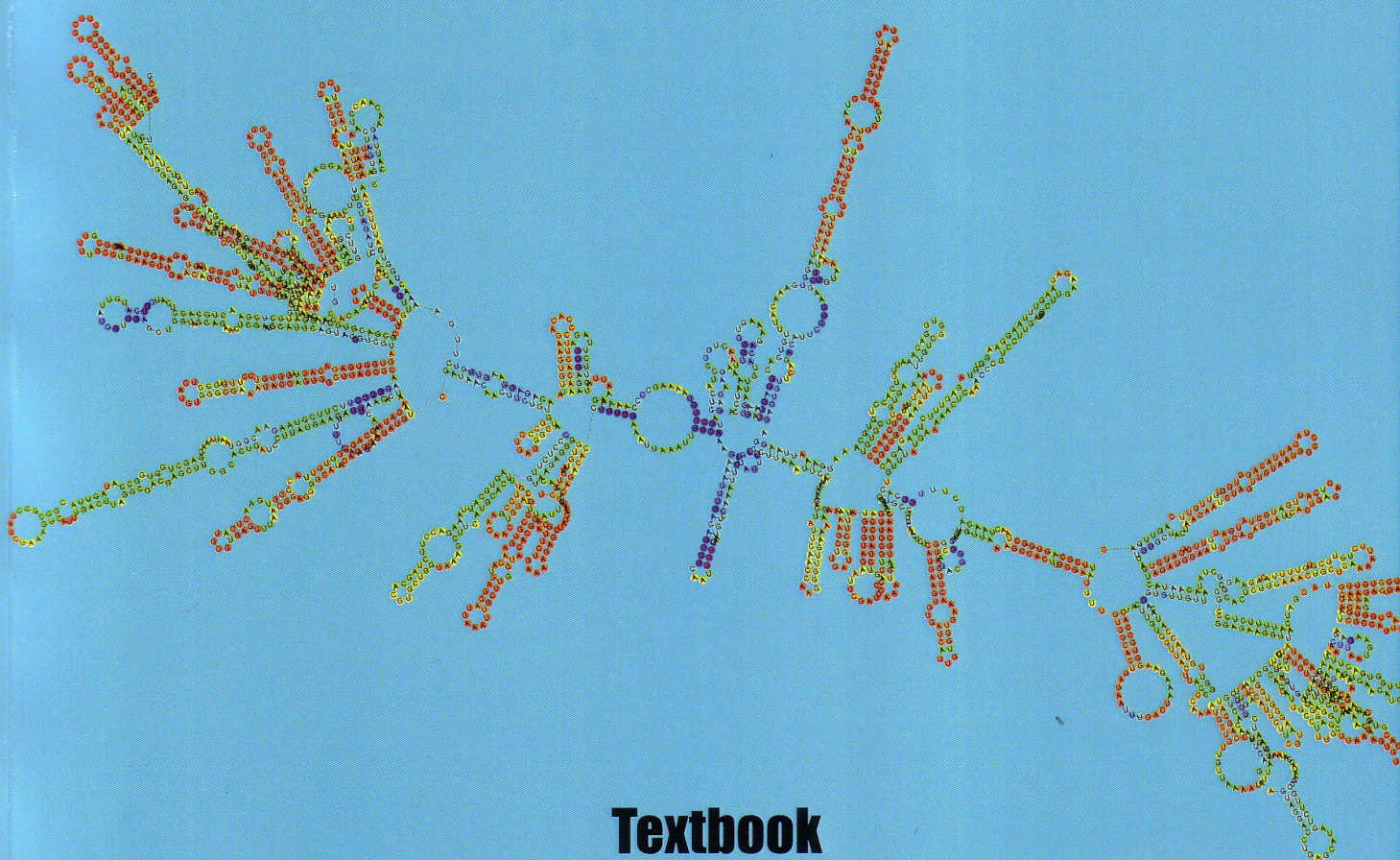


A.Yu. Pyrkova
A.T. Ivachshenko

BIOINFORMATICS **IN BIOTECHNOLOGY**



Textbook



КАЗАК
УНИВЕРСИТЕТИ
БАСПАҮЙІ

MINISTRY OF EDUCATION
AND SCIENCE OF THE REPUBLIC OF KAZAKHSTAN
AL-FARABI KAZAKH NATIONAL UNIVERSITY



QAZAQ
UNIVERSITY
PUBLISHING HOUSE

A.Yu. Pyrkova
A.T. Ivachshenko

BIOINFORMATICS IN BIOTECHNOLOGY

Textbook for educational program
6B05103 – Biotechnology

Almaty
«Qazaq University»
2020

UDC 60(075)
LBC 30.16я73
P 99

*Recommended to use in organizations of higher and postgraduate education
for students in the field of training of personnel «Biological and related sciences»
by Educational and Methodological Association of the Republican Educational
and Methodological Council on the basis of Al-Farabi KazNU.
Protocol No. 1 of the meeting of EMA REMC dated January 22, 2020.*

Reviewers:

doctor of biological sciences, professor,
head of laboratory of Cell Engineering of Institute of Plant Biology and Biotechnology

K.R. Uteulin

doctor of pedagogical sciences, professor of Computer Science Department
of Al-Farabi Kazakh National University

N.N. Kerimbaev

PhD of specialty 03.00.28 «Bioinformatics», leading researcher
of LLP «Scientific and Production Center of Microbiology and Virology»

A.S. Turmagambetova

Pyrkova A.Yu., Ivachshenko A.T.

P 99 Bioinformatics in biotechnology: textbook for educational program
6B05103 – Biotechnology / A.Yu. Pyrkova, A.T. Ivachshenko. – Almaty:
Qazaq University, 2020. – 149 p.

ISBN 978-601-04-4501-7

This textbook is intended for students of the educational program 6B05103 – Biotechnology and contains theoretical material required for students of the bachelor's degree to learn in the discipline «Bioinformatics in biotechnology», as well as is a guide to the performance of practical and control works. The textbook can also be used in the training of students of the master's degree of the educational program 7M05109 – Biotechnology.

**UDC 60(075)
LBC 30.16я73**

ISBN 978-601-04-4501-7

© Pyrkova A.Yu., Ivachshenko A.T., 2020
© Al-Farabi KazNU, 2020

CONTENTS

1. INTRODUCTION TO BIOINFORMATICS	8
1.1. Analysis of genetic sequences	9
1.2. Genomes annotation	10
1.3. Computational evolutionary biology	10
1.4. Assessment of biological diversity	10
1.5. Bioinformatics and computational biology	11
Questions and exercises	11
2. GREEDY ALGORITHMS	13
2.1. An activity-selection problem	13
2.2. When does greedy algorithm apply	15
2.2.1. The principle of greedy choice	15
2.2.2. Optimality for subproblems	16
2.3. Knapsack problem	16
2.3.1. Discrete knapsack problem	16
2.3.2. Fractional knapsack problem	16
2.4. Huffman codes	18
2.4.1. Variable-length code	18
2.4.2. Prefix codes	18
2.4.3. Constructing Huffman codes	20
Questions and exercises	21
3. DYNAMIC PROGRAMMING	23
3.1. Matrix-chain multiplication	23
3.1.1. Step 1: structure of optimal placement of parentheses	25
3.1.2. Step 2: Recurrent relation	25
3.1.3. Step 3: Calculate optimal cost	26
3.1.4. Step 4: Construct an optimal solution	29
3.2. Longest common subsequence	30
3.2.1. Structure of the longest common subsequence	30
3.2.2. Recurrent formula	31
3.2.3. Calculation of the LCS length	32
3.2.4. Construction of LCS	33
Questions and exercises	34
4. BIOLOGICAL DATABASES	35
4.1. Basis of database structures	35
4.2. Databases containing information on biological structures and sequences	36
4.2.1. Database record identifiers	36

4.2.2. Basic databases of biological polymer sequences	36
4.2.3. Protein structure databases	38
4.2.4. Protein structure classification databases	39
4.3. Basic bioinformatics data formats	39
4.3.1. Protein structure file formats	40
4.4. Known database catalog	41
Questions and exercises	44
5. ALIGNMENTS	45
5.1. Sequence Alignment. Introduction	45
5.2. Dotplot matrix of similarity	46
5.3. Dotplot matrices and sequence alignment	48
5.4. Sequence similarity measures	50
5.4.1. Evaluation schemes	51
5.4.1.1. Obtaining the substitution matrix	51
5.4.1.2. BLOSUM matrices	53
5.4.1.3. Weighing of inserts/deletions	53
5.5. Calculation of alignment for two sequences	53
5.5.1. Variations and generalizations	54
5.5.2. Approximate methods for fast database search	54
5.6. Dynamic programming algorithm to construct optimal pair alignment of sequences	55
5.7. Multiple sequence alignment	61
5.8. Mathematical model of multiple alignment of nucleotide and amino acid sequences	61
5.9. Multiple alignment algorithm	62
5.10. Phylogenetic Tree Construction Algorithm	65
5.11. Software implementation of multiple alignment algorithms and dendrograms on Java language using MPJ tools	67
Questions and exercises	74
6. GENOMICS	75
6.1. Bioinformatics in genomics	75
6.2. Genome sequencing	76
6.3. Genome annotation	76
6.4. Comparative genomics	78
6.5. Functional genomics	79
6.6. Parallelized algorithm for prediction of miRNA binding sites in mRNA – MirTarget Program	80
6.6.1. Mathematical model of scanning genes	80
6.6.2. Algorithm for scanning of genes on the cluster	81
6.6.3. Comparative analysis of the linear and parallelized algorithms for scanning genes	82
6.7. Tools from the "Sequence Manipulation Suite" program for decision genomics problems	84
Questions and exercises	87
7. PROTEOMICS	89
7.1. Bioinformatics in proteomics	89
7.2. Protein Identification through Database Searching	91
7.3. Tools from the "Sequence Manipulation Suite" program for decision proteomics problems	93
Questions and exercises	94

8. GRAPHS, PHYLOGENETIC TREES	95
8.1. Graphs	95
8.1.1. Bridge Obsession Problem	97
8.1.2. Eulerian Cycles.....	98
8.1.3. Hamiltonian Cycles	99
8.1.4. Shortest Path Problem.....	99
8.2. Phylogenetic trees.....	100
8.2.1. Clustering methods.....	103
8.2.1.1. Ultrametric trees	103
8.2.1.2. Agglomerative clustering.....	104
8.2.1.3. Neighbor Joining	106
8.2.2. Cladistic methods	110
8.2.3. Problem of variable rate of evolution	111
8.2.4. Computing analysis	112
Questions and exercises.....	113
9. DNA SEQUENCING.....	115
9.1. Shortest Superstring Problem	116
9.2. DNA Arrays as an Alternative Sequencing Technique.....	118
9.3. Sequencing by Hybridization.....	119
9.4. SBH as a Hamiltonian Path Problem.....	121
9.5. SBH as an Eulerian Path Problem	122
Questions and exercises.....	125
PRACTICAL TASKS	127
Practical task 1. Greedy algorithms	127
Practical task 2. Dynamic programming.....	129
Practical task 3. Phylogenetic trees and alignment	130
Practical task 4. Genomics.....	133
Practical task 5. Proteomics.....	134
Practical task 6. DNA sequencing	134
CONTROL TASKS.....	137
Control task 1	137
Control task 2	139
CONTROL QUESTIONS.....	143
GLOSSARY OF BIOLOGICAL TERMS.....	147
LIST OF USED SOURCES	148

LIST OF USED SOURCES

1. Lesk, A.M.: Introduction to Bioinformatics. – Oxford University Press, Oxford (2002).
2. Ignasimitu, S. Basics of Bioinformatics. – M.-Izhevsk: SRC "Regular and Chaotic Dynamics," Institute of Computer Research, 2007. – 320 p. (on Russian language)
3. Cormen, T.H., Leiserson, C.E., Rivest, R.L. Stein, C. Introduction to Algorithms. – Massachusetts Institute of Technology, USA, 2009.
4. Jones, N.C., Pevzner, P.A. An Introduction to Bioinformatics Algorithms. Massachusetts Institute of Technology Press, Massachusetts (2004) Hekmat Sh. Communication Networks. – PragSoft Corporation, 2005. – 164 p.
5. Leon, D., Markel, S. Sequence Analysis in a Nutshell. – O'Reilly & Associates, Inc., USA, 2003.
6. Niyazova R., Berillo O., Atambayeva Sh., Pyrkova A., Alybayeva A. and Ivashchenko A. MiR-1322 Binding Sites in Paralogous and Orthologous Genes // BioMed Research International. – 2015. – Article ID 962637. – 8 pages. <http://www.hindawi.com/journals/bmri/aa/962637>.
7. A. Ivashchenko, O. Berillo, A. Pyrkova, R. Niyazova, Sh. Atambayeva. The Properties of Binding Sites of miR-619-5p, miR-5095, miR-5096, and miR-5585-3p in the mRNAs of Human Genes // BioMed Research International. – Volume 2014 (2014). – Article ID 720715. – 8 pages. <http://dx.doi.org/10.1155/2014/720715>.
8. A. Ivashchenko, O. Berillo, A. Pyrkova, R. Niyazova Binding Sites of miR-1273 Family on the mRNA of Target Genes // BioMed Research International. – Volume 2014 (2014). – Article ID 620530. – 11 pages. <http://dx.doi.org/10.1155/2014/620530>.
9. S. Atambayeva, R. Niyazova, A. Ivashchenko, A. Pyrkova, I. Pinsky, A. Alimniyazova and S. Labeit The Binding Sites of miR-619-5p in the mRNAs of Human and Orthologous Genes // BMC Genomics. – Volume 18 (2017). – Iss. 1. – Article ID 428. – 10 pages. <https://bmcbgenomics.biomedcentral.com/track/pdf/10.1186/s12864-017-3811-6>.
10. Anatoly Ivashchenko, Olga Berillo, Anna Pyrkova, Raigul Niyazova, Shara Atambayeva The binding sites of unique miRNAs in the human mRNAs // Journal of Biotechnology // European Biotechnology Congress 2014. – Volume 185. – Supplement, September 2014. – Pages S54. <http://dx.doi.org/10.1016/j.jbiotec.2014.07.182>.
11. Raigul Niyazova, Olga Berillo, Shara Atambayeva, Anna Pyrkova, Almagul Ryskulova, Anatoly Ivashchenko Multiple miRNA binding sites in CDS of human mRNAs // Journal of Biotechnology// European Biotechnology Congress 2014. – Volume 185. – Supplement, September 2014. – Pages S54. <http://dx.doi.org/10.1016/j.jbiotec.2014.07.183>.
12. MirTarget Program – Certificate of Entering Information in the State Register of Rights to Objects Protected by Copyright. – No. 6598. – 2019. – November 22 on the software complex MirTarget version 1.0 (gene scanning) (authors: Pyrkova A.Yu., Ivashchenko A.T.).

Educational issue

Pyrkova Anna Yurevna,
Ivachshenko Anatoliy Timofeevich

BIOINFORMATICS IN BIOTECHNOLOGY

Textbook

Editor *L. Strautman*
Typesetting *G. Kaliyeva*
Cover design *Y. Gorbunov*

Cover design photos were used from sites
www.background-2672597_960_720.com

IB No.13389

Signed for publishing 10.03.2020. Format 70x100 1/12. Offset paper.
Digital printing. Volume 12,41 printer's sheet. 300 copies. Order No.8204.
Publishing house «Qazaq University»
Al-Farabi Kazakh National University
KazNU, 71 Al-Farabi, 050040, Almaty
Printed in the printing office of the «Qazaq University» publishing house.



9 786010 445017