Gaëtan BENOIT

Postdoctoral researcher in bioinformatics

7 rue de la Piquetière
35700, Rennes, FR
+33 6 72 08 80 73
gae.benoit@gmail.com
Personal website

Current position

2023-Present **Postdoctoral Fellow with Karel Brinda (research engineer)**, *Inria*, Rennes, France
- Exploring the computational challenges of resistance diagnostics using ultra-fast nearest neighbor identification among previously characterized genomes.

Education

2014-2017 **PhD in computer science**, *Université Rennes 1*, Rennes, France
- Large-scale *de novo* comparative metagenomics. Supervised by Claire Lemaitre.

2012-2014 **MSc in computer science**, *Université Rennes 1*, Rennes, France
- Software engineering, machine learning, data modeling and indexing.

2010-2012 **BSc in software engineering**, *Université Rennes 1*, Rennes, France

2008-2010 **HND in electrical and industrial computer engineering**, *IUT*, Poitiers, France

Experience

2021-2023 **Postdoctoral Fellow with Christopher Quince**, *Earlham Institute*, Norwich, UK
- Developed "metaMDBG", a fast and low-memory C++ assembler for HiFi metagenomics data.

2018-2021 **Independent video-game developer**, Lille, France
- Created "Eat'n Eaten", a strategy game based on ecosystem simulation.

2014-2017 **Graduate researcher with Claire Lemaitre**, *Inria*, Rennes, France
- Developed tools in C++ and python to quickly compute ecological distances between numerous metagenomics datasets without *a priori* knowledge.

2016 **3 months visiting predoctoral fellow with Robert Finn**, *EMBL-EBI*, Hinxton, UK
- Investigated how *de* novo comparative metagenomics fits within EBI metagenomic portal.

2014 **6 months intern with Guillaume Rizk**, *Inria*, Rennes, France
- Developed a read compressor in C++ based on probabilistic de-Bruijn graph.

2013 **3 months intern with Guillaume Rizk**, *Inria*, Rennes, France
- Developed a memory efficient read corrector in C++ based on bloom filter.

Technical skills

- Software development
- C++
- Python
- Git, GitHub
- Conda, Bioconda
- Unix, bash/shell

Bioinformatics skills

- Metagenomics
- Assembly graph
- R
- Minimizer
- K-mer sketches
- Bloom filter

Projects

2021-Present **MetaMDBG: a scalable assembler for HiFi metagenomics reads.**

- Developed "metaMDBG" a fast and low-memory C++ assembler of HiFi metagnomics data using the minimizer de-Brujin graph.
- Developed the first implementation of the minimizer de-Bruijn graph in C++.
- Devised new strategies to estimate organims abundance in the graph.
- Created novel assembly methods based on abundance filters, allowing to remove complexity in the graph created by sequencing errors, inter-genomic repeats, and strain variability.
- Assembled hundreds of species in a single circular contig in different environments (human gut, anaerobic digester and sheep gut).
- GitHub: github.com/GaetanBenoitDev/metaMDBG

2018-2021 **Eat'n Eaten: an ecosystem simulator video-game.**

- Created "Eat'n Eaten", an ecosystem sim game, developed in C#, powered by Unity engine.
- Adapted ecosystem processes, such as the food chain and species evolution, in a fun and educational way.
- Developed for multiple platforms (desktop and mobile).
- Collaborated with an artist based in Portugal.
- Received audience award at FLIP 2019 (Festival Ludique International de Parthenay).
- Published on Steam: store.steampowered.com/app/1263650/Eatn_Eaten/

2014-2017 **Simka and SimkaMin: large-scale *de novo* comparative metagenomics tools.**

- Developed "simka" in C++, a software that compute traditional ecological distances by replacing species counts by k-mer counts.
- Developed the first scalable and exact k-mer counting algorithm of multiple datasets, using a fully parallelized disk-based sort merge algorithm.
- Developed "simkaMin" in C++, a very fast estimator of the Jaccard and Bray-Curtis distances based on k-mer subsampling, using minhash sketches.
- Designed software for parallel processing across a cluster (custom job scheduler written in Python).
- Created R scripts to visualize distance matrix as heatmap, hierarchical clustering and PCoA.
- Worked in collaboration with bioanalysts from Genoscope and Tara Oceans consortium. (Richter *et al.* 2022).
- Compared hundreds of plankton samples from the Tara Oceans project on Curie supercomputer (TGCC).
- Used to compare samples containing the most ancient DNA ever sequenced (Kjær *et al.* 2022).
- GitHub: github.com/GaetanBenoitDev/simka

2014 **Leon: read compressor based on probabilistic de-Bruijn graph.**

- Developed "leon" in C++, a compressor of Illumina genomic short-reads data.
- Created a *de novo* reference genome from raw data as a de-Bruijn graph.
- Compacted read coverage by reprentating reads as position in the reference graph and potential variations.
- Stored the reference in a bloom filter to achieve low-memory footprint and reference compression.
- Compressed the final representation using context model and arithmetic encoder.
- GitHub: github.com/GaetanBenoitDev/leon

2013 **Bloocoo: memory-efficient read corrector.**

- Developed "bloocoo" in C++, a corrector of Illumina short-read data.
- Used k-mer counting to distinguish genomic from erroneous k-mers.
- Used a bloom filter to query genomic k-mers in a lightweight fashion.
- Implemented several correction algorithms able to handle false positives from the bloom filter.
- GitHub: github.com/GaetanBenoitDev/bloocoo

Publications

2023 **Benoit G**, Chikhi R, Quince C. Assembling Long and Accurate Metagenomic Reads with Minimizer de Bruijn Graphs. *Accepted in Nature Biotechnology.*

2022 Richter DJ, Watteaux R, Vannier T, Leconte J, Frémont P, Reygondeau G, Maillet N, Henry N, **Benoit G**, ..., Jaillon O. Genomic evidence for global ocean plankton biogeography shaped by large-scale current systems. *Elife.*

2020 **Benoit G**, Mariadassou M, Robin S, Schbath S, Peterlongo P, Lemaitre C. SimkaMin: fast and resource frugal de novo comparative metagenomics. *Bioinformatics.*

2016 **Benoit G**, Peterlongo P, Mariadassou M, Drezen E, Schbath S, Lavenier D, Lemaitre C. Multiple comparative metagenomics using multiset k-mer counting. *PeerJ Computer Science.*

2015 **Benoit G**, Lemaitre C, Lavenier D, Drezen E, Dayris T, Uricaru R, Rizk G. Reference-free compression of high throughput sequencing data with a probabilistic de Bruijn graph. *BMC bioinformatics.*

Conference presentations

2023 EBAME, Brest, Long-read metagenomics assembly. Lecture and demo.

2017 JOBIM, Lille, Simka: large-scale *de novo* comparative metagenomics. Oral presentation.

2023 RCAM, Paris, Fast kmer-based method for estimating the similarity between numerous metagenomic datasets. Oral presentation.

2015 EBAME, Brest, Fast kmer-based method for estimating the similarity between numerous metagenomic datasets. Oral presentation.

2015 JOBIM, Clermont-Ferrand, Fast kmer-based method for estimating the similarity between numerous metagenomic datasets. Poster presentation. Best poster award.

2014 ECCB, Strasbourg, Bloocoo, a memory efficient read corrector. Poster presentation.

Teaching experience

2016-2017 **Teaching fellow**, *Université Rennes 1*, Rennes, France
Introductory to Python programming language (64h).

2015-2016 **Teaching fellow**, *Université Rennes 1*, Rennes, France
Algorithms for bioinformatics. (64h)

2015-2016 **Guest Lecturer and Consultant**, *Maison pour la science - Université Rennes 1*, Rennes, France
Indexing and querying large music database, Application to Shazam app (32h).