THE FEDERAL RESEARCH CENTER INSTITUTE OF CYTOLOGY AND GENETICS
SIBERIAN BRANCH OF THE RUSSIAN ACADEMY OF SCIENCES

THE 7th INTERNATIONAL YOUNG SCIENTISTS SCHOOL "SYSTEMS BIOLOGY AND BIOINFORMATICS" SBB-2015

Abstracts

Novosibirsk, Russia
June 22–25, 2015

International Program Committee

Prof. Nikolay A. Kolchanov	Academician RAS, Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia (Chairman)
Prof. Klaus Mayer	Helmholtz Zentrum Muenchen, Neuherberg, Germany
Prof. Inna Lavrik	Otto von Guericke University, Magdeburg, Germany
Prof. Evgenii Rogayev	University of Massachusetts, USA
Dr. Maxim Zakhartsev	Hohenheim University, Institute of Plant Physiology and Biotechnology, Germany
Dr. Yuriy Orlov	Novosibirsk State University, Russia
Prof. Dmitriy Sherbakov	Limnological institute, Irkutsk, Russia
Dr. Irina Abnizova	Wellcome Trust Sanger Institute, UK
Prof. Rene te Boekhorst	University of Hertfordshire, UK
Prof. Ancha Baranova	George Mason University, USA
Prof. Haihua Bai	Inner Mongolia University for the Nationalities, China
Prof. Guoliang Li	Huazhong Agricultural University, Wuhan, China
Prof. Ming Chen	Zhejiang University, Hangzhou, China

Local Organizing Committee

Ilya Akberdin, PhD
Dmitry Afonnikov, PhD
Nadezhda Glebova
Tatyana Karamysheva, PhD
Galina Kiseleva, PhD
Ekaterina Kulakova
Andrey Kharkevich
Sergey Lashin, PhD
Victoria Mironova
Anna Onchukova
Yuriy Orlov, PhD
Olga Popik
Erlan Tokpanov
Svetlana Zubova (chair)

Contacts

The Federal Research Center Institute of Cytology and Genetics
The Siberian Branch of the Russian Academy of Sciences
RUSSIAN FEDERATION
630090, Novosibirsk
Prospekt Lavrentyeva 10
Tel: +7(383) 363-49-80
Fax: +7(383) 333-12-78
E-mail: sbb2015@ig.sbras.ru
URL: http://conf.nsc.ru/SBB2015/ru

The School was financially supported from Russian Science Foundation grant "Programmed cell death induced via death receptors: Delineating molecular mechanisms of apoptosis initiation via molecular modeling" № 14-44-00011.

ISBN 978-5-91291-023-4

MSCANNER - A GENOME-WIDE MOTIF FINDING TOOL

D.E. Romanov
Academy of Biology and Biotechnology, Rostov-on-Don, Russia
e-mail: rdme@ya.ru

Key words: motif finding, genome-wide motif finding.

Motivation and Aim: There are many motif finding algorithms each with its strong points wherther speed or accuracy. But it is often needed not just find motif, but also compare its location with location of other elements of a sequence, like genes, introns or promoters. The using of various custom formats of data representation hampers such kind of the research.

To illustrate this issue, lets consider the following list of motif finding software:

- GLAM2SCAN - a tool for finding occurrences of a gapped motifs in a sequence database [1];
- MAST - a tool for searching biological sequence databases for sequences that contain an occurrence of each motif in a given set of motifs [2];
- miRanda - an algorithm for the detection of potential microRNA target sites in genomic sequences [3].

These programs allow to find any kind of motifs, whereas miRanda was initially created to detect microRNA target sites. The problem here is that every of these programs uses its own custom input data format and produces output in its own custom format. Despite of input and output format of these programs being human readable, it makes complicated an automatic data processing.

Methods and Algorithms: To address this issues the software package **mscanner** was inroduced. **mscanner** reads motif sequences and genomic DNA sequences from input files and writes search results into the output file. Motif sequences should be in FASTA format and genomic sequences should be either in GENBANK or FASTA format. Output file is always written in GENBANK format.

By now only wrapping for GLAM2SCAN is implemented. Algorithm of finding motifs is made according the following procedure (quotation from GLAM2 software package manual): «By alternately applying glam2 and glam2mask several times, it is possible to find the strongest, second-strongest, third-strongest, etc. motifs in a set of sequences».

Another problem is that none of aforementioned motif finding programs (GLAM2SCAN, MAST and miRanda) could perform a whole genome motif search because of running out of system resources, in particular, the memory. To circumvent this issue the «divide and conquer» algorithm was applied. Initial long sequence was splitted into numerous overlapping fragments with length small enough to perform the search. Every fragment was searched for the occurences of the motifs and finding results were assembled then in a single output file. Such technique allows to study sequences with virtually arbitrary length.

Results: In order to verify the correctness of the proposed software pipeline the sequence of the human genome was scanned for occurrences of motifs of Alu elements. The computational results showed high consistency with publicly available experimental results of fluorescence *in situ* hybridisation of human chromosomes with probes for Alu elements.

It is worth noting, that the using of the standart data formats allows the researcher to *easily* watch the results in genome browsers, or to parse output files with help of such bioinformatic software packages, as BioPython or BioJava.

Availability: Available on request from the authors.

References

1. M.C. Frith, N.F.W. Saunders, B. Kobe, T.L. Bailey (2008) Discovering sequence motifs with arbitrary insertions and deletions, PLoS Computational Biology, 4(5): e1000071.
2. T.L. Bailey, M. Gribskov (1998) Combining evidence using p-values: application to sequence homology searches, Bioinformatics, 14(1): 48-54.
3. A.J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, D.S. Marks, (2003) MicroRNA targets in Drosophila,