovacevic finished computer science in 2007 at the School of Electrical Engineering where he also obtained his PhD in 2016. He worked for 5 years for Intel as algorithm developer and 5 years he's been with Seven Bridges bioinformatics company and teaching Genome informatics course at School of Electrical Engineering in Belgrade. His current interest includes machine learning applied in genomics, precision medicine, neoantigen discovery and bioinformatics in general.

Marko Mišić is an associate professor at the University of Belgrade, School of Electrical Engineering, Republic of Serbia. He completed B.Sc. M.Sc. and Ph.D. in the field of Computer Engineering and Informatics, at the same university. His main areas of interest are parallel programming with an emphasis on GPU programming, algorithms, data structures, bioinformatics and complex network analysis.

Predrag Obradović is a senior teaching assistant at the University of Belgrade, School of Electrical Engineering, Republic of Serbia. He completed B.Sc. M.Sc. in the field of Software Engineering, at the same university, where he pursues Ph.D. in Computer Engineering and Informatics. His research interests include social and complex network analysis, bioinformatics, and machine learning.

Course info

- 13 classes (lecture + exercise) - ~2.5h
- Exam will have both theoretical and practical part
  - 40% on the exam
  - 60% during the semester
    - 40% project assignment (with presentation)
    - 20% - exercises (homework)
- Last class - presentation of student assignments
- Exercise will follow lectures - examples in python Jupyter notebook

Course info - syllabus

1	29. 2. 2024.	Course info. Bioinformatics and genomics definitions. Exercise: Introduction to Python and Jupiter environment, Python structures.
2	7. 3. 2024.	Unix commands in bioinformatics. Git. Exercise: Pandas library
3	14. 3. 2024.	Molecular biology basics. Genome sequencing technologies. Python exercise: Performance measurement. Writing tests.
4	21. 3. 2024.	Portable and reproducible bioinformatic analysis. Describing bioinformatic tools in Common Workflow Language.
5	28. 3. 2024.	Executing bioinformatic analysis locally and on the cloud. Variant calling. Cancer analysis.
6	4. 4. 2024	Single cell RNA analysis. Scanpy library.
7	11. 4. 2024.	Spatial transcriptomics analysis and algorithms.
8	18. 4. 2024.	Burrows-Wheeler Transform and FM Index.
9	25. 4. 2024.	Approximate string matching, Edit distance, Dynamic programming, Global alignment.
10	9. 5. 2024.	Variation on global alignment (end-space-free variant, longest common substring) , local alignment, gaps. Practical: BLAST, Bowtie.
11	16. 5. 2024.	De-Bruijn graph, scaffolding, error correction. (Jovana K, guest lecturer)
12	23. 5. 2024.	Presentation of the student projects

Marko, Vladimir, Predrag, Jovana

Literature

- Vince Buffalo: **Bioinformatics Data Skills**
- Dan Gusfield: **Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology,** Cambridge University Press
- Pavel Pevzner, Neils Jones: **An Introduction to Bioinformatics Algorithms (Computational Molecular Biology)**, MIT Press
- R. Durbin, S. Eddy, A. Krogh, G. Mitchinson: **Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids** , Cambridge University Press
- Veli Mäkinen, Djamal Belazzougui, Fabio Cunial, Alexandru I. Tomescu: **Genome-Scale Algorithm Design: Biological Sequence Analysis in the Era of High-Throughput Sequencing**, Cambridge University press

What is bioinformatics?

Unprecedented wealth of **biological data** has been generated by high-throughput sequencing technologies

- The human genome project - 20 universities (US, UK, JP, FR, DE, PRC)
- 1000 Genomes Project - International research effort on genetic variation
- 100,000 Genomes Project – UK Government project
- The 3,000 rice genomes project
- Personal genome - $100 per human whole genome sequencing in the nearest future

The huge demand for analysis and interpretation of these data

What is bioinformatics?

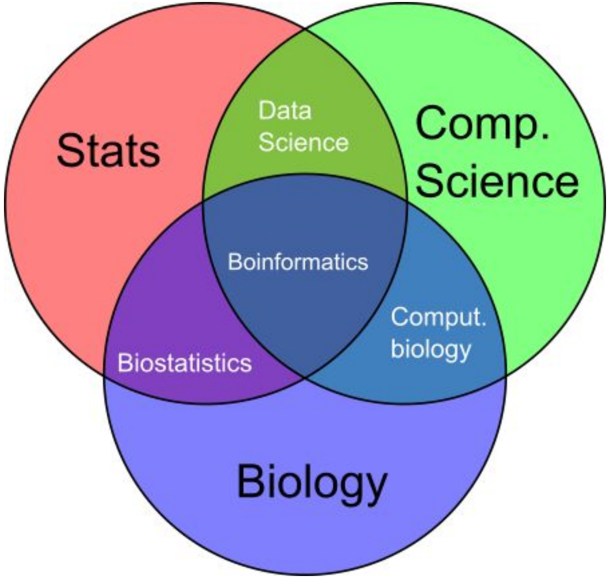
Bioinformatics, n. The science of information and information flow in biological systems, esp. of the use of computational methods in genetics and genomics.
(Oxford English Dictionary)

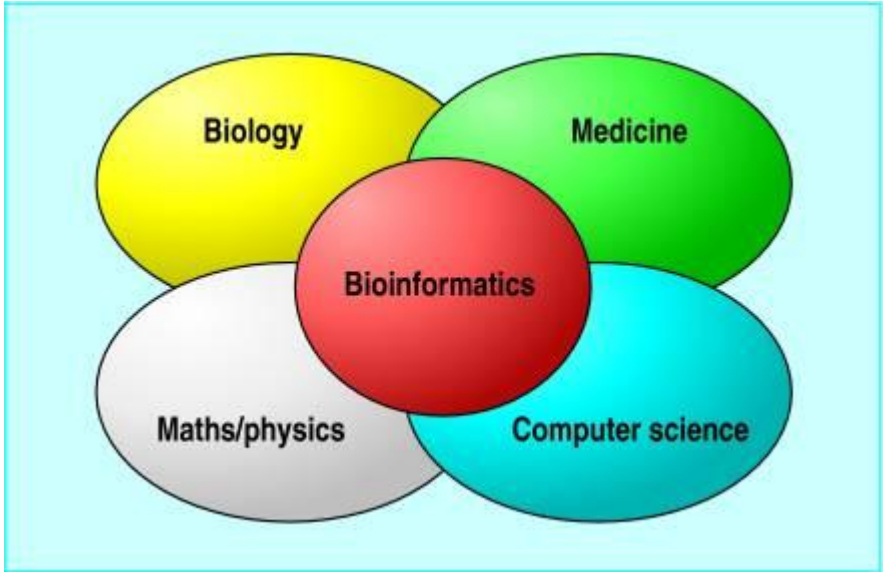
Bioinformatics - using statistical and computing methods that aim to solve biological problems.

Bioinformatics is defined as **the application of tools of computation and analysis to the capture and interpretation of biological data**.
(NIH)

What is bioinformatics?

An **interdisciplinary field**, which harnesses computer science, mathematics, physics, and biology





What is bioinformatics

"I do not think all biological computing is bioinformatics, e.g. mathematical modelling is not bioinformatics, even when connected with biology-related problems. In my opinion, bioinformatics has to do with management and the subsequent use of biological information, particular genetic information."

-- Richard Durbin

Bioinformatics in practice: Develops methods and software tools for storing, retrieving, organizing and analyzing biological data.

Success story - the Humane Genome Project

An international scientific research projects with the goal of determining the base pairs that make up human DNA (1990 - 2003)

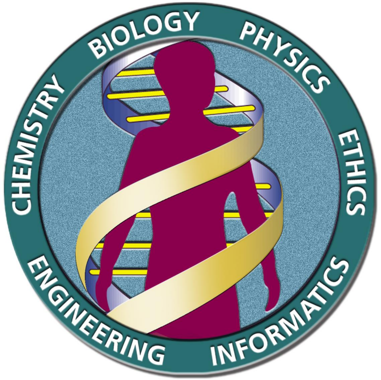
Identifying, mapping and sequencing all of the genes of the human genome from both a physical and a functional standpoint

One of the main achievements of bioinformatics to date

Included 92% of the genome

Level "complete genome" was achieved in May 2021

The final gapless assembly was finished in January 2022



Omics

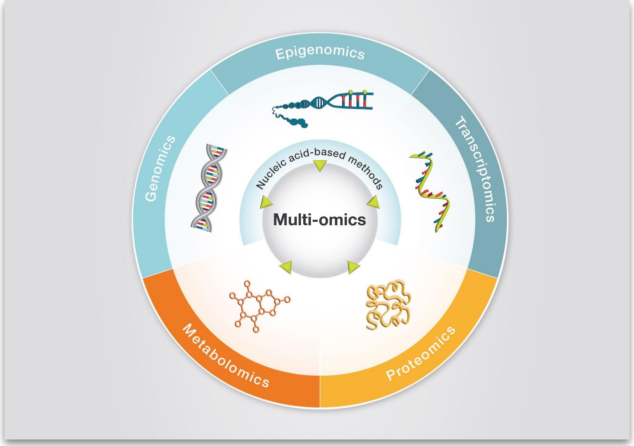
The branches of science known informally as **omics** are various disciplines in computational biology, allowed by **bioinformatics** development

Omics data refers to data generated from high-throughput sequencing technologies used to study the various "omes" of an organism

Four major omics:

- Genomics - all the genetic material
- Transcriptomics - all the RNA molecules
- Proteomics - all the proteins
- Metabolomics - all the small molecules

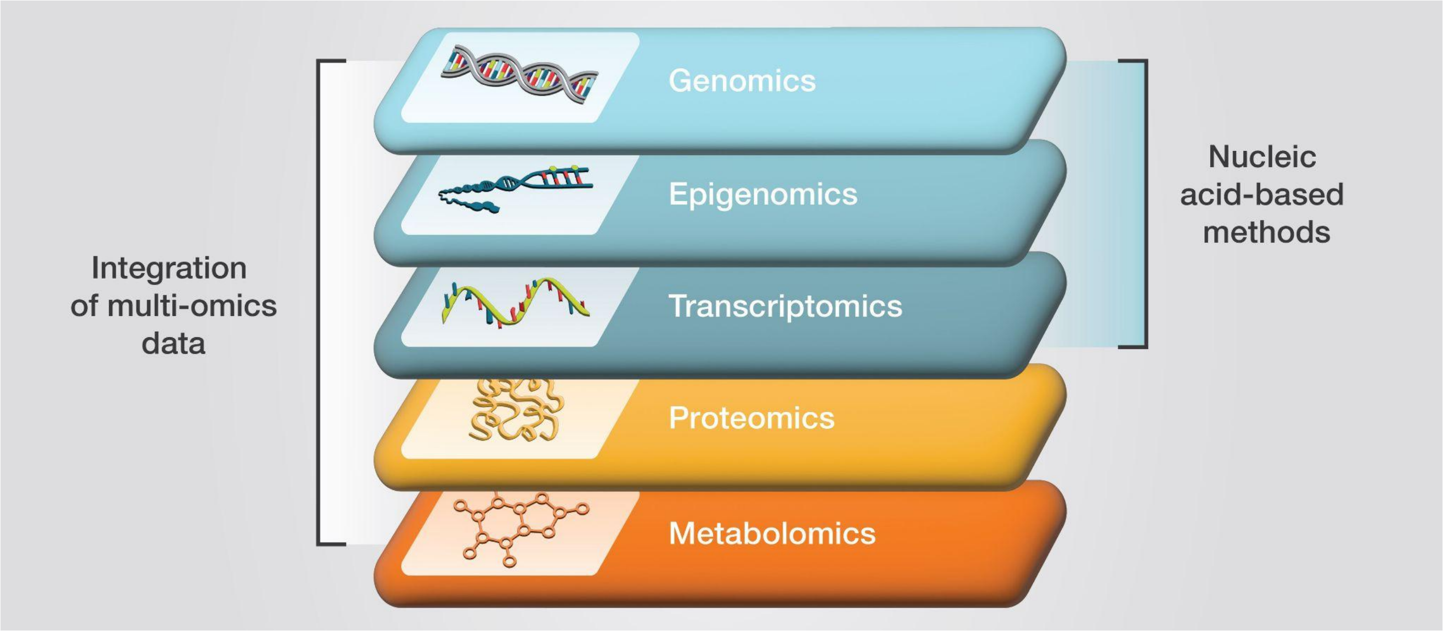
Doubling every 7 months!



Omics - the big picture

Multi-omic approach	Molecular read-out	Results	Technology
Genomics	Genes (DNA)	Genetic variants, gene presence or absence, genome structure	sequencing, exome sequencing
Epigenomics	Modifications of DNA	Location, type, or degree of reversible DNA modifications	Modification-sensitive PCR and qPCR, next-generation sequencing, mass spectrometry
Transcriptomics	RNA and/or cDNA	Gene expression, gene presence or absence, splice sites, RNA editing sites	RT-PCR (reverse transcription-PCR) and RT-qPCR, gene arrays, RNA-sequencing
Proteomics	Protein	Abundance of peptides, peptide modifications, and interactions between peptides	Mass spectrometry, western blotting, and ELISA
Metabolomics	Metabolites	Abundance of small molecules such as carbohydrates, amino acids, and fatty acids	Mass spectrometry, nuclear magnetic resonance (NMR) spectroscopy, and HPLC

Omics - the big picture



Molecular Biology - A Commonality between Various Omes in Multi-omics Approaches

Omics - some appliations

Oncology - combining proteomic, genomic, and transcriptomic data uncovered genes that are significant contributors to various cancers

Alzheimer's disease - discovering distinct differences between genetic predisposition and environmental contributions to Alzheimer’s disease

Drug discovery - multi-omics has helped in the identification and verification of drug targets

Cellular biology - improving the understanding of basic cellular biology, discovering new cell types

Infectious diseases - transcriptomics, proteomics, and antigen receptor analyses were combined to reveal insights into the immune response to COVID-19 infection and potential therapeutic target

Bioinformatics goals

The **primary goal** of bioinformatics is to increase the understanding of biological processes

Focus on developing and applying **computationally intensive techniques** to achieve this goal:

- Pattern recognition
- Data mining
- Machine learning algorithms
- Visualization

Bioinformatics focus

Major research **efforts** in the field include:

- Sequence alignment
- Gene finding
- Genome assembly
- Drug design & discovery
- Protein structure alignment & structure prediction
- Prediction of gene expression and protein–protein interactions
- Genome-wide association studies
- The modeling of evolution
- Cell division/mitosis mechanisms

Bioinformatics databases

Bioinformatics databases are used for efficient management of biological data, storing the results of the analyses, research and applications

- Biological sequence analysis
- Structure analysis
- Next Generation Sequencing
- Network Analysis: Metabolic Pathway Databases, Interaction Analysis Databases, Functional Networks

Both **empirical** and **predicted** data

Different **storage** formats: text, semi-structured data, XML, tables…

Bioinformatics principles

Golden rule of bioinformatics:

- **Never ever trust your tools (or data)**

Adapt **robust** and **reproducible** practices:

- Document each step
- Write down key facts
- Conclude the work done to easy understand it again in few months (make figures, reports)
- Automate manual tasks with script
- Always ask yourself: How much time would it take me to do it one more time?
- Release or publish your code and tools

**Vince Buffalo: Bioinformatics Data Skills*

Bioinformatics principles

Write code for humans, write (meta)data for computers:

- Make assertions (be loud)
- Let the code test the code
- Use existing libraries whenever possible
- Frequently used scripts -> tools
- Use code versioning (git)

Essentially, follow good **software engineering** practices!

**Vince Buffalo: Bioinformatics Data Skills*

Genomics 101

Genome: “The complete set of genes or genetic material present in a cell or organism.” *(Oxford English Dictionary)*

- “Blueprint” or “recipe” of life
- Human genome - 6 billions of base-pairs (A, C, T, G) letters
  - Can be imagined as a string 6 billion letters long

ACTGTGTCACATCGAGAGAGATCACAACACATAGATTACGATCGTAACGTAACGTAAC
ACCCAAATATACGAGTGAGGGGTGGGGACCCCCCCCCCCCACACACATTTTACAA
ACCTAGATCACCATACAGATATAAGAGAGAGANACGTACGTACACAATTACAAATTAAC
AACACAAAGTACTTATACATACACATGGGACCCATAGCACACACAGATATTTATAATAT
ATAGAGAGACAATGTCGTGCTGCAGTAA...

Genomics: contrast with biology & genetics*

* Everything on this slide is a gross generalization

Biology & Genetics	↔	**Genomics**
Targeted studies of one or a few genes	scope	Studies considering all genes in a genome
Targeted, low-throughput experiments	technology	Global, high-throughput experiments
Clever experimental design, painstaking experimentation	hard part	Tons of data, uncertainty, computation

Used with permission of Ben Langmead

Resources and additional reads

Presentation available at: github.com/vladimirkovacevic/gi-2024-etf

- A Computer Scientist’s Guide to Cell Biology, A Travelogue from a Stranger in a Strange Land
- Genomics 101, Edition 2016
- Bioinformatics at COMAV - SNP Calling
- Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM
- High-Throughput Sequencing Technologies - Review paper
- Vince Buffalo: Bioinformatics Data Skills
- Dan Gusfield: Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology, Cambridge
- Pavel Pevzner, Neils Jones: An Introduction to Bioinformatics Algorithms (Computational Molecular Biology), MIT
- R. Durbin, S. Eddy, A. Krogh, G. Mitchinson: Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press
- Veli Mäkinen, Djamal Belazzougui, Fabio Cunial, Alexandru I. Tomescu: Genome-Scale Algorithm Design: Biological Sequence Analysis in the Era of High-Throughput Sequencing, Cambridge University press
- Molekularna biologija 1; Dušanka Savić-Pavićević, Gordana Matić; NNK International, 2020
- RNA-seqlopedia; Cresko Lab, University of Oregon
- ThermoFisher Scientific - Molecular Biology —A Commonality between Various Omes in Multi-omics Approaches