Questions of the State Exam by Specialty
Department of Bioinformatics, MIPT, 2020

Block 1. Probabilistic-theoretical and discrete-mathematical foundations of bioinformatics

1.1 Probability theory and statistics

1. Sequences of independent tests; Bernoulli scheme; geometric distribution.

2. Random variable; distribution of a random variable; mathematical expectation and variance of random variables

3. Conditional probability. Total probability and Bayes formulas.

4. Bayesian estimates of the frequencies of rare events. Pseudo-reports.

5. Conditional distribution. Conditional expectation; covariance and correlation

6. Laws of Large Numbers, Central Limit Theorem

7. Normal distribution

8. Discrete Markov chain. Stationary distribution.

9. Construction of the variation series of the sample and the empirical distribution function.

10. Construction of sample characteristics and ordinal statistics.

11. Statistical estimates of distribution parameters. Unbiasedness, consistency. Confidence intervals.

12. Plausibility. Likelihood function. Construction of estimates of distribution parameters using the maximum likelihood method.

13. Testing hypotheses. Calculation of type I and II errors.

14. Testing hypotheses. Construction of the Neumann-Pearson criterion.

15. Hypotheses about the form of distribution laws. Consent criteria.

16. Implementation of permutation statistical tests.

17. Odds ratio and relative risk.

18. Fisher's exact criterion.

19.FDR-method of accounting for multiple comparisons

20. Estimates of the sensitivity and specificity of the marker

21. Design of associative genetic studies.

1.2 Basic algorithms

1. K-measures. Search for the most common subsequences.

2. Search for motifs in the binding sites of transcription factors. Median row. A greedy algorithm for finding motives. A randomized algorithm for finding motives. Gibbs sampler.

3. Algorithms used in the assembly of genomes. Euler's way. Hamiltonian way.

4. Sequencing of cyclic peptides. Brute –force algorithm for cyclic peptide sequencing. Branch and bound method for cyclic peptide sequencing.

5. Aligning sequences. Dynamic programming. The Manhattan Tourist Challenge. Evaluation of alignments.

6. Dynamic programming on the graph. Finding the most difficult path between two peaks.

7. Dynamic programming on the graph. Finding the sum of weights in paths between two vertices.

8. Combinatorial algorithms. Algorithms for assessing rearrangements in the genome. Synthenic blocks.

9. Algorithms for constructing phylogenetic trees.

10. Clustering algorithms.

11. Algorithms for mapping readings. BWT and suffix trees

12. Hidden Macro Models

13. Algorithms for the identification of peptides using mass spectrometry.

Block 2. Practical aspects of bioinformatics

2.1 Databases and Internet resources

1. NCBI server. Databases containing information on the structure of genomes: BioProject, BioSample, SRA, GenBank, Genome, Nucleotide, Gene, Protein. Online inquiries.
2. Databases containing genetic information: dbSNP, ClinVar, OMIM, GTEx

3. Databases NCBI containing information about the literature: PubMed, PMC, Books. Effective search for publications. Relationship between PMID and doi.
4. Database NCBI GEO. Datasets, their device. Data escrow. Data analysis tools. GEO2R.
5. Genomic browsers: IGV. Genomic browser related databases: UCSC, Ensemble, GeneCode
6. UniProt database
7. Tools for determining the coding regions of the genome. GeneMark
8. Tools for functional annotation of genomes based on its domain structure: Pfam, InterPro

2.2 Practical methods of bioinformatics

1. Prediction of protein function based on its interactions with partner proteins. Data analysis of the IntAct Molecular Interaction Database.

2. Problems of bioinformatics arising in mass spectrometry. Data structure of the project "The humanprotein atlas". PeptideAtlas.

3. Basic principles of sequencing technologies. Types of sequencing technologies, biological mechanisms that can be studied on their basis

4. RNA-seq technology, questions that can be solved with its help. RNA-seq implementation difficulties

5. Stages of RNA-seq data analysis (bowtie, DESeq programs, RPKM / FPKM measure, idea of splice-junction search and parsing into isoforms, search for differentially expressed genes)
6. Normalization of RNA-seq data. Motivation and difficulty. Quantile normalization. Household genes normalization. TMM, RLE, MRN methods.
7. Bisulfite sequencing
8. Method of chromatin immunoprecipitation, ChIP-seq

9. ChIP-seq data analysis (peak search (MACS2), functional analysis, motive search, sequence-LOGO visualization)
10.Analysis of ChIP-seq data for histone modifications (MACS2 broadpeak, NGS-plot program, segmentation using chromHMM)
11. Search for regulatory areas (DNase-seq, ATAC-seq), specifics of reading and processing data

12. "Good practices" of experiment design based on sequencing (cues, Input, spike-in) and analysis of the data obtained

13. Conservatism of the gene, evolution of the locus. Orthologues, paralogs, homologues. Analysis of paralogs in the human genome. BUSCO database.

Block 3. Bioinformatics and applications

3.1 Fundamentals of Molecular Biology

1. Central dogma of molecular biology
2. The structure of eukaryotic genes (structural (exon) and regulatory (promoter, enhancer, insulator) gene elements.
3. Pseudogenes, their classification. Processed pseudogenes. Mechanisms of functional action of processed pseudogenes.
4. Repetitive sequences in DNA. Tandem replays: microsatellites, minisatellites, and satellites. Diseases of trinucleotide repeat expansion.

5. Dispersed repeats: transposons and retrotransposons. Opening mobile elements. Alu family of repeats.

6. Coding and non-coding RNAs. lncRNA, miRNA. Regulation by miRNA. CLASH data.

7. Antisense interactions.

8. Mechanisms providing phenotypic differences in genetically identical cells of a multicellular organism

9. Stages of transcription and processing of RNA, levels of regulation of RNA expression

10. Basic mechanisms of regulation of transcription

11. Modification of DNA and chromatin; epigenetics

12.DNA methylation and its regulatory role

13. DNA packing in a cell, nucleosome structure

14. Modifications of histone proteins and their regulatory role, histone code

15. Transcription initiation factors

16. Post-translational protein modifications.

3.2 Basic cell biology

1. Cell theory. Modern postulates of cell theory.
2. The concept of stem cells. Embryonic stem cells.
3. Epigenetics of stem cells.
4. Technology of genetic knockout.
5. Cell reprogramming. Induced pluripotent cells.
6. Genetic and epigenetic features of reprogrammed somatic cells.
7. Application of reprogramming technology to study the mechanisms of diseases and the search for new methods of therapy
8. Using bioinformatics methods to develop criteria for reprogramming.
9. Cell cycle. Mitosis and meiosis
10. Potential of differentiation of stem and somatic cells.
11. Structure of mammalian cells
12. Functions of eukaryotic cell organelles

3.3 Genetic foundations of medicine

1. Genetic bases of diseases. Etiology (causes) of monogenic and multifactorial diseases. The role of genetic predisposition in the development of multifactorial diseases.
2. Contribution of epigenetic modifications to the development of diseases.
3. Gene therapy: history and current state of the issue.
4. Pharmacogenomics. Applicability criteria in medicine.
5. Genetic tests and analyzes in personalized medicine.
6. Types of gene mutations, their pathological effects. Chromosomal aberrations and disease examples.
7. Cancer diseases and their genetic causes.

8. Biochemical methods used to diagnose hereditary diseases and identify carriers of pathological genes.

9. Gene therapy of hereditary diseases through somatic cells (principles, methods, results).

10. Orphan diseases. Methods for identifying the genetic causes of orphan diseases.

11. Checking the significance of the coefficient of determination.

12. Assessment of genetic risk

13. Estimates of heritability from GWAS results

14. Creation of foundations for personalized medicine in the field of diagnosis and treatment of diseases.

15. Functional biomarkers (genes) for diagnosis and metagenomic analysis.

3.4 Metagenomics and Microbiology

1. The role of the human microbiota in maintaining the homeostasis of the body. Connection with diseases.

2. Diversity of human microbiota.

3. Modern methods of analyzing the diversity of the microbiome.

4. Metatranscriptomics, metaproteomics and metabolomics.

5. Functional metagenomics. New genes, microbial pathways, antibiotic resistance studies.

6. Classification of bacteria toxin-antitoxin systems, biotargets and mechanisms of action.

7. Probiotics - application, mechanisms of action, prospects for use.

8. Comparative genomics of bacteria of the human intestinal microbiota.

9. Bacterial transplantation.

3.5 Genetics and evolution

1. Factors of microevolution.
2. Effective population size.
3. Subdivided population. The Walund effect.
4. Models of migration.
5. Phylogenetic and phylogeographic analysis - similarities and differences.
6. Methods for describing the gene pool based on whole genome data
7. Types of genetic markers.
8. The main features of the structure of the world gene pool
9. Practical applications of population research in medicine and forensic science.
10. Analysis of ancient DNA.
11. The relationship of genetics and related sciences about population. Features of interdisciplinary research.

4. Machine learning

1. Machine learning tasks. The task of teaching with and without a teacher. Classification and regression problems. Examples of ML application in biology and genetics.

2. Metrics in machine learning problems. Cross-validation. Cross-validation problems on biological data.

3. Method of k-nearest neighbors. Distance functions in machine learning problems. Bias-variance tradeoff. Bias-variance tradeoff using KNN as an example.

4. Linear regression. Logistic regression. Regularization. L1 and L2 regularization. Elastic net.

5. Support vector machine. Kernel trick.

6. Decisive trees.

7. The concept of an ensemble. Bagging. Random subspace method. Random forest.

8. Methods for assessing the importance of features.

9. Stochastic gradient descent. Gradient boosting.

10. Neural networks. Universal approximator theorem. Backpropagation algorithm.

11. Convolutional neural networks. Recurrent neural networks.

12. Learning without a teacher. Clustering algorithms.

13. The problem of dimensionality reduction. PCA, T-SNE, UMAP. Limitations of methods.

14. Data preprocessing. Categorical signs and work with them