GATB: Toolbox for developing efficient NGS software
Erwan Drezen, G Rizk, R Chikhi, Charles Deltel, C Lemaitre, P Peterlongo,
D Lavenier

To cite this version:
Erwan Drezen, G Rizk, R Chikhi, Charles Deltel, C Lemaitre, et al.. GATB: Toolbox for developing efficient NGS software. 9th Brazilian Symposium on Bioinformatics, BSB 2014, Oct 2014, Belo Horizonte, Brazil. hal-01088828

HAL Id: hal-01088828
https://hal.archives-ouvertes.fr/hal-01088828
Submitted on 4 Dec 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Improve the Memory Usage for de Bruijn Graphs

K. Salikhov, G. Sacomoto, G. Kucherov


G. Collet, G. Rizk, R. Chikhi, D. Lavenier

based on a Bloom filter

R. Chikhi, G. Rizk.

usage

G. Rizk, D. Lavenier, R. Chikhi,

positive answers

data structure) and by using a CFP additional structure that avoids false

GATB makes this graph compact by using

Strength of GATB

The core data structure of GATB is a
de Bruijn graph that encodes the main
information from the sequencing reads.

Strength of GATB

GATB makes this graph compact by using

a Bloom filter (a space efficient probabilistic
data structure) and by using a CFP additional structure that avoids false
positive answers from the Bloom filter due to its probabilistic nature.

3. Compact de Bruijn graph data structure

The core data structure of GATB is a
de Bruijn graph that encodes the main
information from the sequencing reads.

Strength of GATB

GATB makes this graph compact by using

a Bloom filter (a space efficient probabilistic
data structure) and by using a CFP additional structure that avoids false
positive answers from the Bloom filter due to its probabilistic nature.

4. Workflow

Here is a typical workflow when working with GATB

GATB-CORE transforms the reads into a de Bruijn graph,
saves it in a HDF5 file that can be opened by other tools
developed with the GATB-CORE API.

5. GATB helps you as a NGS user

GATB’s de Bruijn graph: a basis for families of tools

► Data error correction
► Assembly
► Biological motif detection

Several tools based on GATB are already available

Bioocoo K-mer spectrum based read error corrector for large datasets

Minia Short read assembler based on a de Bruijn graph. Results are
of similar contiguity and accuracy to other de Bruijn assemblers (e.g. Velvet)

DiscoSNP Discover Single Nucleotide Polymorphism (SNP) from non-
assembled reads

TakeABreak Detects inversion breakpoints without a reference genome by
looking for fixed size topological patterns in the de Bruijn graph

6. GATB helps you as a NGS developer

The GATB C++ library gives you the opportunity to quickly
develop new NGS tools that fit your needs.

Major facts about the GATB C++ library

► Object Oriented Design
► Simple and powerful graph API
► Simple and powerful multithreading model
► HDFS usage for data storage
► Fully documented with numerous code samples
► Complete test suite

License & Web Site

GATB is released under the GNU Affero General Public License.

Proprietary licencing for software editors or services providers is currently being studied.

For more details on GATB:
http://gatb.inria.fr

Partners

Publications


R. Chikhi, G. Rizk: Space-efficient and exact de Bruijn graph representation based on a Bloom filter. Algorithms for Molecular Biology 2013, 8:22


Erwan Drézen1, Guillaume Rizk1, Rayan Chikhi2, Charles Deltel3, Claire Lemaitre1, Pierre Peterlongo1 and Dominique Lavenier1

1 INRIA/IRISA/GenScale, Campus de Beaulieu, 35042 Rennes cedex 2 Department of Computer Science and Engineering, Pennsylvania State University, USA

1 9th Brazilian Symposium on Bioinformatics, BSB 2014

Tool Box for developing efficient NGS software

9th Brazilian Symposium on Bioinformatics, BSB 2014

http://gatb.inria.fr

For more details on GATB:
http://gatb.inria.fr