Polina **Tikhonova**

Moscow, Russia

Scholar | tikhonova.polly@mail.ru | PollyTikhonova | polina-tikhonova |
orcid.org/0000-0003-2353-6070

About

Bioinformatician, data scientist. Currently working as a researcher in a metagenomics laboratory at a research institute in Moscow.
In the Research Experience section the major projects, I participated in, are outlined. Apart from them, I performed small data analysis and programming tasks which could not be treated as separate projects but some of them were used for publications.

Work & Education

10/2018 - present	**Researcher,** Laboratory of Genomic Studies and Computational Biology, Federal Research and Clinical Center of Physical-Chemical Medicine	*RCPCM*
2016 - 2018	**Master,** Faculty of Computer Science, "Data Analysis in Biology and Medicine", National Research University Higher School of Economics, Moscow, Russia	*HSE*
2012 - 2016	**Bachelor,** Faculty of Mathematics, National Research University Higher School of Economics, Moscow, Russia	*HSE*
2008 - 2012	**School,** Lyceum "School of Physics and Technology", Obninsk, Kaluga region, Russia	

Additional Education

2 term, 2017	**Course "Machine Learning",** Skoltech, Moscow	
1 term, 2015	**Course "Data Structures and Algorithms",** School of Data Analysis, Yandex, Moscow	
1 term, 2015	**Course "Python for Data Analysis",** National Research University Higher School of Economics	*HSE*

Research Experience

10/2020-present	**Metagenomics web-server,** metagenomics data analysis pipelines *metagenomics, 16S, NGS, data analysis, dada2, phyloseq, statistics, CoDa, Python, R, Django*	*RCPCM*
2020	**Internship,** Phylogenomics of Marinimicrobia *phylogenomics, ocean biology, phylogenetic tree, iTOL, checkM, MarkerFinder, FastTree* At the laboratory of Dr. Frank Aylward at Virginia Tech, remotely	*VT*
10/2019 - 02/2020	**HMG-proteins binding sites,** motifs, correlation with G-quadruplexes within TADs regions *quadruplexes, TADs, MEME, motifs, proteins, Fisher test, Monte-Carlo simulations, ATAC-seq, G4-seq, ChIP-seq, CTCF, bedtools, Python, R* In collaboration with a Laboratory of Artificial Antibodies Publication in process	*RCPCM*
10/2019	**ImGQFinder,** a Python tool for finding G/C-quadruplexes in a sequence *quadruplexes, CLI, Python* https://github.com/RCPCM-GCB/ImGQFinder	*RCPCM*
06/2019 - 08/2019	**Summer Internship ,** a family-based genome association analysis *GWAS, neurodevelopmental disorders, plink, family-based associations, TDT* At the laboratory of Dr. Elena Grigorenko at University of Houston, remotely	*UH*
06/2019 - 06/2020	**KnowYourHeart,** searching metabolic markers for different levels of alcohol consumption *metabolites, CoDa (compositional data), differential expression, robustness, statistics, correlation, t-test, Benjamini-Hochberg, fdr, PCA, MDS, Python, R* In collaboration with The London School of Hygiene & Tropical Medicine https://knowyourheart.science	*LSHTM, RCPCM*
10/2018 - 04/2019	**Phigaro,** a phage searching tool *prophages, metagenomic sequences, pVOGs, GC content, CLI, cross-validation, HMMER, Prodigal, What the Phage, Python, pypi, bioconda, Git Actions, ReadTheDocs, Docker* https://github.com/bobeobibo/phigaro	*RCPCM*
2017 - 2018	**Analysis of Mg-binding Sites in RNA Structures using ML methods,** Master Thesis *RNA biology, 3D structures, PyMol, machine learning, RandomForest, k-means, Python* https://pollytikhonova.github.io/coursework/	*HSE*

Publications

2020 **"Phigaro: high throughput prophage sequence annotation"**
Starikova E., Tikhonova P., Prianichnikov N., Rands C., Zdobnov E., Ilina E., Govorun V.
Bioinformatics, 36 (12), 3882-3884
doi: 10.1093/bioinformatics/btaa250 PMID: 32311023

2020 **"Evaluation of the Levels of Metabolites in Feces of Patients with Inflammatory Bowel Diseases"**
Zhgun E.S., Kislun Y.V., Kalachniuk T.N., Veselovsky V.A., Urban A.S., Tikhonova P.O., Pavlenko A.V., Ilchenko G.N., Ilina E.N. *Biomedical chemistry*, 66(3), 233-240
doi: 10.18097/PBMC20206603233 PMID: 32588829 [Russian version]
doi: 10.1134/S1990750820040113 [English version]

2020 **"The Effect of Intestinal Microbiome on the Effectiveness of Antitumor Immunotherapy"**
Olekhnovich E. I., Manolov A. I., Pavlenko A. V., Konanov D. N., Fedorov D. E., Tikhonova P. O., Glushchenko O. E., Ilina E. N. *Biomedical chemistry*, 66(1), 54-63
doi: 10.18097/PBMC20206601054 PMID: 32116226 [Russian version]
doi: 10.1134/S1990750820030105 [English version]

2019 **"Diversity of RNA pseudoknots exists only for short stems"**
Baulin E., Korinevskaya A., Tikhonova P., Roytberg M. *Mathematical Biology and Bioinformatics,* 2019. V. 14(S).
doi: 10.17537/2019.14.t37

Conferences

2019 **"Phigaro: a tool for phage detection in a bacterial sequences"**
Tikhonova P., Starikova E., Prianichnikov N., Rands C., Zdobnov E., Govorun V.
IX Russian Symposium "Proteins and Peptides", p.146, 2019, Sochi (Dagomys)
http://www.rusbiochem.org/files/uploaded/DAG2019_AbstractBook_Vol209122019.pdf

2018 **"Machine learning for Mg2+-binding sites prediction in RNA structures"**
Baulin E., Tikhonova P., Roytberg M. *Proceedings of the International Conference "Mathematical Biology and Bioinformatics".* Vol. 7. Pushchino: IMPB RAS, 2018. Paper No. e61.
doi: 10.17537/icmbb18.35

2017 **"Short stems in RNA secondary structure",**
Tikhonova P., Baulin E., Roytberg M.
Moscow Conference on Computational Molecular Biology (July) 27-30, 2017. Moscow
https://www.elibrary.ru/item.asp?id=32563281

Teaching

autumn, 2018 **Additional education "Introduction to Python"**
Moscow State University of Geodesy and Cartography

summer, 2018 **"Introduction to data analyses and machine learning methods",** Refresher Courses
SberBank

spring, 2018 **Course "Machine Learning": seminars,** MIPT
https://ml-mipt.github.io/

Non-academic activities

April 3-5, 2020 **EcoHackathon,** OpenRecycleTeam, the winners.
An Ecology Hackathon. Developed an application (chat-bot) for gathering information about hardly recycable types of wrappings.

1998 - 2012 **"Snowdrop",** classical ballet team, a dancer, a soloist.

Sona Hovsepyan Dance Company

04/17/16 **Time in the void,** modern ballet dance, a dancer
The laureate of the Student Festival "Festos-2016"

06/29/15 **The Pomegranate Taste,** a modern ballet play, a soloist
https://www.hse.ru/en/news/179552819.html

03/02/14 **The Human Voice,** a modern ballet play, a dancer

Polina **Tikhonova**

tikhonova.polly@mail.ru

Research Experience

November 24, 2020

I got my first research experience when I was studying at master's program "Data Analysis in Biology and Medicine" at Higher School of Economics. It was my master's degree, which was devoted to a part of structural RNA biology, specifically, to recognizing the Magnesium ion bindings using machine learning methods. At that work I tried to apply all new knowledge I have got at my master's program and an external machine learning course at SkolTech University. I needed to overcome problems related to data processing, choosing the datasets, features and models. Also, this project demanded from me to be creative. For example, at the case when I needed to compute the number of ions and its coordinates only based on information about predicted elements of RNA strains which were presumably placed close to the ions. And I invented an algorithm which iteratively computed k-means with different number of clusters till a coverage number (the percent of predicted ions covered by the current k-means) reaches a threshold. Finally, I was able to depict the ions and structures as 3D models using PyMOL.

Unfortunately, the results were not stunning. Poor quality of the datasets that we had could be one of the reasons. Especially, as far as I know, my thesis assistant and his student tried my model on their datasets and the results were pretty good. Moreover, when I am returning in my thoughts to this project and I think that one of the crucial features that we did not try was to include energetic modelling of the structures. We were tried to compute something at the very beginning, but I did not have enough resources at that time to develop energetic and machine learning models at the same time.

Also, I have two external research experiences. The first was a summer internship under a supervision of Oxana Naumova at Dr. Grigorenko's genetic lab at the University of Houston last year. Her laboratory researches the problems related to neurodevelopment disorders. It was a GWAS family-based study. It was very unusual and interesting experience for me, I have got to know about some statistical and other methods that allows to detect the significant SNPs.

Another external experience is a research with Dr. Frank Aylward from Biology Department of Virginia Tech. This work is an extension of the published by Dr. Aylward's Marinimicrobia research - a phylogenetic analyses of new Marinimicrobia sequences. I built a tree which combined the sequences from the published research and the new ones to understand if they lay as a separate cluster and then I will be working out metabolic features of new sequences.

The last but not least source of my research experience is my job. I work in a metagenomics laboratory in Federal Research and Clinical Center of Physical-Chemical Medicine. Among all microbiome related data analyses tasks it would be reasonable to highlight the three biggest projects.

The first one, was a development of a phage searching tool - Phigaro. When I joined this project, a first version of the tool was already developed. The tool is a pipeline consisted of two third-party programs (Prodigal and HMMER) and core computational algorithm which detects the prophage sequences. My task was to improve the algorithm's performance. Initially, the algorithm computed the phage score profile based on pVOGs, but this profile was too noisy. I proposed to add GC profile, which made prophage peaks much more distant. During this project I experienced all steps of Python tool development, including the publication, uploading the tool to PyPi, bioconda, Docker, auto testing and updates with Git

Actions and organization of comfortable documentation with ReadTheDocs.

The second research experience was a part of big long-term international project “Know Your Heart” (PI Dr. David Leon). The primary idea of the project was to find the reasons of high mortality in Russia. There was collected a huge dataset consisted of various types of data: microbiome, metabolome, blood samples, extensive metadata. Also, there was a subset of people with an increased alcohol consumption. An international team was organized with our colleges from The London School of Hygiene Tropical Medicine (England) to try to find some crucial differences in microbiome and metabolome datasets within people with different alcohol consumption (from zero to hazardous and harmful). I was a part of this team and metabolomics dataset was under my responsibility. I did an extensive data analyses including usage of CoDa (compositional data) technics, statistics methods (t-tests, robustness, FDR) and revealed some metabolites that could indicate the differences between alcohol consumption. As the next step we are going to combine the results of microbiome and metabolome analysis to build a single classification model.

Another project was devoted to protein binding motifs and their cooccurrence with G-quadruplexes within TAD (topologically associating domain) regions. This project was my collaboration with another laboratory in the institute - laboratory of Artificial Antibodies. My collaborators were very good biologists from wet lab, they were setting up an experiment with a group of proteins and needed a person who would help them to analyze the current situation. I was such a person. They had some developments from different years and projects, that I may need for my work. One of them was a tool for finding G-quadruplexes in a sequence (ImGQFinder), which they used for the similar project. But I found out that it was very hard to use this tool, because it has some serious limitations which could make the results different. So, before performing the statistical analysis I needed to fix these technical problems (rewrote the tool). Another little difficulty, that we met in this project, was the fact that we used the open data for the analysis – sequences of different cell lines that corresponded to different proteins. And for the sustainability of the results, of course, we needed to eliminate the possible noise that could occur due to different methods of post processing. So, I needed to find out all the details of the ATACseq process to be able accurately process raw data.

Right now, at work a team was created to build a huge web-system which will allow to combine and gather at one place all metadata and medical data from doctors and managers, get raw data from laboratories with all the characteristics and processes description that they have, and the analytical part, which will automatically calculate the raw data, and allows the user to apply some of our data analyses pipelines just by clicking on a couple of buttons and get the report. I am a member of the team, who leads the analytical part of the system. Thus, I am going to construct and implement some of the analysis pipelines, which I partly already constructed during my work with microbiome and metabolome datasets.