Federal State Autonomous Educational Institution of Higher Education "Moscow Institute of Physics and Technology (National Research University)"

APPROVED
Head of the Phystech School of Biological and Medical Physics
D.V. Kuzmin

Work program of the course (training module)

course: Data Visualization/Визуализация данных
major: Applied Mathematics and Physics
specialization: Applied Bioinformatics/Прикладная биоинформатика
Phystech School of Biological and Medical Physics
Chair of Bioinformatics and Systems Biology
term: 1
qualification: Master

Semester, form of interim assessment: 2 (spring) - Exam

Academic hours: 45 AH in total, including:
lectures: 30 AH.
seminars: 15 AH.
laboratory practical: 0 AH.

Independent work: 60 AH.

Exam preparation: 30 AH.

In total: 135 AH, credits in total: 3

Authors of the program:

S.A. Bruskin, candidate of biological sciences, associate professor
E.V. Chekalin

The program was discussed at the Chair of Bioinformatics and Systems Biology 04.06.2020

Annotation

This discipline was introduced to teach data analysis and visualization using language R. Discipline objectives: to teach the basics of the programming language R; teach you how to use IDE R-studio; acquaint with the basic packages for data analysis using R; the formation of basic bioinformatics skills among students and the acquisition of practical experience necessary for conducting independent research in the field of systems biology.

1. Study objective

Purpose of the course

Study practical skills in using the R, IDE R-studio language and basic packages for data analysis.

Tasks of the course

- Study the basics of the R programming language;
- Learn to use IDE R-studio;
- Introduce basic data analysis packages with R;
- The development of students' basic bioinformatic skills and the acquisition of practical experience is necessary for conducting independent scientific research in the field of system biology.

2. List of the planned results of the course (training module), correlated with the planned results of the mastering the educational program

Mastering the discipline is aimed at the formation of the following competencies:

Code and the name of the competence	Competency indicators
Gen.Pro.C-2 Acquire an understanding of current scientific and technological challenges in professional settings, and scientifically formulate professional objectives	Gen.Pro.C-2.1 Assess the current state of mathematical research within professional settings
	Gen.Pro.C-2.2 Assess the relevance and practical importance of research in professional settings
	Gen.Pro.C-2.3 Understand professional terminology used in modern scientific and technical literature and present scientific results in oral and written form within professional communication
Pro.C-3 Use research and testing equipment (devices and installations, specialized software) in a selected subject field	Pro.C-3.1 Understand the operating principles of the equipment and specialized software
	Pro.C-3.2 Conduct an experiment (simulation) using research equipment (software)
	Pro.C-3.3 Evaluate the accuracy of the experimental (numerical) results

3. List of the planned results of the course (training module)

As a result of studying the course the student should:

know:

- Main packages of the R software environment;
- Basic R syntax;

be able to:

- Obtain an ability of programing in R language
- Implement and debug bioinformatics algorithms;
- Implement statistical analysis in a software environment R.

master:

- skills of working with large volumes of biological data;
- the culture of planning and implementation of multi-stage bioinformatic analysis.

4. Content of the course (training module), structured by topics (sections), indicating the number of allocated academic hours and types of training sessions

4.1. The sections of the course (training module) and the complexity of the types of training sessions

№	Topic (section) of the course	Types of training sessions, including independent work			
		Lectures	Seminars	Laboratory practical	Independent work
1	Introduction to the programming language R and IDE R-studio.	2	1		6
2	Syntax of R	2	1		6
3	Operators and functions	2	1		6
4	Basic graphics in R	2	1		6
5	Implementing graphs with ggplot2	4	2		6
6	Correlation and linear regression	4	2		6
7	PCA and Heatmaps	2	1		6
8	Clustering in R;	4	2		6
9	NGS and search for differentially expressed genes;	4	2		6
10	Packages for working with sequences. The final lesson.	4	2		6
AH in total		30	15		60
Exam preparation		30 AH.			
Total complexity		135 AH., credits in total 3			

4.2. Content of the course (training module), structured by topics (sections)

Semester: 2 (Spring)

1. Introduction to the programming language R and IDE R-studio.

Introductory lesson. The main commands in R. Setting up the R-studio workspace. Install packages and updates.

2. Syntax of R

The implementation of the code on R. Syntax R. The main objects in R, the concept of variables, array, matrix, data.frame, list, array.

3. Operators and functions

Operators in R. Implementation of functions in the language R, functions of the first level. Functions of the nth level. It is assumed that, based on the results of the module, students will be able to implement their own functions in the R language, as well as manipulate variables using the basic operators in R.

4. Basic graphics in R

Fundamentals of graphics in R. Construction of two-dimensional graphs on the plane using the base () package. Boxplot, barplot, pie plot, dot plot, histogram. Adjustable parameters of the plot and par () object. Legend on base () charts. The parameters of the axes of the plot in the base plot.

5. Implementing graphs with ggplot2

Implementing graphs using the ggplot2 package. Layers in ggplot2 objects. Additional features of the ggplot2 package: geom_point, geom_abline, geom_polygon, geom_rect. Facet and grid on ggplot2 charts. Adjustable environment settings graphics. Main (), axis labesl. Legend in ggplot2

6. Correlation and linear regression

The concept of correlation. Pearson and Spiren correlation. Data visualization. Linear regression. Multiple linear regression. R2 and F-statistics. Testing models. Test and training sample. ANOVA. Glm, generalized linear model. Logit regression and AIC. Amendment for multiple comparison. FDR, Bonferroni amendment.

7. PCA and Heatmaps

Analysis of the main components. General principles of PCA implementation. The prcomp functions of the stats package. Pca3d package Heat maps. Packages heatmap and heatmap.2. Data visualization with PCA and heatmap.

8. Clustering in R;

Basic clustering algorithms. Euclidean, manhattan, maximum, canberra. Hierarchical clustering. The distance between the clusters. Comlete, Single, Average linkage. Dendrogram as an object R. K-means, k-medoids. Self-organizing map. Silhouette.

9. NGS and search for differentially expressed genes;

General principles of the next generation sequencing. Trimming and QC Reed. FastQC and trimmomatic. Overview of alignment algorithms. BWA, Bowtie, STAR. Alignment of reads on the genome. Detection of alternative splicing. Cufflinks. Cuffdiff. Normalization of reads. DEXseq, EdgeR, limma. Log fold-change. Search for differentially expressed genes and identification of reliable DEGs.

10. Packages for working with sequences. The final lesson.

The main functions of the package Seqinr, Ape, annotation of genes using the package GenomicRanges, GenomicAlignments. Go-enrichent. Hypergeometric distribution.

5. Description of the material and technical facilities that are necessary for the implementation of the educational process of the course (training module)

Equipment required for lectures and seminars: computer and multimedia equipment (projector, sound system),
UNIX server with a separate account for each student.

6. List of the main and additional literature, that is necessary for the course (training module) mastering

Main literature
Provided at the department:
Team, R. C. (2013). R: A language and environment for statistical computing.
Shipunov, A. B., Baldin, E. M., Volkova, P. A., Korobeinikov, A. I., Nazarova, S. A., Petrov,

Additional literature
Provided at the department:
S. V., & Sufiyanov, V. G. (2012). Visual statistics. Use R !. M .: DMK Press, 298, 1.

7. List of web resources that are necessary for the course (training module) mastering

Scientific bibliographic and patent databases in the field of physico-chemical biology, available on the Internet in free mode –
Science Citation Index (Web of Science), Medline (PubMed), Scientific Electronic Library (NEB), Russian Patent DB of FGU FIPS and American USPAFULL patent database;
email addresses of major scientific publishers who provide access to the full text of current and archival issues of these journals.
http://stackoverflow.com/
https://stat.ethz.ch/pipermail/r-help/

https://www.biostars.org/
https://www.statmethods.net/

8. List of information technologies used for implementation of the educational process, including a list of software and information reference systems (if necessary)

Internet access. For some of the lessons, you need Zoom. Google Drive to access course materials. The presence of smartphones / laptops during classes is encouraged to participate in interactive exercises.

9. Guidelines for students to master the course

A student who studies discipline must, on the one hand, master a general conceptual apparatus, and on the other hand, must learn to apply theoretical knowledge in practice.
As a result of studying the discipline, the student should know the basic definitions of the discipline, be able to apply this knowledge to solve various problems.

Successful learning requires:
- visits to all classes provided by the curriculum for the discipline;
- conducting the abstract of occupations;
- intense independent work of the student.

Independent work includes:
- reading recommended literature;
- study of educational material, preparation of answers to questions intended for self-study;
- solving problems offered to students in the classroom;
- preparation for performance of tasks of the current and intermediate certification.

An indicator of possession of the material is the ability to answer questions on discipline topics without an outline.

It is important to achieve an understanding of the material being studied, and not its mechanical memorization. If it is difficult to study individual
topics, questions, you should seek advice from the teacher.

Intermediate control of students' knowledge in the form of problem solving in accordance with the subject of classes is possible

SUPPLEMENT

Assessment funds for course (training module)

major: Applied Mathematics and Physics
specialization: Applied Bioinformatics/Прикладная биоинформатика
Phystech School of Biological and Medical Physics
Chair of Bioinformatics and Systems Biology
term: 1
qualification: Master

Semester, form of interim assessment: 2 (spring) - Exam

Authors:

S.A. Bruskin, candidate of biological sciences, associate professor
E.V. Chekalin

1. Competencies formed during the process of studying the course

Code and the name of the competence	Competency indicators
Gen.Pro.C-2 Acquire an understanding of current scientific and technological challenges in professional settings, and scientifically formulate professional objectives	Gen.Pro.C-2.1 Assess the current state of mathematical research within professional settings
	Gen.Pro.C-2.2 Assess the relevance and practical importance of research in professional settings
	Gen.Pro.C-2.3 Understand professional terminology used in modern scientific and technical literature and present scientific results in oral and written form within professional communication
Pro.C-3 Use research and testing equipment (devices and installations, specialized software) in a selected subject field	Pro.C-3.1 Understand the operating principles of the equipment and specialized software
	Pro.C-3.2 Conduct an experiment (simulation) using research equipment (software)
	Pro.C-3.3 Evaluate the accuracy of the experimental (numerical) results

2. Competency assessment indicators

As a result of studying the course the student should:

know:

- Main packages of the R software environment;
- Basic R syntax;

be able to:

- Obtain an ability of programing in R language
- Implement and debug bioinformatics algorithms;
- Implement statistical analysis in a software environment R.

master:

- skills of working with large volumes of biological data;
- the culture of planning and implementation of multi-stage bioinformatic analysis.

3. List of typical control tasks used to evaluate knowledge and skills

During the current control, the student should be able to answer the following questions:

1. Generate two random Poisson sets of 200 numbers, one with an average of 0.5, the other with an average 3. Is there a linear relationship between them?
2. What structure is suitable for storing temperature values measured hourly in five patients per day. Create such a structure and fill it with random data. When did the second patient have a temperature higher than 40 degrees?
3. Create a new numeric variable new_var in the mtcars data that contains units in the lines if the car has at least four carburetors (the variable "carb") or more than six cylinders (the variable cyl.). In the lines in which the condition is not met, should be zeros.
4. In the mtcars data frame, create a new column (variable) called even_gear, in which there will be ones if the value of the variable (gear) is even, and zero if the number is odd.
5. Create 3 linear variables of the same length with arbitrary names, containing Check whether the sum of the first two numbers is strictly greater than the third number. The result of the comparison (TRUE or FALSE) save in a new variable named result.
6. Transform the mtcars data frame into a sheet and create a new sheet element called even_gear, in which there will be one if the value of the variable (gear) is even and zero if the number is odd.
7. Create a structure for storing the name, surname, age and gender of three people and fill it. How to make a simple search by last name?

8. Drunk goes on a bridge with a width of l steps. Every pitch drunk shifts randomly one step to the right or left. If a drunk stepped over the edge of the bridge - he falls. Write a function that counts how many steps a drunk will do before falling. The width of the bridge and the position of a drunk - the initial parameters. Drunken_path (l), output: ‘Our boozer will fall at n-d step’.
9. Write a function that receives two experimentally measured dependencies (x1, y1, x2, y2) at the input, approximates them by the direct least squares method, and returns the intersection coordinates.
10. Write a function that compares a set of words of the same length and returns a pairwise distance matrix — the proportion of mismatched letters.
11. Write a function that draws the trajectory of the moon relative to the sun. Assume that the earth moves evenly with the speed V1 around the Sun in an orbit of radius R1. The moon moves evenly at a speed V2 around the earth in an orbit of radius R1.
12. Write a function that draws a given set of points on a plane and connects with straight lines those that are less than a specified

During the class, interactive discussions can take place in the course chats, which will be homework. It is possible to perform patent search as an independent task. Successful completion of all tasks in the course and the completion of control slices of knowledge gives an advantage in the differential credit.

4. Evaluation criteria

Intermediate certification for the discipline "Data visualization» is carried out in the form of an exam (credit). Exam (test) is held in written (oral) form.

1. Generate two random Poisson sets of 200 numbers, one with an average of 0.5, the other with an average 3. Is there a linear relationship between them?
2. What structure is suitable for storing temperature values measured hourly in five patients per day. Create such a structure and fill it with random data. When did the second patient have a temperature higher than 40 degrees?
3. Create a new numeric variable new_var in the mtcars data that contains units in the lines if the car has at least four carburetors (the variable "carb") or more than six cylinders (the variable cyl.). In the lines in which the condition is not met, should be zeros.
4. In the mtcars data frame, create a new column (variable) called even_gear, in which there will be ones if the value of the variable (gear) is even, and zero if the number is odd.
5. Create 3 linear variables of the same length with arbitrary names, containing Check whether the sum of the first two numbers is strictly greater than the third number. The result of the comparison (TRUE or FALSE) save in a new variable named result.
6. Transform the mtcars data frame into a sheet and create a new sheet element called even_gear, in which there will be one if the value of the variable (gear) is even and zero if the number is odd.
7. Create a structure for storing the name, surname, age and gender of three people and fill it. How to make a simple search by last name?
8. Drunk goes on a bridge with a width of l steps. Every pitch drunk shifts randomly one step to the right or left. If a drunk stepped over the edge of the bridge - he falls. Write a function that counts how many steps a drunk will do before falling. The width of the bridge and the position of a drunk - the initial parameters. Drunken_path (l), output: ‘Our boozer will fall at n-d step’.
9. Write a function that receives two experimentally measured dependencies (x1, y1, x2, y2) at the input, approximates them by the direct least squares method, and returns the intersection coordinates.
10. Write a function that compares a set of words of the same length and returns a pairwise distance matrix — the proportion of mismatched letters.
11. Write a function that draws the trajectory of the moon relative to the sun. Assume that the earth moves evenly with the speed V1 around the Sun in an orbit of radius R1. The moon moves evenly at a speed V2 around the earth in an orbit of radius R1.
12. Write a function that draws a given set of points on a plane and connects with straight lines those that are less than a specified distance (maximum distance is a function parameter). The size of the points is proportional to the number of edges included in it (the maximum size is a function parameter). The color of the lines depends on the distance (the color corresponding to the minimum and maximum distance is the function parameters).

13. Write a function that finds all occurrences of the word (w) in the text (t) with no more than n errors. W, t and n are function parameters. The function returns the position of the beginning of the occurrences.

14. Write a function that finds all occurrences of the word (w) in the text (t) with no more than n errors. W, t and n are function parameters. The function returns the position of the beginning of the occurrences.

15. Write a function that draws a random broken line. The step length and rotation is determined (from the previous direction) is determined randomly on the basis of a uniform distribution. The range of lengths and rotations, as well as the number of steps, are the function parameters ..

16. Write a function that generates all possible sequences of a given length from a given alphabet.

17. For the AirPassengers data embedded in R, calculate the moving average with a smoothing interval of 10. Type the result (the first value in the output should be the average for the elements 1:10, in the second value the average for the elements 2:11, etc., in the latter - the average for the elements 135: 144). Save all the average values in the variable moving_average.

Examples of exam tasks:

Task 1. Transform the mtcars data frame into a sheet and create a new sheet element called even_gear, in which there will be one if the value of the variable (gear) is even and zero if the number is odd.

Task 2. Create a structure for storing the name, surname, age and gender of three people and fill it. How to make a simple search by last name?

Task 3. Drunk goes on a bridge with a width of l steps. Every pitch drunk shifts randomly one step to the right or left. If a drunk stepped over the edge of the bridge - he falls. Write a function that counts how many steps a drunk will do before falling. The width of the bridge and the position of a drunk - the initial parameters. Drunken_path (l), output: ‘Our boozer will fall at n-d step’.

Task 4. Write a function that receives two experimentally measured dependencies (x1, y1, x2, y2) at the input, approximates them by the direct least squares method, and returns the intersection coordinates.

Task 5. Write a function that compares a set of words of the same length and returns a pairwise distance matrix — the proportion of mismatched letters.

Can be left unchanged

The mark is excellent (10 points) - it is given to a student who has shown comprehensive, systematic, deep knowledge of the curriculum of the discipline, who has an interest in this subject area, has demonstrated the ability to confidently and creatively put them into practice in solving specific problems, and a free and proper substantiation of decisions.

The mark is excellent (9 points) - it is given to a student who has shown comprehensive, systematic, in-depth knowledge of the curriculum of the discipline and the ability to confidently put them into practice in solving specific problems, free and proper substantiation of the decisions made.

The mark is excellent (8 points) - given to a student who has shown comprehensive, systematic, in-depth knowledge of the curriculum of the discipline and the ability to confidently apply them in practice in solving specific problems, correct justification of decisions made, with some shortcomings.

A mark is good (7 points) - it is put up for a student, if he knows the material firmly, sets it up competently and in essence, knows how to apply the knowledge gained in practice, but does not competently substantiate the results obtained.

Evaluation is good (6 points) - it is put up to a student, if he knows the material firmly, sets it up correctly and in essence, knows how to apply this knowledge in practice, but admits some inaccuracies in the answer or in solving problems.

A mark is good (5 points) - it is given to a student, if he basically knows the material, correctly and essentially sets it out, knows how to apply this knowledge in practice, but allows a sufficiently large number of inaccuracies to answer or solve problems.

Grade satisfactorily (4 points) is given to a student who has shown the fragmented, fragmented nature of knowledge, insufficiently correct formulations of basic concepts, violations of the logical sequence in the presentation of program material, but at the same time he has mastered the main sections of the curriculum necessary for further education and can apply knowledge is modeled in a standard situation.

Grade satisfactorily (3 points) - given to a student who showed the fragmented, scattered nature of knowledge, making mistakes in formulating basic concepts, disrupting the logical sequence in presenting program material, poorly masters the main sections of the curriculum required for further education and even applies the knowledge gained in a standard situation.

The rating is unsatisfactory (2 points) - is given to a student who does not know most of the main content of the curriculum of the discipline, makes gross mistakes in the wording of the basic principles and does not know how to use this knowledge when solving typical tasks.

Unsatisfactory mark (1 point) - is given to a student who does not know the main content of the discipline's curriculum, makes gross errors in the wording of the basic concepts of the discipline and does not have any skills to solve typical practical problems.

5. Methodological materials defining the procedures for the assessment of knowledge, skills, abilities and/or experience

When conducting an oral differential test, the student is given 60 minutes to prepare. Interrogation of a student on a ticket on an oral differential test should not exceed one astronomical hour.