*BMC Bioinformatics* 2020, **21**(Suppl 20):567
https://doi.org/10.1186/s12859-020-03838-2

BMC Bioinformatics

MEETING ABSTRACTS Open Access

# Selected abstracts of "Bioinformatics: from Algorithms to Applications 2020" conference

Russia. 27-28 July 2020

Published: 17 December 2020

## I1

### Fourth International Conference "Bioinformatics: From Algorithms to Applications" (BiATA 2020)

Alla Lapidus[1,*], Anton Korobeynikov[1]
[1]Center for Algorithmic Biotechnologies, Saint Petersburg State University, Saint Petersburg, Russia, 199034

**Correspondence**: Alla Lapidus - a.lapidus@spbu.ru
*BMC Bioinformatics* 2020, **21(Suppl 20)**: I1

International Conference "Bioinformatics: from Algorithms to Applications" (BiATA) is one of the few international conferences that bring together both the programmers and algorithm developers creating tools for a wide spectrum of modern bioinformatics studies and the researchers conducting those experiments interested in finding reliable and easy to use tools for data analysis.
COVID-19 this year forced conferences online. Virtual conferences offer new experience that on a positive side allowed more participants to attend meetings that would be too hard or too expensive to attend in person. That's exactly what happened to BiATA-2020 that we had had to run remotely this year: we had a pleasure to welcome 475 scientists from all over the world versus 100 participants in previous years!
Follow our traditions, BiATA2020 promotes active application of bioinformatics in numerous fields of research, identifies new trends in the fields of bioinformatics, computational genomics and transcriptomics, as well as in discovery of biologically active molecules.
Topics covered within the framework of the conference include but are not limited to:

- Sequencing technologies
- Molecular sequence analysis
- Computational genomics
- Genome assembly
- Transcriptomics
- Metagenomics
- Agrigenomics
- Viromics
- Natural Products Discovery

The Fourth international conference "Bioinformatics: from Algorithms to Applications" was held on July 27–28, 2020 and was accompanied by the 2-day online workshop that included metagenomic data analysis and annotation using MGnify [1].

**Reference**

1 Mitchell AL, et al., MGnify: the microbiome analysis resource in 2020. Nucleic Acids Res. 2020 Jan 8;48(D1):D570–D578. https://doi.org/10.1093/nar/gkz1035

to promote elucidation of biological mechanisms which made this bacterial genus associated with public health risks.

## O4

### Do multiple long-distance transfers shape TBEV spread pattern?

Andrei A. Deviatkin[1,2*], Yulia A. Vakulenko[3,4], Ivan S. Kholodilov[5], Galina G. Karganova[5,6], Alexander N. Lukashev[1,3]
[1]Laboratory of Molecular Biology and Biochemistry, Institute of Molecular Medicine, Sechenov First Moscow State Medical University, 119048 Moscow, Russia; [2]Laboratory of Postgenomic Technologies, Izmerov Research Institute of Occupational Health, 105275 Moscow, Russia; [3]Martsinovsky Institute of Medical Parasitology, Tropical and Vector Borne Diseases, Sechenov First Moscow State Medical University, 119435 Moscow, Russia; [4]Department of Virology, Faculty of Biology, Lomonosov Moscow State University, 119234 Moscow, Russia; [5]Laboratory of Biology of Arboviruses, Chumakov Institute of Poliomyelitis and Viral Encephalitides (FSBSI "Chumakov FSC R&D IBP RAS), 108819 Moscow, Russia; [6]Department of Organization and Technology of Immunobiological Preparations, Institute for Translational Medicine and Biotechnology, Sechenov First Moscow State Medical University, 119991 Moscow, Russia

**Correspondence**: Andrei A. Deviatkin - andreideviatkin@gmail.com
BMC Bioinformatics 2020, **21(Suppl 20)**: O4

Tick-borne encephalitis (TBE) is viral zoonosis transmitted by the bite of infected ticks. In 1999, phylogenetic analysis demonstrated clear separation of TBE viruses into three subtypes, that were called based on its distribution: European, Siberian, and Far-Eastern. It is now becoming apparent that the actual spread of these viruses may differ from the nominal. Herein, 848 TBEV sequences (1028 nt E-gene fragments) were analyzed to indicate all long-distance virus transfers, that can be revealed from the sequence data. Threshold of 500 km was used for the selection of long-distance virus transfers. Noteworthy, ticks are not able to spread the infection on their own over such a distance. In other words, these long-distance virus transmissions were caused by vector-assisted tick transmission. In all subtypes and most of the smaller groups in these subtypes, there were a lot of recent long-distance virus transfers, that was revealed by Bayesian evolutionary analysis. Moreover, this is suggested to be a systematic pattern, rather than anecdotal events. For example, 19 out of 125 known sequences the Far-Eastern subtype were obtained in Japan. Genetic diversity of viruses found within this country was comparable with the diversity of the whole subtype. At the same time, this subtype is distributed throughout Japan, China, South Korea, Russia, Estonia and Latvia. The above arguments allow us to state that long transfers may be considered as a normal and abundant pattern in TBEV spreading.

**Acknowledgements**
This research was funded by the Russian Science Foundation (Grant # 19-75-00013).

## O5

### A rigorous approach to pairwise distance analysis of a protein family via multi-dimensional scaling and its application to the genealogy of squalene synthase paralogues of green algae

Robert B. Moore*[1], Michael Barnathan[2†], Brian Fristensky[3†], Yan Li[4,5†], Gregory Knowles[4†], Paul Gardner-Stephen[4], Angelo Bueti[4], Peter Anderson[4], Jianguang Qin[4], and Andrew S Ball[1]
[1]School of Science, RMIT University, Bundoora, Victoria 3083, Australia; [2]Temple University, Philadelphia, Pennsylvania 19122, USA; [3]Department of Plant Science, University of Manitoba, Winnipeg, Manitoba R3T 2N2, Canada; [4]School of Biological Sciences, Flinders University, Bedford Park, South Australia 5042, Australia; [5]Norwegian Institute for Water Research, Oslo NO-0349, Norway (current address)

**Correspondence**: Robert B. Moore - science@academicmail.org
*BMC Bioinformatics* 2020, **21(Suppl 20)**: O5

[†]Michael Barnathan, Brian Fristensky, Yan Li and Gregory Knowles authors contributed equally to this work.

**Background** Gene duplication resulting in paralogues is an imperfectly understood process in terms of the order of duplications facilitating the evolution of new functions. An example is the Squalene Synthase-Like (SSL) gene family of the oil-bearing green alga *Botryococcus braunii* race B. Squalene synthases combine two half reactions of catalysis, being the formation of presqualene disphosphate (PSPP) and then the subsequent synthesis of a triterpenoid, in this case squalene. It was previously established that the SSL paralogues have separated these two half reactions.
**Materials and methods** The squalene synthase (SS) and SSL genes of the organism were sequenced using Illumina reads and SOAP and Velvet assembly, in such a way that the rest of the genome was essentially ignored but this gene family was retrieved in complete detail. By this means the full set of paralogues were determined. Secondly, the genetic "distances" between four genes (as in-silico proteins) so recovered were compared pairwise with each other, as well as with the same set of genes published by a previous group, and with other green algal SS genes. Finally, the pairwise distance comparisons were input into a novel algorithm for Multi-Dimensional Scaling (MDS), which in combination with standard substitution matrices and a simple

averaging model for evolutionary rate of the genes, enabled a tree to be derived. Specifically, 5 dimensions were used in the MDS.
**Results** The order of evolution SS → SSL2 → SSL1 /SSL3 was determined, and inference of an evolutionary scenario was made. Further, an alignment process necessitated reannotation of key squalene synthases from green algal model organisms *Chlamydomonas* and *Volvox*, with support also obtained from the homologous proteins of non-model green algae *Micromonas* and *Ostreococcus*.
**Conclusions** Gene *B. braunii* SS, also known as BSS, has diverged further than a typical green algal SS, because selection on BSS was relaxed through duplication to create SSL2 which has all the functions of BSS except that PSPP synthesis is down-regulated without disappearing. C-terminal analysis indicated that green algal SSs may be membrane associated via two transmembrane alpha-helices plus an additional putative anchoring region. By this C-terminal indicator, BSS and SSL2 proteins may be in the same compartment/organelle as each other, explaining the somewhat relaxed selection on each. By contrast SSL1 and SSL3 which together generate a triterpenoid isomer named botryococcene, are lacking the C-terminal transmembrane alpha helices when analysed bioinformatically. This may indicate they have migrated to a different compartment or are in some way separated from the site of squalene synthesis, that has facilitated their separate evolution. SSL1 and SSL3 have stochastically recombined with each other which may have also facilitated the evolution of their new combined function—botryocccene synthesis, and are predicted to exist as a heterodimer.

## O6

### Black cat in a dark room: search for new viruses in metagenomes

Yulia Yakovleva[#1,2*], Alexey Zabelkin[#3], Maria Skazina[2,4], Artyom Kaltovich[2], Dmitry Antipov[5,6], Mikhail Rayko[5,6]
[1]Department of Cytology and Histology, Saint Petersburg State University, Saint Petersburg, Russia; [2]Bioinformatics Institute, Saint Petersburg, Russia; [3]ITMO University, Saint Petersburg, Russia; [4]Department of Applied Ecology, Saint Petersburg State University, Saint Petersburg, Russia; [5]Center for Algorithmic Biotechnology, Saint Petersburg State University, Saint Petersburg, Russia; [6]Institute of Translational Biomedicine, Saint Petersburg State University, Saint Petersburg, Russia

**Correspondence**: Yulia Yakovleva - st041958@student.spbu.ru
*BMC Bioinformatics* 2020, **21(Suppl 20)**: O6

[#]Yulia Yakovleva and Alexey Zabelkin have contributed equally.

Detection of hidden viral diversity is a challenging task, which goes beyond the standard protocol of processing metagenomic data. Meanwhile, publicly available databases contain a large amount of metagenomic data—the promising source of novel viral genomes, which remains largely understudied. Here we present the new pipeline for detecting full-length viral genomes from assembled metagenomes.
Viral genomes represent cyclic or linear molecules with the ends containing repeated sequences. Both types could be recognized as cyclic sequences. We detect such contigs by searching repeats ranging from 50 to 200 bp using Knuth-Morris-Pratt algorithm. This algorithm takes linear time depending on the maximum length of the allowed repeat, which permits to process large amounts of data and reduce its dimensionality. We classify cyclic sequences as viral or non-viral based on predicted gene content using viralVerify tool. For each selected viral contig we identify the capsid and terminase genes based on HMM profiles. We aligned found protein sequences against the NCBI nr database with Diamond. The protein sequences, both queries and hits, belonging to each HMM profile were clustered with CD-HIT v4.8.1 (span 80%, identity 50%). The resulting centroid sequences were aligned using MAFFT v7.310 with default parameters, followed by phylogeny reconstruction using UPGMA and RAxML v8.2.11 separately. Clusters that do not contain any hits were classified as previously unknown. The completeness of viral contigs was inspected with viralComplete and CheckV. We tested our pipeline on assembled metagenomes from NCBI Assembly database. More than 170 Gb of data representing about 1300 metagenomes derived from seawater, soil and biofilms habitats were analyzed.
Our analysis revealed that the diversity of viruses is much greater than we know up to date. Hundreds of new viruses clusters were detected. For example, we identified 3 new representatives of the Siphoviridae and Podoviridae bacteriophage families from 10 biofilm-derived metagenomes. Our approach allows us to detect full-length viral genomes with a lower chance of false-positive results. In the future, the user of our pipeline can submit metagenome assemblies or raw reads to the input and receive annotated viral genomes from the data. Further analysis of metagenomes from other habitats is indispensable.
Project is available on GitHub: https://github.com/Yulia-Yakovleva/metavirome.

**Acknowledgments**
This work was supported by Saint Petersburg State University (project ID 51555639).

## O7

### The 3C criterion: Contiguity, Completeness and Correctness to assess de novo genome assemblies

Jose Arturo Molina-Mora[1*], Fernando García[1]
[1]Centro de Investigación en Enfermedades Tropicales y Facultad de Microbiología, Universidad de Costa Rica, Costa Rica