CZU: 004+57 DOI: 10.46727/c.steam-2023.p328-333

BIOINFORMATICS AND APPLICATIONS IN GENOMIC RESEARCH

Stela GOLOVCO, PhD student in Inflammation, Immunity and Cancer
https://orcid.org/0000-0002-5307-1932
Department of Medicine, University of Verona

Abstract. The article presents general information about bioinformatics, its fields, describes the IT tools used in data processing. Also, the stapes of realizing a bioinformatics data processing are presented and an example of an application is described in the realization of which the author was involved.
Keywords: bioinformatics, bioinformatics tools, gene, biological research.

Rezumat. În articol se prezintă informații generale despre bioinformatică, domeniile ei, se descriu instrumentele informatice utilizate în prelucrarea datelor. De asemenea, sunt prezentate etapele realizării unei prelucrări a datelor bioinformatice și se descrie un exemplu de aplicație la realizarea căreia a fost implicată autoarea.
Cuvinte cheie: bioinformatica, instrumente bioinformatice, gene, cercetare biologică.

What is Bioinformatics?

Bioinformatics is an interdisciplinary field that processes biological data using computational tools and algorithms to extract novel relevant information and interpret the results. It facilitates the storage, retrieval, and analysis of biological data, enabling researchers to make significant discoveries in various areas, including evolutionary biology, molecular medicine, and agricultural science.

Bioinformatics is a rapidly evolving field with diverse applications across various domains of biological research. Some of the prominent research fields within bioinformatics include:

- *Genomics*: Genomics involves the study of the structure, function, and evolution of genomes.
- *Proteomics*: Proteomics focuses on the large-scale study of proteins, including their structures, functions, and interactions.
- *Transcriptomics*: Transcriptomics involves the study of the transcriptome, including mRNA, non-coding RNA, and other RNA transcripts.
- *Pharmacogenomics*: Pharmacogenomics aims to understand how an individual's genetic makeup influences their response to drugs.
- *Systems Biology*: Systems biology aims to understand biological systems as integrated and interconnected networks of genes, proteins, and other molecular components.

A general workflow in bioinformatics involves several key steps, from data acquisition and preprocessing to analysis and interpretation. Follows an overview of a

typical bioinformatics workflow that may vary depending on the specific research analysis being conducted.

1. *Data collection*: The process begins with the collection of biological data such as extraction of DNA or RNA, protein sequences or other types of biological information from public databases, experimental data, or literature.
2. *Sequencing*: The process of determining the precise order of nucleotides within a DNA or RNA molecule. There are several sequencing techniques including Sanger Sequencing, Next-Generation Sequencing (NGS), Third-Generation Sequencing.
3. *Raw data acquisition*: The nucleotide sequences produced by the sequencing are stored in a FASTA format file, which is a standard text-based format where each sequence has a unique identifier.
4. *Data analysis*: The sequenced data requires preprocessing and analysis in order to ensure the quality of the data and uncover patterns, relationships and structures within complex biological data. This step is performed by a bioinformatician applying various computational tools and algorithms.
5. *Data visualization and interpretation*: Visualization tools are used to present the data, analysis results, and models in a comprehensible format, such as graphs, charts, and diagrams. The results are interpreted in the context of the research question.

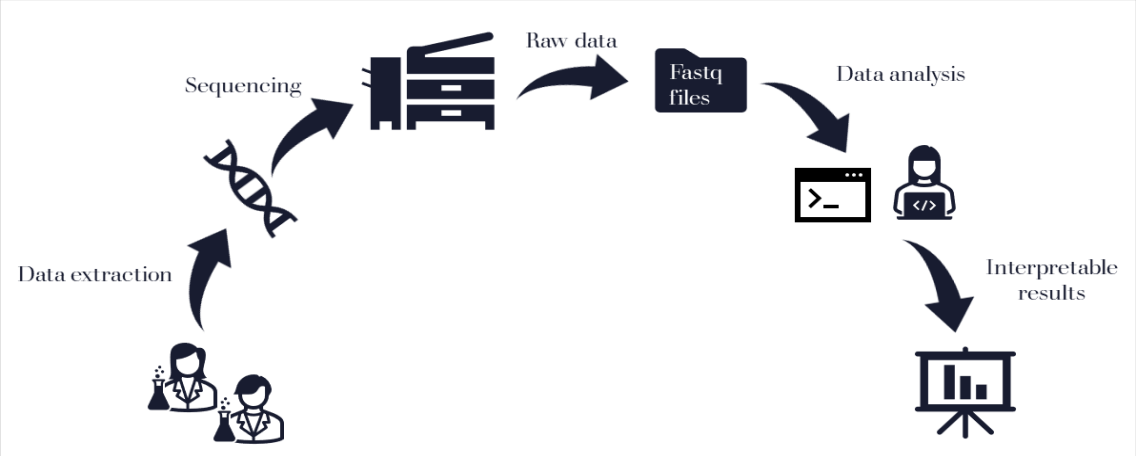

Figure 1. Key steps of General workflow in bioinformatics

Bioinformatics tools

Among the most known and used bioinformatics tools we find the *RStudio platform*, which is a functional integrated development environment (IDE) available for the R. R is an open source language used to deal with statistical computing and graphics. It is a free software available for most common operating systems (Window, Linux, MacOS) and it is organized in packages divided by the different usage and area of interested. This software allows us to explain and visualize the data through graphic representations such as, diagrams, plots, heatmaps, in a clear and understandable manner.

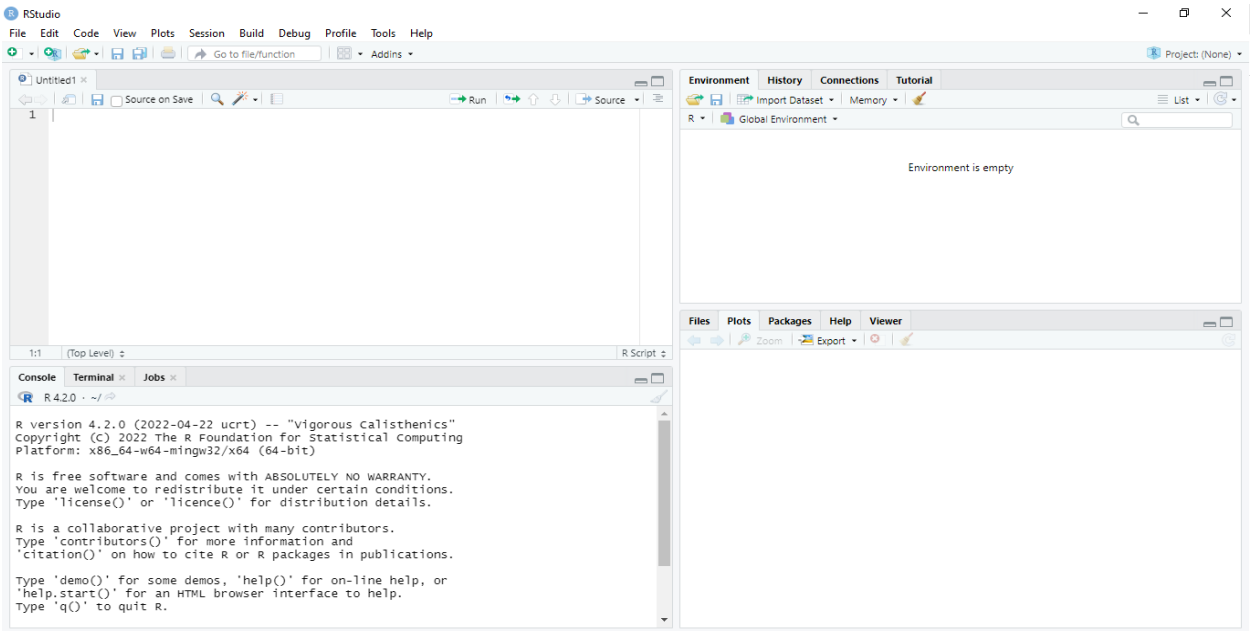

Figure 2. The *RStudio platform*

R system is available from the Comprehensive R Archive Network https://cran.r-project.org/

Bioconductor

Bioconductor is a collection of R packages for the analysis and comprehension of high throughput genomic data. It contains packages for expression and other microarrays, sequence analysis, imaging, and other domains. The *Bioconductor* web site is at https://www.bioconductor.org/ and provides installation, package repository, help, and other documentation.

Bioconductor is well suited to handle extensive data and to perform specific data analysis including: sequence analysis such as the alignment of biological sequences to reference sequence to identify similar regions that may have functional or structural relationships (Biostrings package); genomic annotation used to identify functional elements within a genome sequence, such as genes, regulatory elements and to detect different types of variants, such as singular nucleotide variants, indels, copy number alterations and structural variants (VariantAnnotation, GenomicFeatures, biomaRt, BSgenome packages); gene expression analysis uses RNA data to detect which genes are expressed in a sample and which are not, to estimate transcripts amount to identify differentially expressed genes among different groups (*Limma*, *edgeR*, *Glimma*, *DESeq2* and *GSEA* packages).

Gene Set Enrichment Analysis (GSEA)

GSEA is another computational method used for analyzing and interpreting genomic expression data. It helps to determine whether a defined set of genes shows statistically

significant differences between two biological states, such as disease and control conditions, or between different phenotypes.

Example of application

Follows an example of its possible application on the Lung Squamous Cell Carcinoma expression data downloaded from the cBio Cancer Genomics Portal using the *cbioportalR* and *cgdsr* packages of Bioconductor.

The research question was *to find if there is a correlation between the copy number alterations of the Rictor gene and different curated gene sets from online pathway databases*. After an initial data collection and preprocessing, the samples were divided into two groups:

1. 51 samples with *Rictor* high level amplification (>4 copies) and
2. 103 samples with 2 copies of *Rictor*, for a total of 154 samples.

For each sample, we downloaded the mRNA expression data, RSEM (Batch normalized from Illumina HiSeq_RNASeqV2).

	CNV	Phenotype
TCGA.18.3406.01	0	Diploid
TCGA.18.3407.01	0	Diploid
TCGA.18.3408.01	0	Diploid
TCGA.18.3414.01	2	Amplification
TCGA.18.3416.01	0	Diploid
TCGA.18.4083.01	2	Amplification
TCGA.18.4086.01	0	Diploid
TCGA.18.4721.01	0	Diploid
TCGA.21.1076.01	0	Diploid
TCGA.21.1079.01	2	Amplification
TCGA.21.5784.01	2	Amplification
TCGA.21.5787.01	2	Amplification
TCGA.22.1000.01	0	Diploid
TCGA.22.1005.01	0	Diploid
TCGA.22.1016.01	2	Amplification

Showing 1 to 15 of 154 entries, 2 total columns

	A1BG	A1BG.AS1	AADAC	AADACL2	AADACP1	ABCC4	ABCC5
TCGA.18.3406.01	741.6930	212.8270	84.2520	4.7244	200.0000	501.5750	1933.860
TCGA.18.3407.01	46.7127	44.5766	51.8554	0.9515	18.0780	349.6670	10059.000
TCGA.18.3408.01	1.1864	16.5861	218.8940	16.6098	42.7110	2289.7800	45697.800
TCGA.18.3414.01	40.2769	21.5987	131.0690	132.9150	128.4840	393.5760	2473.690
TCGA.18.3416.01	55.9910	64.0191	34.0320	0.5579	16.7370	266.6770	18239.500
TCGA.18.4083.01	35.6266	50.4037	294.7860	126.5770	12.1951	1351.1400	9005.050
TCGA.18.4086.01	61.3810	66.4499	150.0870	79.2013	22.1764	836.3660	8530.380
TCGA.18.4721.01	43.2331	76.3346	13.1579	4.6992	15.0376	632.5190	15486.800
TCGA.21.1076.01	57.4080	81.9628	182.0340	44.3213	33.6367	506.1340	5228.730
TCGA.21.1079.01	154.3280	169.1460	7.6931	0.4274	229.5120	359.0130	1125.760
TCGA.21.5784.01	144.1340	55.9593	101.7360	3.0963	16.3663	300.3430	7096.320
TCGA.21.5787.01	36.5497	27.2844	2.9240	0.0000	2.1930	106.7250	331.506
TCGA.22.1000.01	108.7860	85.4322	175.3180	0.0000	6.6408	566.6850	2062.200
TCGA.22.1005.01	185.7220	186.1790	89.5795	0.0000	14.6252	488.1170	546.618
TCGA.22.1016.01	112.5460	174.5120	124.2940	33.0912	1.2107	370.0570	1246.570
TCGA.22.1017.01	61.4863	72.7287	26.2169	0.0000	1.1566	559.0360	794.988

Showing 1 to 16 of 154 entries, 19387 total columns

Figure 3. Data collection about the samples

Name	Type	Value
c2	list [6229] (GSEABase::GeneSetCollecti	List of length 6229
BIOCARTA_FEEDER_PATHWAY	S4 (GSEABase::GeneSet)	S4 object of class GeneSet
BIOCARTA_PROTEASOME_PATHWAY	S4 (GSEABase::GeneSet)	S4 object of class GeneSet
BIOCARTA_KREB_PATHWAY	S4 (GSEABase::GeneSet)	S4 object of class GeneSet
ST_INTERFERON_GAMMA_PATHWAY	S4 (GSEABase::GeneSet)	S4 object of class GeneSet
ST_WNT_CA2_CYCLIC_GMP_PATHWAY	S4 (GSEABase::GeneSet)	S4 object of class GeneSet
ST_DIFFERENTIATION_PATHWAY_IN_PC12_CELLS	S4 (GSEABase::GeneSet)	S4 object of class GeneSet
ST_TUMOR_NECROSIS_FACTOR_PATHWAY	S4 (GSEABase::GeneSet)	S4 object of class GeneSet
ST_ERK1_ERK2_MAPK_PATHWAY	S4 (GSEABase::GeneSet)	S4 object of class GeneSet
ST_GA12_PATHWAY	S4 (GSEABase::GeneSet)	S4 object of class GeneSet
ST_G_ALPHA_S_PATHWAY	S4 (GSEABase::GeneSet)	S4 object of class GeneSet
ST_G_ALPHA_I_PATHWAY	S4 (GSEABase::GeneSet)	S4 object of class GeneSet
ST_IL_13_PATHWAY	S4 (GSEABase::GeneSet)	S4 object of class GeneSet
ST_P38_MAPK_PATHWAY	S4 (GSEABase::GeneSet)	S4 object of class GeneSet

Figure 4. Description of data variables

Next step was to perform gene set enrichment analysis using the *GSEABase* and *GSVA* packages, highly used in the interpretation of gene expression data. It offers tools for handling gene sets and creating custom gene set collections. In our case, a collection of

annotated gene sets from the *The Molecular Signatures Database* (MSigDB) was used to identify the pathways associated with specific gene expression profiles.

Gsva function calculates enrichment scores for the defined gene sets. The obtained matrix provides the enrichment scores, in terms of zscore, for each input gene set or pathway in each sample.

The final step was to create a linear model using *limma* package that describes the relationship between different gene expression levels and our groups of interest. We also computed empirical Bayes moderation for a better estimation of the statistical parameters and differential gene expression. Then we extracted the most significantly differentially expressed gene sets by applying the statistical correction Benjamini & Hochberg (BH) and setting a *pvalue* cutoff equal to 0.05.

	TCGA.18.3406.01	TCGA.18.3407.01	TCGA.18.3408.01	TCGA.18.3414.01	TCGA.18.3416.01	TCGA.18.4083.01
BIOCARTA_FEEDER_PATHWAY	-0.56327450	-0.564038792	-0.2828065865	-1.5183524728	-0.31615542	-0.067532718
CARTA_PROTEASOME_PATHWAY	4.35790839	-1.508085386	3.3446944633	-0.0380644308	1.51209489	-0.390841371
BIOCARTA_KREB_PATHWAY	0.70371427	-0.368417885	0.4314694977	3.1020511990	0.07433884	-0.690813734
INTERFERON_GAMMA_PATHWAY	3.17915558	0.703781735	0.0354613989	-0.6005266269	2.51113683	0.786171238
/NT_CA2_CYCLIC_GMP_PATHWAY	-0.12296314	0.265531722	-0.2265145647	-1.1626060936	-1.56278671	-1.115811342
ITION_PATHWAY_IN_PC12_CELLS	-0.34749830	-0.220019488	0.7671670104	0.8033075859	-1.04853801	0.540332353
R_NECROSIS_FACTOR_PATHWAY	-0.27444219	-0.217717986	-1.8047846210	-0.3846263525	1.38714097	-2.037111746
ST_ERK1_ERK2_MAPK_PATHWAY	-0.24243749	0.270572342	-0.1075275220	0.2080261562	0.28792283	-0.987868293
ST_GA12_PATHWAY	-0.65486205	-1.283794011	0.3605972638	0.0949300032	-1.80585690	-1.379261882
ST_G_ALPHA_S_PATHWAY	0.36014869	-0.076157724	1.9136660040	1.5789066184	-0.63947429	-1.403102427
ST_G_ALPHA_I_PATHWAY	-2.47530929	-0.710735140	-0.4394499468	-0.2918449105	-1.92684220	1.239378880
ST_IL_13_PATHWAY	0.38612174	0.463068419	-0.4525799293	-1.0858278215	-0.19163092	-0.003437427
ST_P38_MAPK_PATHWAY	1.16926354	-0.700067777	1.3858495359	1.6749071800	-0.56161194	-1.237219605
ST_JAK_STAT_PATHWAY	-0.74211469	-0.199684863	-0.6178642266	-0.3889023184	0.15399203	0.174434735
NULE_CELL_SURVIVAL_PATHWAY	0.86078441	0.832378012	1.5395395131	1.4123981026	-0.34368363	-2.092124062
ST_ADRENERGIC	-0.36931464	-0.681273404	0.5828395542	0.4183275978	-0.78464553	-0.927326194
INTEGRIN_SIGNALING_PATHWAY	-2.25282185	-0.057738798	-0.2449535712	-1.6087761557	-1.09513631	-0.414089309
ST_GAQ_PATHWAY	-2.02973830	-1.487868601	-1.9599387552	-0.4353607747	-1.06153514	-0.873033150
ST_GA13_PATHWAY	-0.30021776	-0.673583786	0.4138088212	-1.3062965337	-0.64977915	-0.559272780
ST_STAT3_PATHWAY	-0.80376237	0.509892193	-1.3206590902	-1.0419578486	-0.98167049	-1.083276043

Showing 1 to 21 of 6.228 entries, 154 total columns

Figure 5. Linear model using *limma* package

The gene set enrichment analysis was performed using seven gene set collections from MSigDB (hallmark, C2, C4, C5, C6, C7, C8), but only “C2” gene set collection showed statistically significant results. "C2" category includes curated gene sets from online pathway databases, publications in PubMed, and knowledge of domain experts. The PETRETTO_LEFT_VENTRICLE_MASS_QTL_CIS_DN gene set was excluded since it was not relevant to the aim of the study nor compatible with the cellular nature of our case study.

Figure 6 shows a graphical representation of the gene expression profile among the two groups. Functions from the *ComplexHeatmap* package were used to create an informative heatmap for visualizing the expression profile for each pathway in every sample. The heatmap reports zscore values for every pathway in each sample. To highlight the enrichment difference among the two groups for a single pathway, we grouped the samples according to their genomic characterization.

This highlights that, although there is a statistically significant enrichment difference for each geneset, there is heterogeneity within each group. In fact, not all samples from the same group showed a comparable enrichment score between them. Furthermore,

comparing the enrichment scores of the different genesets within the same group it is evident that these are not correlated with each other. For example, not all samples that showed positivity for the NIKOLSKY_BREAST_CANCER_5P15_AMPLICON geneset had a positive enrichment score for the PHONG_TNF_TARGETS_DN geneset.

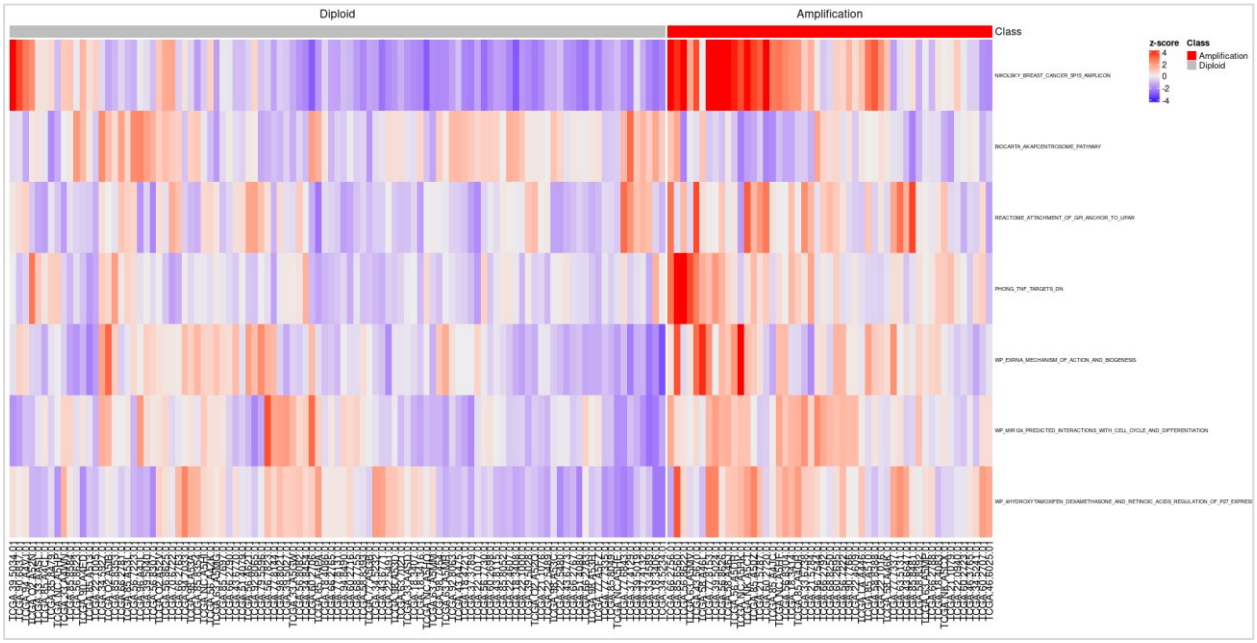

Figure 6. Heatmap visualization of differential expression profiles

Conclusion and Future Prospects

Bioinformatics is an ever-changing field that combines biology, computer science, and information engineering. Bioinformatics experts are required in all sectors of biomedical organizations, biotechnology, pharmaceutical, hospital, research institutions, and industry. For example, with the use of genomic data, researchers can identify specific genetic variations that influence an individual's response to particular treatments. This could lead to the development of more tailored therapies and improved patient outcomes. Moreover, bioinformatics is essential in designing and optimizing biological systems for various applications, including the prediction of protein structures, drug discovery processes and the production of novel biomaterials.

References

1. CARLSO, M. et. al. High-throughput sequence analysis with R and Bioconductor. June, 2022. 83 p.
2. LIBERZON, A et. al. Molecular signatures database (MSigDB) hallmark gene set collection. In: *Cell Syst*. 2015 Dec 23, 1(6), pp. 417–425.