RESEARCH CENTRE

Grenoble - Rhône-Alpes

IN PARTNERSHIP WITH:

Université Claude Bernard (Lyon 1), Institut national des sciences appliquées de Lyon, Centrum Wiskunde & Informatica, Université de Rome la Sapienza

2021 ACTIVITY REPORT

Project-Team ERABLE

European Research team in Algorithms and Biology, formaL and Experimental

IN COLLABORATION WITH: Laboratoire de Biométrie et Biologie Evolutive (LBBE)

DOMAIN

Digital Health, Biology and Earth

THEME

Computational Biology

Contents

Project-Team ERABLE

Creation of the Project-Team: 2015 July 01

Keywords

Computer sciences and digital sciences

A3. – Data and knowledge

A3.1. – Data

A3.1.1. – Modeling, representation

A3.1.4. – Uncertain data

A3.3. – Data and knowledge analysis

A3.3.2. – Data mining

A3.3.3. – Big data analysis

A7. – Theory of computation

A8.1. – Discrete mathematics, combinatorics

A8.2. – Optimization

A8.7. – Graph theory

A8.8. – Network science

A8.9. – Performance evaluation

Other research topics and application domains

B1. – Life sciences

B1.1. – Biology

B1.1.1. – Structural biology

B1.1.2. – Molecular and cellular biology

B1.1.4. – Genetics and genomics

B1.1.6. – Evolutionnary biology

B1.1.7. – Bioinformatics

B1.1.10. – Systems and synthetic biology

B2. – Health

B2.2. – Physiology and diseases

B2.2.3. – Cancer

B2.2.4. – Infectious diseases, Virology

B2.3. – Epidemiology

1 Team members, visitors, external collaborators

Research Scientists

- Marie-France Sagot [Team leader, Inria, Senior Researcher, HDR]
- Laurent Jacob [CNRS, Researcher]
- Solon Pissis [CWI, Researcher]
- Blerina Sinaimeri [Inria, Researcher]
- Leen Stougie [CWI, Senior Researcher]
- Alain Viari [Inria, Senior Researcher]

Faculty Members

- Hubert Charles [INSA Lyon, Professor, HDR]
- Roberto Grossi [Université de Pise - Italie, Professor]
- Giuseppe Francesco Italiano [Université internationale libre d'études sociales Guido Carli - Italie, Professor]
- Vincent Lacroix [Univ Claude Bernard, Associate Professor]
- Alberto Marchetti-Spaccamela [Sapienza Universita di Roma, Professor]
- Arnaud Mary [Univ Claude Bernard, Associate Professor]
- Sabine Peres [Univ Claude Bernard, Professor, from Oct 2021, HDR]
- Nadia Pisanti [Université de Pise - Italie, Associate Professor]
- Cristina Vieira [Univ Claude Bernard, Professor, HDR]

Post-Doctoral Fellows

- Audric Cologne [CNRS]
- Scheila Mucha [Inria]

PhD Students

- Marianne Borderes [MaaT Pharma Lyon, CIFRE, Until July 2021]
- Nicolas Homberg [Institut national de recherche pour l'agriculture, l'alimentation et l'environnement]
- Maxime Mahout [Univ Paris-Saclay]
- Luca Nesterenko [CNRS, From Oct 2021]
- Camille Sessegolo [Univ Claude Bernard, until Oct 2021]
- Antoine Villie [CNRS]
- Yishu Wang [Univ Claude Bernard, until Sep 2021]

Interns and Apprentices

- Konstantinn Bonnet [Univ Claude Bernard, from Feb 2021 until Jul 2021]
- Eric Cumunel [Univ Claude Bernard, until Nov 2021]
- Johanna Trost [CNRS]

Administrative Assistant

- Anouchka Ronceray [Inria]

External Collaborator

- Susana Vinga [Instituto Superior Técnico / Université de Lisbonne]

2 Overall objectives

Cells are seen as the basic structural, functional and biological units of all living systems. They represent the smallest units of life that can replicate independently, and are often referred to as the building blocks of life. Living organisms are then classified into unicellular ones – this is the case of most bacteria and archea – or multicellular – this is the case of animals and plants. Actually, multicellular organisms, such as for instance human, may be seen as composed of native (human) cells, but also of extraneous cells represented by the diverse bacteria living inside the organism. The proportion in the number of the latter in relation to the number of native cells is believed to be high: this is for example of 90% in humans. Multicellular organisms have thus been described also as "superorganisms with an internal ecosystem of diverse symbiotic microbiota and parasites" (Nicholson *et al.*, Nat Biotechnol, 22(10):1268-1274, 2004) where symbiotic means that the extraneous unicellular organisms (cells) live in a close, and in this case, long-term relation both with the multicellular organisms they inhabit and among themselves. On the other hand, bacteria sometimes group into colonies of genetically identical individuals which may acquire both the ability to adhere together and to become specialised for different tasks. An example of this is the cyanobacterium *Anabaena sphaerica* who may group to form filaments of differentiated cells, some – the heterocysts – specialised for nitrogen fixation while the others are capable of photosynthesis. Such filaments have been seen as first examples of multicellular patterning.

At its extreme, one could then see life as one collection, or a collection of collections of genetically identical or distinct self-replicating cells who interact, sometimes closely and for long periods of evolutionary time, with same or distinct functional objectives. The interaction may be at equilibrium, meaning that it is beneficial or neutral to all, or it may be unstable meaning that the interaction may be or become at some time beneficial only to some and detrimental to other cells or collections of cells. The interaction may involve living systems, or systems that have been described as being at the edge of life such as viruses, or else living systems and chemical compounds (environment). It also includes the interaction between cells within a multicellular organism, or between transposable elements and their host genome.

The application objective of ERABLE is, through the use of mathematical models and algorithms, to better understand such close and often persistent interactions, with a longer term aim of becoming able in some cases to suggest the means of controlling for or of re-establishing equilibrium in an interacting community by acting on its environment or on its players, how they play and who plays. This objective requires to identify who are the partners in a closely interacting community, who is interacting with whom, how and by which means. Any model is a simplification of reality, but once selected, the algorithms to explore such model should address questions that are precisely defined and, whenever possible, be exact in the answer as well as exhaustive when more than one exists in order to guarantee an accurate interpretation of the results within the given model. This fits well the mathematical and computational expertise of the team, and drives the methodological objective of ERABLE which is to substantially and systematically contribute to the field of exact enumeration algorithms for problems that most often will be hard in terms of their complexity, and as such to also contribute to the field of combinatorics in as much as this may help in enlarging the scope of application of exact methods.

The key objective is, by constantly crossing ideas from different models and types of approaches, to look for and to infer "patterns", as simple and general as possible, either at the level of the biological application or in terms of methodology. This objective drives which biological systems are considered, and also which models and in which order, going from simple discrete ones first on to more complex continuous models later if necessary and possible.

3 Research program

3.1 Two main goals

ERABLE has two main sets of research goals that currently cover four main axes. We present here the research goals.

The first is related to the original areas of expertise of the team, namely combinatorial and statistical modelling and algorithms, although more recently the team has also been joined by members that come from biology including experimental.

The second set of goals concern its main Life Science interest which is to better understand interactions between living systems and their environment. This includes close and often persistent interactions between two living systems (symbiosis), interactions between living systems and viruses, and interactions between living systems and chemical compounds. It also includes interactions between cells within a multicellular organism, or interactions between transposable elements and their host genome.

Two major steps are constantly involved in the research done by the team: a first one of modelling (*i.e.* translating) a Life Science problem into a mathematical one, and a second of algorithm analysis and design. The algorithms developed are then applied to the questions of interest in Life Science using data from the literature or from collaborators. More recently, thanks to the recruitment of young researchers (PhD students and postdocs) in biology, the team has become able to start doing experiments and producing data or validating some of the results obtained on its own.

From a methodological point of view, the main characteristic of the team is to consider that, once a model is selected, the algorithms to explore such model should, whenever possible, be exact in the answer provided as well as exhaustive when more than one exists for a more accurate interpretation of the results. More recently, the team has also become interested in exploring the interface between exact algorithms on one hand, and probabilistic or statistical ones on the other such as used in machine learning approaches, notably "interpretable" versions thereof.

3.2 Different research axes

The goals of the team are biological and methodological, the two being intrinsically linked. Any division into axes along one or the other aspect or a combination of both is thus somewhat artificial. Following the evaluation of the team at the end of 2017, four main axes were identified, with the last one being the more recently added one. This axis is specifically oriented towards health in general. The first three axes are: (pan)genomics and transcriptomics in general, metabolism and (post)transcriptional regulation, and (co)evolution.

Notice that the division itself is based on the biological level (genomic, metabolic/regulatory, evolutionary) or main current Life Science purpose (health) rather than on the mathematical or computational methodology involved. Any choice has its part of arbitrariness. Through the one we made, we wished to emphasise the fact that the area of application of ERABLE is important for us. *It does not mean that the mathematical and computational objectives are not equally important*, but only that those are, most often, motivated by problems coming from or associated to the general Life Science goal. Notice that such arbitrariness also means that some Life Science topics will be artificially split into two different Axes. One example of this is genomics and the main health areas currently addressed that are intrinsically inter-related.

Axis 1: (Pan)Genomics and transcriptomics in general

Intra and inter-cellular interactions involve molecular elements whose identification is crucial to understand what governs, and also what might enable to control such interactions. For the sake of

clarity, the elements may be classified in two main classes, one corresponding to the elements that allow the interactions to happen by moving around or across the cells, and another that are the genomic regions where contact is established. Examples of the first are non coding RNAs, proteins, and mobile genetic elements such as (DNA) transposons, retro-transposons, insertion sequences, etc. Examples of the second are DNA/RNA/protein binding sites and targets. Furthermore, both types (effectors and targets) are subject to variation across individuals of a population, or even within a single (diploid) individual. Identification of these variations is yet another topic that we wish to cover. Variations are understood in the broad sense and cover single nucleotide polymorphisms (SNPs), copy-number variants (CNVs), repeats other than mobile elements, genomic rearrangements (deletions, duplications, insertions, inversions, translocations) and alternative splicings (ASs). All three classes of identification problems (effectors, targets, variations) may be put under the general umbrella of genomic functional annotation.

Axis 2: Metabolism and (post)transcriptional regulation

As increasingly more data about the interaction of molecular elements (among which those described above) becomes available, these should then be modelled in a subsequent step in the form of networks. This raises two main classes of problems. The first is to accurately infer such networks. Assuming such a network, integrated or "simple", has been inferred for a given organism or set of organisms, the second problem is then to develop the appropriate mathematical models and methods to extract further biological information from such networks.

The team has so far concentrated its efforts on two main aspects concerning such interactions: metabolism and post-transcriptional regulation by small RNAs. The more special niche we have been exploring in relation to metabolism concerns the fact that the latter may be seen as an organism's immediate window into its environment. Finely understanding how species communicate through those windows, or what impact they may have on each other through them is thus important when the ultimate goal is to be able to model communities of organisms, for understanding them and possibly, on a longer term, for control. While such communication has been explored in a number of papers, most do so at a too high level or only considered couples of interacting organisms, not larger communities. The idea of investigating consortia, and in the case of synthetic biology, of using them, has thus started being developed in the last decade only, and was motivated by the fact that such consortia may perform more complicated functions than could single populations, as well as be more robust to environmental fluctuations. Another originality of the work that the team has been doing in the last decade has also been to fully explore the combinatorial aspects of the structures used (graphs or directed hypergraphs) and of the associated algorithms. As concerns post-transcriptional regulation, the team has essentially been exploring the idea that small RNAs may have an important role in the dialog between different species.

Axis 3: (Co)Evolution

Understanding how species that live in a close relationship with others may (co)evolve requires understanding for how long symbiotic relationships are maintained or how they change through time. This may have deep implications in some cases also for understanding how to control such relationships, which may be a way of controlling the impact of symbionts on the host, or the impact of the host on the symbionts and on the environment (by acting on its symbiotic partner(s)). These relationships, also called *symbiotic associations*, have however not yet been very widely studied, at least not at a large scale.

One of the problems is getting the data, meaning the trees for hosts and symbionts but even prior to that, determining with which symbionts the present-day hosts are associated. This means that at the modelling step, we need to consider the possibility, or the probability of errors or of missing information. The other problem is measuring the stability of the association. This has generally been done by concomitantly studying the phylogenies of hosts and symbionts, that is by doing what is called a *cophylogeny* analysis, which itself is often realised by performing what is called a *reconciliation* of two phylogenetic trees (in theory, it could be more than two but this is a problem that has not yet been addressed by the team), one for the symbionts and one for the hosts with which the symbionts are associated. This consists in mapping one of the trees (usually, the symbiont tree) to the other. Cophylogeny inherits all the difficulties of phylogeny, among which the fact that it is not possible to check the result against the "truth" as this is now lost in the past. Cophylogeny however also brings new problems of its own which are to estimate the frequency of the different types of events that could lead

to discrepant evolutionary histories, and to estimate the duration of the associations such events may create.

Axis 4: Health in general

As indicated above, this is a recent axis in the team and concerns various applications to human and animal health. In some ways, it overlaps with the three previous axes as well as with Axis 5 on the methodological aspects, but since it gained more importance in the past few years, we decided to develop more these particular applications. Most of them started through collaborations with clinicians. Such applications are currently focused on three different topics: (i) Infectiology, (ii) Rare diseases, and (iii) Cancer. A fourth topic started a few years ago in collaboration with researchers from different universities and institutions in Brazil, and concerns tropical diseases, notably related to *Trypanosoma cruzi* (Chagas disease). This topic will be developed more strongly from 2022 on, notably through the collaboration with Ariel Silber, full professor at the Department of Parasitology of the University of São Paulo, with whom we have projects in common, and since the middle of 2021 a PhD student in co-supervision with M.-F. Sagot from ERABLE. This student is Gabriela Torres Montanaro. Both Gabriela and Ariel will be visiting ERABLE at different occasions in 2022, sometimes for long periods especially in the case of Gabriela.

Among the other three topics, infectiology is the oldest one. It started by a collaboration with Arnaldo Zaha from the Federal University of Rio Grande do Sul in Brazil that focused on pathogenic bacteria living inside the respiratory tract of swines. Since our participation in the H2020 ITN MicroWine, we started interested in infections affecting plants this time, and more particularly vine plants. Rare Diseases on the other hand started by a collaboration with clinicians from the Centre de Recherche en Neurosciences of Lyon (CNRL) and is focused on the Taybi-Linder Syndrome (TALS) and on abnormal splicing of U12 introns, while Cancer rests on a collaboration with the Centre Léon Bérard (CLB) and Centre de Recherche en Cancérologie of Lyon (CRCL) which is focused on Breast and Prostate carcinomas and Gynaecological carcinosarcomas.

The latter collaboration was initiated through a relationship between a member of ERABLE (Alain Viari) and Dr. Gilles Thomas who had been friends since many years. G. Thomas was one of the pioneers of Cancer Genomics in France. After his death in 2014, Alain Viari took the (part time) responsibility of his team at CLB and pursued the main projects he had started.

Notice however that as concerns cancer, at the end of 2021 (October 1st), a new member joined the ERABLE team as full professor in the LBBE - University of Lyon, namely Sabine Peres. Sabine has also been working on cancer, in her case from a perspective of metabolism, in collaboration with Laurent Schwartz (Assistance Publique - Hôpitaux de Paris) and with Mario Jolicoeur, (Polytechnique Montréal, Canada).

Within Inria and beyond, the first two applications and the fourth one (Infectiology, Rare Diseases, and Tropical diseases) may be seen as unique because of their specific focus (resp. microbiome and respiratory tract of swines / vine plants on one hand, and TALS on the other). In the first case, such uniqueness is also related to the fact that the work done involves a strong computational part but also experiments that in some cases (respiratory tract of swines) is *performed within ERABLE itself*.

4 Application domains

4.1 Biology and Health

The main areas of application of ERABLE are: (1) biology understood in its more general sense, with a special focus on symbiosis and on intracellular interactions, and (2) health with a special emphasis for now on infectious diseases, rare diseases, cancer, and since more recently, tropical diseases notably related to *Trypanosoma cruzi*.

5 Social and environmental responsibility

5.1 Footprint of research activities

There are three axes on which we would like to focus in the coming years.

Travelling is essential for the team, that is European and has many international collaborations. We would however like to continue to develop as much as possible travelling by train or even car. This is something we do already, for instance between Lyon and Amsterdam by train, and that we have done in the past, such as for instance between Lyon and Pisa by car, and between Rome and Lyon by train, or even in the latter case once between Rome and Amsterdam!

Computing is also essential for the team. We would like to continue our effort to produce resource frugal software and develop better guidelines for the end users of our software so that they know better under which conditions our software is expected to be adapted, and which more resource-frugal alternatives exist, if any.

Having an impact on how data are produced is also an interest of the team. Much of the data produced is currently only superficially analysed. Generating smaller datasets and promoting data reuse could avoid not only data waste, but also economise on computer time and energy required to produce such data.

5.2 Expected impact of research results

As indicated earlier, the overall objective of the team is to arrive at a better understanding of close and often persistent interactions among living systems, between such living systems and viruses, between living systems and chemical compounds (environment), among cells within a multicellular organism, and between transposable elements and their host genome. There is another longer-term objective, much more difficult and riskier, a “dream” objective whose underlying motivation may be seen as social and is also environmental.

The main idea we thus wish to explore is inspired by the one universal concept underlying life. This is the concept of survival. Any living organism has indeed one single objective: to remain alive and reproduce. Not only that, any living organism is driven by the need to give its descendants the chance to perpetuate themselves. As such, no organism, and more in general, no species can be considered as “good” or “bad” in itself. Such concepts arise only from the fact that resources, some of which may be shared among different species, are of limited availability. Conflict thus seems inevitable, and “war” among species the only way towards survival.

However, this is not true in all cases. Conflict is often observed, even actively pursued by, for instance, humans. Two striking examples that have been attracting attention lately, not necessarily in a way that is positive for us, are related to the use of antibiotics on one hand, and insecticides on the other, both of which, especially but not only the second can also have disastrous environmental consequences. Yet cooperation, or at least the need to stop distinguishing between “good” (mutualistic) and “bad” (parasitic) interactions appears to be, and indeed in many circumstances is of crucial importance for survival. The two questions which we want to address are: (i) what happens to the organisms involved in “bad” interactions with others (for instance, their human hosts) when the current treatments are used, and (ii) can we find a non-violent or cooperative way to treat such diseases?

Put in this way, the question is infinitely vast. It is not completely utopic. We had the opportunity in recent years to discuss such question with notably biologists with whom we were involved in two European projects (namely BachBerry, and MicroWine). In both cases, we had examples of bacteria that are "bad" when present in a certain environment, and "good" when the environment changes. In one of the cases at least, related to vine plants, such change in environment seems to be related to the presence of other bacteria. This idea is already explored in agriculture to avoid the use of insecticide. Such exploration is however still relatively limited in terms of scope, and especially, has not yet been fully investigated scientifically.

The aim will be to reach some proofs of concepts, which may then inspire others, including ourselves on a longer term, to pursue research along this line of thought. Such proofs will in themselves already require to better understand what is involved in, and what drives or influences any interaction.

6 Highlights of the year

The research of all team members, in particular of PhD students or Postdocs, is important for us and we prefer not to highlight any in particular.

7 New software and platforms

We indicate in this section the new methods we developed in 2021 but also the older ones that continue to be used and that are being constantly maintained by the researchers of the team. This indeed represents a great part of our effort and time, and is important in general.

7.1 New software

7.1.1 BrumiR

Name: A toolkit for de novo discovery of microRNAs from sRNA-seq data.

Keywords: Bioinformatics, Structural Biology, Genomics

Functional Description: BRUMIR is an algorithm that is able to discover miRNAs directly and exclusively from sRNA-seq data. It was benchmarked with datasets encompassing animal and plant species using real and simulated sRNA-seq experiments. The results show that BRUMIR reaches the highest recall for miRNA discovery, while at the same time being much faster and more efficient than the state-of-the-art tools evaluated. The latter allows BRUMIR to analyse a large number of sRNA-seq experiments, from plant or animal species. Moreover, BRUMIR detects additional information regarding other expressed sequences (sRNAs, isomiRs, etc.), thus maximising the biological insight gained from sRNA-seq experiments. Finally, when a reference genome is available, BRUMIR provides a new mapping tool (BRUMIR2REFERENCE) that performs a posteriori an exhaustive search to identify the precursor sequences.

URL: https://github.com/camoragaq/BrumiR

Contact: Carol Moraga Quinteros

Participants: Carol Moraga Quinteros, Marie-France Sagot

7.1.2 Caldera

Keywords: Genomics, Graph algorithmics

Functional Description: CALDERA extends DBGWAS by performing one test for each closed connected subgraph of the compacted De Bruijn graph built over a set of bacterial genomes. This allows to test the association between a phenotype and the presence of a causal gene which has several variants. CALDERA exploits Tarone's concept of testability to avoid testing sequences which cannot possibly be associated with the phenotype.

URL: https://github.com/HectorRDB/Caldera_Recomb

Contact: Laurent Jacob

7.1.3 Capybara

Name: equivalence ClAss enumeration of coPhylogenY event-BAsed ReconciliAtions

Keywords: Bioinformatics, Evolution

Functional Description: Phylogenetic tree reconciliation is the method of choice in analysing host-symbiont systems. Despite the many reconciliation tools that have been proposed in the literature, two main issues remain unresolved: listing suboptimal solutions (*i.e.*, whose score is "close" to the optimal ones), and listing only solutions that are biologically different "enough". The first issue arises because the optimal solutions are not always the ones biologically most significant, providing many suboptimal solutions as alternatives for the optimal ones is thus very useful. The second one is related to the difficulty to analyse an often huge number of optimal solutions. Capybara addresses both of these problems in an efficient way. Furthermore, it includes a tool for visualising the solutions that significantly helps the user in the process of analysing the results.

URL: https://github.com/Helio-Wang/Capybara-app

Publication: hal-02917341

Contact: Yishu Wang

Participants: Yishu Wang, Arnaud Mary, Marie-France Sagot, Blerina Sinaimeri

7.1.4 C3Part/Isofun

Keywords: Bioinformatics, Genomics

Functional Description: The C3PART / ISOFUN package implements a generic approach to the local alignment of two or more graphs representing biological data, such as genomes, metabolic pathways or protein-protein interactions, in order to infer a functional coupling between them.

URL: http://www.inrialpes.fr/helix/people/viari/lxgraph/index.html

Contact: Alain Viari

Participants: Alain Viari, Anne Morgat, Frédéric Boyer, Marie-France Sagot, Yves-Pol Deniélou

7.1.5 Cassis

Keywords: Bioinformatics, Genomics

Functional Description: Implements methods for the precise detection of genomic rearrangement breakpoints.

URL: http://pbil.univ-lyon1.fr/software/Cassis/

Contact: Marie-France Sagot

Participants: Christian Baudet, Christian Gautier, Claire Lemaitre, Eric Tannier, Marie-France Sagot

7.1.6 Coala

Name: CO-evolution Assessment by a Likelihood-free Approach

Keywords: Bioinformatics, Evolution

Functional Description: COALA stands for "COevolution Assessment by a Likelihood-free Approach". It is thus a likelihood-free method for the co-phylogeny reconstruction problem which is based on an Approximate Bayesian Computation (ABC) approach.

URL: http://team.inria.fr/erable/en/software/coala/

Contact: Blerina Sinaimeri

Participants: Beatrice Donati, Blerina Sinaimeri, Catherine Matias, Christian Baudet, Christian Gautier, Marie-France Sagot, Pierluigi Crescenzi

7.1.7 CSC

Keywords: Genomics, Algorithm

Functional Description: Given two sequences x and y, CSC (which stands for Circular Sequence Comparison) finds the cyclic rotation of x (or an approximation of it) that minimises the blockwise q-gram distance from y.

URL: https://github.com/solonas13/csc

Contact: Nadia Pisanti

7.1.8 Cycads

Keywords: Systems Biology, Bioinformatics

Functional Description: Annotation database system to ease the development and update of enriched BIOCYC databases. CYCADS allows the integration of the latest sequence information and functional annotation data from various methods into a metabolic network reconstruction. Functionalities will be added in future to automate a bridge to metabolic network analysis tools, such as METEXPLORE. CYCADS was used to produce a collection of more than 22 arthropod metabolism databases, available at ACYPICYC (http://acypicyc.cycadsys.org) and ARTHROPODACYC (http://arthropodacyc.cycadsys.org). It will continue to be used to create other databases (newly sequenced organisms, Aphid biotypes and symbionts...).

URL: http://www.cycadsys.org/

Contact: Hubert Charles

Participants: Augusto Vellozo, Hubert Charles, Marie-France Sagot, Stefano Colella

7.1.9 DBGWAS

Keywords: Graph algorithmics, Genomics

Functional Description: DBGWAS is a tool for quick and efficient bacterial GWAS. It uses a compacted De Bruijn Graph (cDBG) structure to represent the variability within all bacterial genome assemblies given as input. Then cDBG nodes are tested for association with a phenotype of interest and the resulting associated nodes are then re-mapped on the cDBG. The output of DBGWAS consists of regions of the cDBG around statistically significant nodes with several informations related to the phenotypes, offering a representation helping in the interpretation. The output can be viewed with any modern web browser, and thus easily shared.

URL: https://gitlab.com/leoisl/dbgwas

Contact: Laurent Jacob

7.1.10 Eucalypt

Keywords: Bioinformatics, Evolution

Functional Description: EUCALYPT stands for "EnUmerator of Coevolutionary Associations in PoLYnomial-Time delay". It is an algorithm for enumerating all optimal (possibly time-unfeasible) mappings of a symbiont tree unto a host tree.

URL: http://team.inria.fr/erable/en/software/eucalypt/

Contact: Blerina Sinaimeri

Participants: Beatrice Donati, Blerina Sinaimeri, Christian Baudet, Marie-France Sagot, Pierluigi Crescenzi

7.1.11 Fast-SG

Keywords: Genomics, Algorithm, NGS

Functional Description: FAST-SG enables the optimal hybrid assembly of large genomes by combining short and long read technologies.

URL: https://github.com/adigenova/fast-sg

Contact: Alex Di Genova

Participants: Alex Di Genova, Marie-France Sagot, Alejandro Maass, Gonzalo Ruz Heredia

7.1.12 Gobbolino-Touché

Keywords: Bioinformatics, Graph algorithmics, Systems Biology

Functional Description: Designed to solve the metabolic stories problem, which consists in finding all maximal directed acyclic subgraphs of a directed graph G whose sources and targets belong to a subset of the nodes of G, called the black nodes.

URL: https://team.inria.fr/erable/en/software/gobbolino/

Contact: Marie-France Sagot

Participants: Etienne Birmelé, Fabien Jourdan, Ludovic Cottret, Marie-France Sagot, Paulo Vieira Milreu, Pierluigi Crescenzi, Vicente Acuña, Vincent Lacroix

7.1.13 HapCol

Keywords: Bioinformatics, Genomics

Functional Description: A fast and memory-efficient DP approach for haplotype assembly from long reads that works until 25x coverage and solves a constrained minimum error correction problem exactly.

URL: http://hapcol.algolab.eu/

Contact: Nadia Pisanti

7.1.14 HgLib

Name: HyperGraph Library

Keywords: Graph algorithmics, Hypergraphs

Functional Description: The open-source library hglib is dedicated to model hypergraphs, which are a generalisation of graphs. In an *undirected* hypergraph, an hyperedge contains any number of vertices. A *directed* hypergraph has hyperarcs which connect several tail and head vertices. This library, which is written in C++, allows to associate user defined properties to vertices, to hyperedges/hyperarcs and to the hypergraph itself. It can thus be used for a wide range of problems arising in operations research, computer science, and computational biology.

Release Contributions: Initial version

URL: https://gitlab.inria.fr/kirikomics/hglib

Contact: Arnaud Mary

Participants: Martin Wannagat, David Parsons, Arnaud Mary, Irene Ziska

7.1.15 KissDE

Keywords: Bioinformatics, NGS

Functional Description: KISSDE is an R Package enabling to test if a variant (genomic variant or splice variant) is enriched in a condition. It takes as input a table of read counts obtained from an NGS data pre-processing and gives as output a list of condition-specific variants.

Release Contributions: This new version improved the recall and made more precise the size of the effect computation.

URL: http://kissplice.prabi.fr/tools/kissDE/

Contact: Vincent Lacroix

Participants: Camille Marchet, Aurélie Siberchicot, Audric Cologne, Clara Benoît-Pilven, Janice Kielbassa, Lilia Brinza, Vincent Lacroix

7.1.16 KisSplice

Keywords: Bioinformatics, Bioinfirmatics search sequence, Genomics, NGS

Functional Description: Enables to analyse RNA-seq data with or without a reference genome. It is an exact local transcriptome assembler, which can identify SNPs, indels and alternative splicing events. It can deal with an arbitrary number of biological conditions, and will quantify each variant in each condition.

Release Contributions: Improvements : The KissReads module has been modified and sped up, with a significant impact on run times. Parameters : –timeout default now at 10000: in big datasets, recall can be increased while run time is a bit longer. Bugs fixed : –Reads containing only 'N': the graph construction was stopped if the file contained a read composed only of 'N's. This is was a silence bug, no error message was produced. –Problems compiling with new versions of MAC OSX (10.8+): KisSplice is now compiling with the new default C++ compiler of OSX 10.8+.

KISSPLICE was applied to a new application field, virology, through a collaboration with the group of Nadia Naffakh at Institut Pasteur. The goal is to understand how a virus (in this case influenza) manipulates the splicing of its host. This led to new developments in KISSPLICE. Taking into account the strandedness of the reads was required, in order not to mis-interpret transcriptional readthrough. We now use BCALM instead of DBG-V4 for the de Bruijn graph construction and this led to major improvements in memory and time requirements of the pipeline. We still cannot scale to very large datasets like in cancer, the time limiting step being the quantification of bubbles.

URL: http://kissplice.prabi.fr/

Contact: Vincent Lacroix

Participants: Alice Julien-Laferrière, Leandro Ishi Soares de Lima, Vincent Miele, Rayan Chikhi, Pierre Peterlongo, Camille Marchet, Gustavo Akio Tominaga Sacomoto, Marie-France Sagot, Vincent Lacroix

7.1.17 KisSplice2RefGenome

Keywords: Bioinformatics, NGS, Transcriptomics

Functional Description: KISSPLICE identifies variations in RNA-seq data, without a reference genome. In many applications however, a reference genome is available. KISSPLICE2REFGENOME enables to facilitate the interpretation of the results of KISSPLICE after mapping them to a reference genome.

URL: http://kissplice.prabi.fr/tools/kiss2refgenome/

Contact: Vincent Lacroix

Participants: Audric Cologne, Camille Marchet, Camille Sessegolo, Alice Julien-Laferrière, Vincent Lacroix

7.1.18 KisSplice2RefTranscriptome

Keywords: Bioinformatics, NGS, Transcriptomics

Functional Description: KISSPLICE2REFTRANSCRIPTOME enables to combine the output of KISSPLICE with the output of a full length transcriptome assembler, thus allowing to predict a functional impact for the positioned SNPs, and to intersect these results with condition-specific SNPs. Overall, starting from RNA-seq data only, we obtain a list of condition-specific SNPs stratified by functional impact.

URL: http://kissplice.prabi.fr/tools/kiss2rt/

Contact: Vincent Lacroix

Participants: Helene Lopez Maestre, Mathilde Boutigny, Vincent Lacroix

7.1.19 MetExplore

Keywords: Systems Biology, Bioinformatics

Functional Description: Web-server that allows to build, curate and analyse genome-scale metabolic networks. METEXPLORE is also able to deal with data from metabolomics experiments by mapping a list of masses or identifiers onto filtered metabolic networks. Finally, it proposes several functions to perform Flux Balance Analysis (FBA). The web-server is mature, it was developed in PHP, JAVA, Javascript and Mysql. METEXPLORE was started under another name during Ludovic Cottret's PhD in Bamboo, and is now maintained by the METEXPLORE group at the Inra of Toulouse.

URL: https://metexplore.toulouse.inra.fr/index.html/

Contact: Fabien Jourdan

Participants: Fabien Jourdan, Hubert Charles, Ludovic Cottret, Marie-France Sagot

7.1.20 Mirinho

Keywords: Bioinformatics, Computational biology, Genomics, Structural Biology

Functional Description: Predicts, at a genome-wide scale, microRNA candidates.

URL: http://team.inria.fr/erable/en/software/mirinho/

Contact: Marie-France Sagot

Participants: Christian Gautier, Christine Gaspin, Cyril Fournier, Marie-France Sagot, Susan Higashi

7.1.21 Momo

Name: Multi-Objective Metabolic mixed integer Optimization

Keywords: Metabolism, Metabolic networks, Multi-objective optimisation

Functional Description: MOMO is a multi-objective mixed integer optimisation approach for enumerating knockout reactions leading to the overproduction and/or inhibition of specific compounds in a metabolic network.

URL: http://team.inria.fr/erable/en/software/momo/

Contact: Marie-France Sagot

Participants: Ricardo Luiz de Andrade Abrantes, Nuno Mira, Susana Vinga, Marie-France Sagot

7.1.22 Moomin

Name: Mathematical explOration of Omics data on a MetabolIc Network

Keywords: Metabolic networks, Transcriptomics

Functional Description: MOOMIN is a tool for analysing differential expression data. It takes as its input a metabolic network and the results of a DE analysis: a posterior probability of differential expression and a (logarithm of a) fold change for a list of genes. It then forms a hypothesis of a metabolic shift, determining for each reaction its status as "increased flux", "decreased flux", or "no change". These are expressed as colours: red for an increase, blue for a decrease, and grey for no change. See the paper for full details: https://doi.org/10.1093/bioinformatics/btz584

URL: https://github.com/htpusa/moomin

Contact: Marie-France Sagot

Participants: Henri Taneli Pusa, Mariana Ferrarini, Ricardo Luiz de Andrade Abrantes, Arnaud Mary, Alberto Marchetti-Spaccamela, Leendert Stougie, Marie-France Sagot

7.1.23 MultiPus

Keywords: Systems Biology, Algorithm, Graph algorithmics, Metabolic networks, Computational biology

Functional Description: MULTIPUS (for "MULTIple species for the synthetic Production of Useful biochemical Substances") is an algorithm that, given a microbial consortium as input, identifies all optimal sub-consortia to synthetically produce compounds that are either exogenous to it, or are endogenous but where interaction among the species in the sub-consortia could improve the production line.

URL: https://team.inria.fr/erable/en/software/multipus/

Contact: Marie-France Sagot

Participants: Alberto Marchetti-Spaccamela, Alice Julien-Laferrière, Arnaud Mary, Delphine Parrot, Laurent Bulteau, Leendert Stougie, Marie-France Sagot, Susana Vinga

7.1.24 Pitufolandia

Keywords: Bioinformatics, Graph algorithmics, Systems Biology

Functional Description: The algorithms in PITUFOLANDIA (PITUFO / PITUFINA / PAPAPITUFO) are designed to solve the minimal precursor set problem, which consists in finding all minimal sets of precursors (usually, nutrients) in a metabolic network that are able to produce a set of target metabolites.

URL: https://team.inria.fr/erable/en/software/pitufo/

Contact: Marie-France Sagot

Participants: Vicente Acuña, Paulo Vieira Milreu, Alberto Marchetti-Spaccamela, Leendert Stougie, Martin Wannagat, Marie-France Sagot

7.1.25 Sasita

Keywords: Bioinformatics, Graph algorithmics, Systems Biology

Functional Description: SASITA is a software for the exhaustive enumeration of minimal precursor sets in metabolic networks.

URL: https://team.inria.fr/erable/en/software/sasita/

Contact: Marie-France Sagot

Participants: Vicente Acuña, Ricardo Luiz de Andrade Abrantes, Paulo Vieira Milreu, Alberto Marchetti-Spaccamela, Leendert Stougie, Martin Wannagat, Marie-France Sagot

7.1.26 Smile

Keywords: Bioinformatics, Genomic sequence

Functional Description: Motif inference algorithm taking as input a set of biological sequences.

Contact: Marie-France Sagot

Participant: Marie-France Sagot

7.1.27 Rime

Keywords: Bioinformatics, Genomics, Sequence alignment

Functional Description: Detects long similar fragments occurring at least twice in a set of biological sequences.

Contact: Nadia Pisanti

Participants: Nadia Pisanti, Marie-France Sagot

7.1.28 Totoro

Name: Transient respOnse to meTabOlic pertuRbation inferred at the whole netwOrk level

Keywords: Bioinformatics, Graph algorithmics, Systems Biology

Functional Description: TOTORO is a constraint-based approach that integrates internal metabolite concentrations that were measured before and after a perturbation into genome-scale metabolic reconstructions. It predicts reactions that were active during the transient state that occurred after the perturbation. The method is solely based on metabolomic data.

URL: https://gitlab.inria.fr/erable/totoro

Contact: Irene Ziska

Participants: Irene Ziska, Arnaud Mary, Marie-France Sagot

7.1.29 Wengan

Name: Making the path

Keyword: Genome assembly

Functional Description: WENGAN is a new genome assembler that unlike most of the current long-reads assemblers avoids entirely the all-vs-all read comparison. The key idea behind WENGAN is that long-read alignments can be inferred by building paths on a sequence graph. To achieve this, WENGAN builds a new sequence graph called the Synthetic Scaffolding Graph. The SSG is built from a spectrum of synthetic mate-pair libraries extracted from raw long-reads. Longer alignments are then built by performing a transitive reduction of the edges. Another distinct feature of WENGAN is that it performs self-validation by following the read information. WENGAN identifies miss-assemblies at differents steps of the assembly process.

URL: https://github.com/adigenova/wengan

Contact: Marie-France Sagot

Participants: Alex Di Genova, Marie-France Sagot

7.1.30 WhatsHap

Keywords: Bioinformatics, Genomics

Functional Description: WHATSHAP is a DP approach for haplotype assembly from long reads that works until 20x coverage and solves the minimum error correction problem exactly. PWHATSHAP is a parallelisation of the core dynamic programming algorithm of WHATSHAP.

URL: https://bitbucket.org/whatshap/whatshap

Contact: Nadia Pisanti

8 New results

8.1 General comments

We present in this section the main results obtained in 2021.

We tried to organise these along the four axes as presented above. Clearly, in some cases, a result obtained overlaps more than one axis. In such case, we chose the one that could be seen as the main one concerned by such results.

On the other hand, we chose not to detail the results on more theoretical aspects of computer science when these are initially addressed in contexts not directly related to computational biology [4, 6, 13, 23, 25, 27, 29, 30, 31, 33, 34, 32, 38] even though they could be relevant for different problems in the life sciences areas of research, or could become more specifically so in a near future, in particular those on string algorithms [3, 10, 11, 16, 24, 28] in the case for instance, but not exclusively, of (pan)genome analysis. Notably, the work presented in [36] on bidirectional string anchors as a new string sampling mechanism.

A few other results of 2021 are not mentioned in this report, not because the corresponding work is not important, but because it was likewise more specialised on a specific topic, such as for instance scheduling [7, 12]. In the same way, also for space reasons, we chose not to detail the results presented in some biological papers of the team when these did not require a mathematical or algorithmic input [5, 19, 20], or because they are too specialised [2], or yet because they are still submitted only [43].

Finally, we wish to call attention to the fact that some members of ERABLE, at CWI and at the University of Pisa, have been working on a theoretical problem which is important in relation to our main area of application. This problem indeed concerns privacy of the information that may be inferred by some of the algorithms developed, and more precisely what has been called string sanitization [26, 37].

The work above has involved the following members of Erable:

Participants: Hubert Charles, Roberto Grossi, Giuseppe Francesco Italiano, Laurent Jacob, Alberto Marchetti-Spaccamela, Nadia Pisanti, Solon Pissis, Leen Stougie, Cristina Vieira.

8.2 Axis 1: (Pan)Genomics and transcriptomics in general

Bubbles and bubble generator

Participants: Audric Cologne, Eric Cumunel, Giuseppe Francesco Italiano, Vincent Lacroix, Arnaud Mary, Marie-France Sagot, Camille Sessegolo, Blerina Sinaimeri.

Bubbles are pairs of internally vertex-disjoint s,t-paths in a directed graph. In de Bruijn graphs built from reads of RNA and DNA data, bubbles represent interesting biological events, such as alternative splicing (AS) and allelic differences (SNPs and indels). However, the set of all bubbles in a de Bruijn graph built from real data is usually too large to be efficiently enumerated and analysed in practice. In particular, despite significant research done in this area, listing bubbles still remains the main bottleneck for tools that detect AS events in a reference-free context. We recently introduced the concept of a bubble generator as a way for obtaining a compact representation of the bubble space of a graph (Acuña *et al.*, *Algorithmica*, 82:898-914, 2019, appeared in 2020). Although this generator was quite effective in finding AS events, preliminary experiments showed that it is about 5 times slower than state-of-art methods. This year, we proposed a new family of bubble generators which improve substantially on the previous one. Indeed, the generators in this new family are about two orders of magnitude faster and are still able to achieve similar precision in identifying AS events. To highlight the practical value of our new generators, we also reported some experimental results on real datasets. This work was presented at IWOCA 2020. An extended version of the paper at IWOCA was submitted to a journal and was just recently accepted and published [1].

In 2021, we continued working on this problem. Indeed, our work raised several new and perhaps intriguing questions. First, we noticed that while for flow graphs our family produces minimum generators, for general graphs it is still open to find a minimum bubble generator. Second, the fast computation of our new generators opens the way to the design of algorithms that efficiently combine the bubbles of a generator in order to find more AS events. Third, we believe that the number of false positives could be reduced by adding more biologically motivated constraints. An example of constraint that can be introduced toward this aim is to give a weight to each edge of the de Bruijn graph based on the reads coverage. A true AS event would then correspond to bubbles in which the edges inside a leg must have similar weights (but different legs may have different coverage). Fourth, when constructing a de Bruijn graph from RNA-seq reads, some filters are applied that are meant to eliminate sequencing errors. These filters remove vertices and edges whose coverage by the set of reads is below some given thresholds. Changing those thresholds has a significant impact on the resulting de Bruijn graph, and hence on the set of solutions. Is it possible to compute in a dynamic fashion a bubble generator when this coverage threshold is changing, without having to recompute everything from scratch?

Finally, in October 2021, Camille Sessegolo defended her PhD [41] co-supervised by V. Lacroix and A. Mary. In the first part of her work, Camille analysed Nanopore long read datasets to understand how this technology can help to study eukaryotic transcriptomes. In particular, she checked whether transcript and gene quantifications obtained with Nanopore data were reliable. She used Spike-in (artificial transcripts from which the real quantification is known) and showed that the most precise quantifications were obtained with the RNA direct protocol. Moreover, she observed that only a fraction of the long reads covered full length transcripts. She then worked on a new model for complex alternative splicing events in non-model species.

De Bruijn graphs

Participants: Giuseppe Francesco Italiano, Blerina Sinaimeri.

In a more purely theoretical work yet still related in particular with the previous problem and with the axis in general as it deals with de Bruijn graphs, we proposed in [35] a new compressed representation for weighted de Bruijn graphs, which is based on the idea of delta-encoding the variations of k-mer abundances on a spanning branching of the graph which is likely to be of practical value. Indeed, as an illustration, when combined with the compressed BOSS de Bruijn graph representation, it encodes the weighted de Bruijn graph of a 16x-covered DNA read-set (60M distinct k-mers, k = 28) within 4.15 bits per distinct k-mer and can answer abundance queries in about 60 microseconds on a standard machine. In contrast, state of the art tools declare a space usage of at least 30 bits per distinct k-mer for the same task, which was confirmed by our experiments. As a by-product of our new data structure, we exhibited efficient compressed data structures for answering partial sums on edge-weighted trees, which might be of independent interest.

Genome assembly

Participants: Hubert Charles, Alex di Genova, Mariana Galvão Ferrarini, Marie-France Sagot, Cristina Vieira.

Generating accurate genome assemblies of large, repeat-rich human genomes has proved difficult using only long, error-prone reads, and most human genomes assembled from such long reads thus add accurate short reads to polish the consensus sequence. We had developed a hybrid assembly method, that we called WENGAN, to address this problem. We had shown in the paper presenting WENGAN that was published in 2020 in *Nature Biotechnology* that WENGAN provides the highest quality at a low computational cost. In 2021, WENGAN was used in a collaboration with biologists, notably from the Insa at Lyon, to assemble the transposable element-rich genome of the cereal pest *Sitophilus oryzae*. This resulted in a paper accepted at *BMC Biology* [21].

Binning

Participants: Marianne Borderes, Mariana Galvão Ferrarini, Marie-France Sagot, Susana Vinga.

The human gut microbiota performs functions that are essential for the maintenance of the host physiology. However, characterising the functioning of microbial communities in relation to the host remains challenging in reference-based metagenomic analyses. Indeed, as taxonomic and functional analyses are performed independently, the link between genes and species remains unclear. Although a first set of species-level bins was built by clustering co-abundant genes, no reference bin set is established on the most used gut microbiota catalog, the Integrated Gene Catalog (IGC). With the aim to identify the best suitable method to group the IGC genes, we had benchmarked nine taxonomy-independent binners implementing abundance-based, hybrid and integrative approaches. To this purpose, we designed a Simulated non-redundant Gene Catalog (SGC) and computed assessment metrics that were adapted for our purpose. We showed that, overall, the best trade-off between the main metrics is reached by an integrative binner. For each approach, we then compared the results of the best-performing binner with our expected community structures and applied the method to IGC. We showed that the three approaches investigated are distinguished by specific advantages, and by inherent or scalability limitations. We concluded from this that hybrid and integrative binners show promising and potentially complementary results but that they require improvements to be used on IGC to recover human gut microbial species. This work was submitted to *NAR Genomics and Bioinformatics* in 2020 and was published in early 2021 [8]. This is the work of the PhD student Marianne Borderes, co-supervised by M.-F. Sagot and S. Vinga (Instituto Superior Técnico, Lisbon), and funded by the ANR Technology Spock with the company MaatPharma, under the supervision of Lilia Boucinha initially, and then from 2019 on of Emmanuel Prestat. This work with S. Vinga was part of the Inria Associated Team Compasso which lasted from 2018 until the end of 2020.

Related to this, in 2021, Marianne Borderes defended her PhD in July. More recent work appears in the manuscript which however, due to confidentiality reasons, is not publicly available, and therefore also cannot be cited as we are not allowed to upload it in HAL. The two other works done in the context of this PhD should be submitted in 2022. They have been delayed in great part due to the fact that the company wishes to deposit patents and this takes time.

Genome Wide Association Study (GWAS)

Participants: Laurent Jacob, Arnaud Mary.

Genome wide association studies (GWAS) which aim to find genetic variants associated with a trait have widely been used on bacteria to identify genetic determinants of drug resistance or hyper-virulence. Recent bacterial GWAS methods usually rely on k-mers, whose presence in a genome can denote variants ranging from single nucleotide polymorphisms to mobile genetic elements. Since many bacterial species include genes that are not shared among all strains, this approach avoids the reliance on a common reference genome. However, the same gene can exist in slightly different versions across different strains, leading to diluted effects when trying to detect its association to a phenotype through k-mer-based GWAS. In a paper that has been submitted in 2021, we proposed to overcome this by testing covariates built from closed connected subgraphs of the De Bruijn graph defined over genomic k-mers. These covariates are able to capture polymorphic genes as a single entity, improving k-mer-based GWAS in terms of power and interpretability. As the number of subgraphs is exponential in the number of nodes in the DBG, a method naively testing all possible subgraphs would result in very low statistical power due to multiple testing corrections, and the mere exploration of these subgraphs would quickly become computationally intractable. The concept of testable hypothesis has successfully been used to address both problems in similar contexts. This concept was leveraged to test all closed connected subgraphs by proposing a novel enumeration scheme for these objects which fully exploits the pruning opportunity offered by testability, resulting in drastic improvements in computational efficiency. This was shown on both real and simulated datasets. We also showed how by considering subgraphs, we could obtain a more powerful and interpretable method. The latter is integrated with existing visual tools to facilitate interpretation.

An implementation of the method, as well as code to reproduce all results is available on request. A preliminary version of the submitted paper [44] is available already in bioRxiv here.

8.3 Axis 2: Metabolism and (post)transcriptional regulation

Metabolism

Participants: Mariana Galvão Ferrarini, Arnaud Mary, Marie-France Sagot, Leen Stougie, Susana Vinga.

In 2020, we mentioned an article that had been submitted and which presented a novel computational method called TOTORO (for "Transient respOnse to meTabOlic pertuRbation inferred at the whole netwOrk level"). TOTORO integrates the concentrations of internal metabolites that were measured before and after a perturbation into a genome-scale metabolic reconstruction in order to predict the reactions that were active during the transient state which occurred after the perturbation. The proposed method is a constraint-based approach that takes the stoichiometry of the network into account. It minimises the change in concentrations for unmeasured metabolites and also the number of active reactions during the transient state to account for a parsimonious assumption. An implementation in C++ is freely available here. TOTORO is able to handle full networks and to consider in the model stoichiometry, cycles, reversible reactions as well as co-factors. This work was also part of the PhD of Irene Ziska (co-supervision between M.-F. Sagot and S. Vinga from IST, Lisbon, Portugal) that was defended in November 2020, and of an Inria Associated Team project (Compasso) with Portugal. For various reasons, the paper is still in revision although the process is now once again well on its way.

Although not related to metabolism, and instead purely theoretical, we mention here also a work done this year on hypergraphs, which represent the mathematical model used to represent metabolic networks, more precisely on sampling hypergraphs with given degrees.

There is a well-known connection between hypergraphs and bipartite graphs, obtained by treating the incidence matrix of the hypergraph as the biadjacency matrix of a bipartite graph. We used this connection in [14] to describe and analyse a rejection sampling algorithm for sampling simple uniform hypergraphs with a given degree sequence. The algorithm presented in the paper uses, as a black box, an algorithm A for sampling bipartite graphs with given degrees, uniformly or nearly uniformly, in (expected) polynomial time. The expected runtime of the hypergraph sampling algorithm depends on the (expected) runtime of the bipartite graph sampling algorithm A, and the probability that a uniformly random bipartite graph with given degrees corresponds to a simple hypergraph. Some conditions were given on the hypergraph degree sequence which guarantee that this probability is bounded below by a positive constant.

Post-transcriptional regulation

Participants: Mariana Galvão Ferrarini, Carol Moraga Quinteros, Marie-France Sagot, Susana Vinga.

MicroRNAs (miRNAs) belong to a class of small non-coding RNAs (ncRNAs) of 18-24 nucleotides in part responsible for post-transcriptional gene regulation in eukaryotes. These evolutionarily conserved molecules influence fundamental biological processes, including cell proliferation, differentiation, apoptosis, immune response, and metabolism. Accurately identifying miRNAs has however proven difficult. In the last decade, with the increasing accessibility of high-throughput sequencing technologies, different methods have been developed to identify miRNAs, but most of them rely exclusively on pre-existing reference genomes. Despite all the advancements in the sequencing technologies and *de novo* assembly algorithms, few complete genomes are available today. This represents a recurrent problem for researchers working on non-model species. The lack of a high-quality reference genome thus reduces the possibilities for discovering novel miRNAs. In a paper that was also submitted in 2020, we introduced BRUMIR, which is a package composed of three tools; 1) a new discovery miRNA tool (BRUMIR-CORE) a

specific genome mapper (BRUMIR2REFERENCE), and 3) a sRNA-seq read simulator (MIRSIM). In particular, BRUMIR-CORE is a *de novo* algorithm based on a de Bruijn graph approach that is able to identify miRNAs directly and exclusively from sRNA-seq data. Due to the end of the PhD of the main participant in this work, Carol Moraga Quinteros, funded by Conicyt and defended in October 2020, plus some other reasons related to the Covid which made progress difficult both in France and in Chile where one of the authors of the paper and (in)formal collaborator of ERABLE works, the revision process of the paper was much delayed but is now accelerating once again. A preprint is available in BioRxiv here, and the code is in GitHub here.

8.4 Axis 3: (Co)Evolution

Phylogeny

Participants: Leen Stougie.

Maximum parsimony distance is a measure used to quantify the dissimilarity of two unrooted phylogenetic trees. It is NP-hard to compute, and very few positive algorithmic results are known due to its complex combinatorial structure. This shortcoming was addressed in [17] where we showed that the problem is fixed parameter tractable. This was done by establishing a linear kernel. After applying certain reduction rules, the resulting instance thus reaches a size that is bounded by a linear function of the distance. As powerful corollaries to this result, we proved that the problem permits a polynomial-time constant factor approximation algorithm; that the treewidth of a natural auxiliary graph structure encountered in phylogenetics is bounded by a function of the distance; and that the distance is within a constant factor of the size of a maximum agreement forest of the two trees, a well studied object in phylogenetics.

Cophylogeny

Participants: Arnaud Mary, Marie-France Sagot, Blerina Sinaimeri, Yishu Wang.

Phylogenetic tree reconciliation is the method of choice for analysing host-symbiont systems. Despite the many reconciliation tools that have been proposed in the literature, two main issues were still unresolved: (i) listing suboptimal solutions (*i.e.* whose score is "close" to the optimal ones) and (ii) listing only solutions that are biologically different "enough". The first issue arises because the optimal solutions are not always the ones biologically most significant; providing many suboptimal solutions as alternatives for the optimal ones is thus very useful. The second one is related to the difficulty to analyse a number of optimal solutions that is often exponential. In 2020, a method, that we called CAPYBARA for "equivalence ClAss enumeration of coPhylogenY event-BAsed ReconciliAtions", was then proposed that addressed both of these problems in an efficient way. Furthermore, CAPYBARA included a tool for visualising the solutions that may significantly help the user in the process of analysing the results. The source code, documentation, and binaries for all platforms are freely available here. This work was published in 2020 as an Application Note in the journal *Bioinformatics*.

The problem of an efficient enumeration of equivalence classes or of one representative per class (without generating all the solutions), although identified as a need in many areas, has been addressed only for very few specific cases. In 2020, we started working on providing a general framework that solves this problem in polynomial delay in a wide variety of contexts, including optimisation ones that can be addressed by dynamic programming algorithms such as is the case of phylogentic tree reconciliation, and for certain types of equivalence relations between solutions. Two papers issued from this work in 2021, one theoretical, and one an extension of the work we have been doing on cophylogeny and phylogenetic tree reconciliation.

In the first paper [39], we provided a general theoretical framework that solves this problem in polynomial delay for a wide variety of contexts, including optimisation ones that can be addressed by

dynamic programming algorithms, and for certain types of equivalence relations between solutions. In order to do this, we went through an intermediate problem, namely the enumeration of coloured subtrees in acyclic decomposable AND/OR graphs (ad-AND/OR graph). This paper was presented at ESA 2021. An extended journal version is currently submitted.

In the second paper [40], we went back to the problem of phylogenetic tree reconciliation. We introduced three different criteria under which two solutions may be considered biologically equivalent, and thus various equivalence relations for grouping the reconciliations that may be considered biologically equivalent. We then used the theoretical framework developed and presented in the first paper [39] to propose polynomial-delay algorithms specifically adapted to the tree reconciliation problem. Although the method corresponds to CAPYBARA, the algorithms, the proofs, and the experiments were presented in detail in this paper for the first time. This paper was accepted at WABI 2021. An extended journal version is currently submitted.

All the above work on cophylogeny and phylogenetic tree reconciliation was part of the PhD of Yishu Wang [42] defended in October of 2021.

There is however new work in preparation, also related to cophylogeny but where this time we propose a method, called AMOCOALA which, for a given pair of host and symbiont trees, estimates the probabilities of the cophylogeny events, where one of the events correspond to what has been called in the literature spread events. We rely for this on an approximate Bayesian computation (ABC) approach. This is an extension of a previous work which led to the method called COALA. The algorithm that we propose, by including spread events, enables multiple associations to be taken into account in a more accurate way, inducing more confidence in the estimated sets of costs and thus in the reconciliation of a given pair of host and symbiont trees. Its rooting in the previous method COALA allows it to estimate the probabilities of the events even in the case of large datasets. A paper is already well advanced and should be submitted in early 2022. In it, we evaluate our method on both synthetic and real datasets. The software will also soon be made publicly available. This is work done in collaboration with Catherine Matias, as was already the work on COALA.

8.5 Axis 4: Health in general

Human SARS-CoV-2

Participants: Leen Stougie.

The work done in previous years by some members of ERABLE on phylogenetic networks was used in 2021 to try to represent the evolutionary history of SARS-CoV-2 in [22].

Rooted phylogenetic networks enable to display complex evolutionary history involving so-called reticulation events, such as genetic recombination. Various methods have been developed to construct such networks, using for example a multiple sequence alignment or multiple phylogenetic trees as input data. Coronaviruses are known to recombine frequently, but rooted phylogenetic networks have not yet been used extensively to describe their evolutionary history. In this paper, we created a workflow to compare the evolutionary history of SARS-CoV-2 with other SARS-like viruses using several rooted phylogenetic network inference algorithms. This workflow includes filtering noise from sets of phylogenetic trees by contracting edges based on branch length and bootstrap support, followed by resolution of multifurcations. We explored the running times of the network inference algorithms, the impact of filtering on the properties of the produced networks, and attempted to derive biological insights regarding the evolution of SARS-CoV-2 from them.

Our results demonstrate that as part of a wider workflow and with careful attention paid to running time, rooted phylogenetic network algorithms are capable of producing plausible networks from coronavirus data. The networks we obtained partly corroborate the existing theories about SARS-CoV-2, and partly produce new avenues for exploration regarding the location and significance of reticulate activity within the wider group of SARS-like viruses. Moreover, our workflow may serve as a model for pipelines in which phylogenetic network algorithms can be used to analyse different datasets and test different hypotheses.

Human and animal

Participants: Mariana Galvão Ferrarini, Marie-France Sagot.

Toxoplasmosis, a protozoan infection caused by *Toxoplasma gondii*, is estimated to affect around 2.5 billion people worldwide. Nevertheless, the side effects of drugs combined with the long period of therapy usually result in discontinuation of the treatment. New therapies should be developed by exploring peculiarities of the parasite's metabolic pathways, similarly to what has been well described in cancer cell metabolism. An example is the switch in the metabolism of cancer that blocks the conversion of pyruvate into acetyl coenzyme A in mitochondria. In this context, dichloroacetate (DCA) is an anticancer drug that reverts the tumor proliferation by inhibiting the enzymes responsible for this switch: the pyruvate dehydrogenase kinases (PDKs). DCA has also been used in the treatment of certain symptoms of malaria; however, there is no evidence of how this drug affects apicomplexan species. In a paper published this year [15], we studied the metabolism of *T. gondii* and showed that DCA also inhibits *T. gondii*'s *in vitro* infection with no toxic effects on host cells. DCA caused an increase in the activity of pyruvate dehydrogenase followed by an unbalanced mitochondrial activity. We also observed morphological alterations in mitochondria and in a few apicoplasts, both of which are essential organelles for parasite survival. To date, the kinases that potentially regulate the activity of pyruvate metabolism in both organelles have never been described. We confirmed the presence in the genome of *T. gondii* of two putative kinases, verified their cellular localisation in the mitochondrion, and provided *in silico* data suggesting that they are potential targets of DCA.

Currently, the drugs used for toxoplasmosis are severely toxic to human cells, and the treatment still lacks effective and safer alternatives. The search for novel drug targets is thus timely. We reported also in the paper that the treatment of *T. gondii* with an anticancer drug, dichloroacetate (DCA), was effective in decreasing *in vitro* infection without being toxic to human cells. Moreover, we verified the mitochondrial localisation of two kinases that possibly regulate the activity of pyruvate metabolism in *T. gondii*, and which had never been studied. DCA increased pyruvate dehydrogenase (PDH) activity in *T. gondii*, followed by an unbalanced mitochondrial activity, in a manner similar to what was previously observed in cancer cells. We thus proposed that the conserved kinases could be potential regulators of pyruvate metabolism and interesting targets for new therapies.

This work was done in the context of the Capes-Cofecub project Ahimsa with researchers at Instituto de Biologia Molecular do Paraná – Fiocruz-PR, Curitiba, Paraná, Brazil, notably Andrea Àvila, and at the Department of Parasitology of the University of São Paulo, Brazil, notably Ariel Mariano Silber.

Human cancer

Participants: Alain Viari.

Besides the work presented above, Erable, and more specifically one of its members, Alain Viari, has continued to be very active in the area of human cancer research. A number of papers have thus been published in 2021 [9, 18].

9 Bilateral contracts and grants with industry

9.1 Bilateral grants with industry

Spock

Title: characterization of hoSt-gut microbiota interactions and identification of key Players based on a unified reference for standardized quantitative metagenOmics and metaboliC analysis frameworK

Industrial Partner: MaatPharma (Person responsible: Lilia Boucinha until 2019, Emmanuel Prestat from 2019).

ERABLE participants: Marie-France Sagot (ERABLE coordinator and PhD main supervisor with Susana Vinga from IST, Lisbon, Portugal, as PhD co-supervisor), Marianne Borderes (beneficiary of the PhD scholarship in MaatPharma, PhD defended in July 2021).

Type: ANR Technology (2018-2021).

Web page: Spock.

9.2 Informal Relations with Industry

Laurent Jacob works with Pendulum Therapeutics (previously Whole Biome) since 2019, with whom he signed an Non Disclosure Agreement and via whom he collaborates with Hector Roux de Bezieux, who is a PhD student in biostatistics at the University of California (PhD defended this year), Berkeley, USA, and was a computational biologist at the company since coming back to France in September 2021. He is now employed by Pendulum.

10 Partnerships and cooperations

10.1 International initiatives

10.1.1 Inria associated team not involved in an IIL or an international program

CAPOEIRA

Title: Computational APproaches with the Objective to Explore intra and cross-species Interactions and their Role in All domains of life

Duration: Initially planned from 2020 to 2022 included, however due to the Covid, Inria proposed to have the Associated Team start again from 2022.

Coordinators: André Fujita (USP) and Marie-France Sagot (Inria)

Partners: Universidade de São Paulo, Inria

Inria contact: Marie-france Sagot

Summary: The CAPOEIRA project covers theoretical computer science (essentially graph theory), mathematics (combinatorics, statistics, and probability), and the development of algorithms to address various biological questions, in particular, the intra and cross-species interactions, which have implications in all aspects of life sciences, including health, ecology, and environment. Two main general topics will be addressed, namely evolution/co-evolution, and biological network (graph/hypergraph) analysis and comparison.

Web page: Capoeira

10.1.2 Participation in other International Programs

ERABLE has currently one project with Brazil where Inria is involved as institution.

Ahimsa

Title: Alternative approacH to Investigating and Modelling Sickness and health

Coordinators: Marie-France Sagot (ERABLE), Andrea Ávila (Instituto de Biologia Molecular do Paraná – Fiocruz-PR, Curitiba, Paraná, Brazil)

ERABLE participant(s): M. Ferrarini, Arnaud Mary, Scheila Mucha, Marie-France Sagot, Blerina Sinaimeri

Type: Capes-Cofecub (2020-2022)

Web page: Ahimsa

Informal International Partners ERABLE participates in the project Network for Organismal Interactions Research (NOIR) funded by Conicyt in Chile within the call International Networking between Research Centers. The project started in 2019 for 2 years. At the the end of 2020, it was extended until June 2021. The coordinator on the Chilean side is Elena Vidal from the Universidad Mayor, Santiago, Chile, and the main Erable participants in 2021 is Marie-France Sagot.

10.1.3 Visits of international scientists

Inria International Chair None in 2021.

Other international visits to the team

Andrea Ávila

Status Researcher

Institution of origin: Instituto Carlos Chagas (ICC/Fiocruz), Curitiba

Country: Brazil

Dates: November 14 to 23, 2021

Context of the visit: Project Capes-Cofecub Ahimsa

Mobility program/type of mobility: Research stay

Helisson Faoro

Status Researcher

Institution of origin: Instituto Carlos Chagas (ICC/Fiocruz), Curitiba

Country: Brazil

Dates: November 14 to 23, 2021

Context of the visit: Project Capes-Cofecub Ahimsa

Mobility program/type of mobility: Research stay

Ariel Mariano Silber

Status Full professor

Institution of origin: University of São Paulo, Department of Parasitology

Country: Brazil

Dates: November 2 to 23, 2021

Context of the visit: Project USP-UdL

Mobility program/type of mobility: Research stay and Lectures

10.1.4 Visits to international teams

No visit was possible in 2021 due to the Covid.

10.2 European initiatives

10.2.1 FP7 & H2020 projects

Olissipo

Title: Fostering Computational Biology Research and innovation in Lisbon

Coordinator: Susana Vinga, INESC-ID, Instituto Superior Técnico, Lisbon.

Other participants: Inria EPI ERABLE, the Swiss Federal Institute of Technology (ETH Zürich) in Switzerland, and the European Molecular Biology Laboratory (EMBL) in Germany

ERABLE participants: Giuseppe Italiano, Vincent Lacroix, Alberto Marchetti-Spaccamela, Arnaud Mary, Marie-France Sagot (ERABLE coordinator), Blerina Sinaimeri, Leen Stougie, Alain Viari.

Type: H2020 Twinning.

Comment: Due to the Covid-19, the start of this project was delayed until January 1st, 2021. It will last until the end of 2023, unless it is extended due to the fact that some of the planned initiatives for the first year may not be realisable, once again because of the Covid-19.

Web pages: Olissipo-Erable and Olissipo

Besides the above, two ERABLE members in the Netherlands, Solon Pissis and Leen Stougie, participate in an H2020 MSCA-RISE project (2020-2022) called Pangaia (Pan-genome Graph Algorithms and Data Integration) coordinated by Paola Bonizzoni, University of Milan, Italy.

In the same way, another H2020 project, in this case an ITN with acronym Alpaca that involves members of ERABLE has been accepted in 2020 but started only in 2021. Two members of ERABLE will host a PhD student in their institutions, namely Solon Pissis at CWI and Nadia Pisanti at the University of Pisa. Other members of ERABLE will be involved in Alpaca.

Finally, a EU-COFUND grant for 14 postdocs was attributed to the NWO Networks Consortium, 2 of which are assigned to CWI to work with the members of Erable at the CWI.

10.2.2 Other european programs/initiatives

By itself, ERABLE is built from what initially were collaborations with some major European Organisations (CWI, Sapienza University of Rome, LUISS University also in Rome, University of Pisa, Free University of Amsterdam) and then became a European Inria Team.

10.3 National initiatives

10.3.1 ANR

ABRomics-PF

Title: A numerical platform on AMR to store, integrate, analyze and share multi-omics data

Coordinators: Philippe Glaser, Pasteur Institute; Claudine Médigue, CEA/IG/Genoscope and CNRS UMR8030; Jacques van Helden, University Aix-Marseille.

ERABLE participants: Laurent Jacob.

Type: ANR.

Duration: 2021-2025.

Web page: ABRomics-PF.

Aster

Title: Algorithms and Software for Third gEneration Rna sequencing

Coordinator: Hélène Touzet, University of Lille and CNRS.

ERABLE participants: Vincent Lacroix (ERABLE coordinator), Audric Cologne, Eric Cumunel, Alex di Genova, Leandro I. S. de Lima, Arnaud Mary, Marie-France Sagot, Camille Sessegolo, Blerina Sinaimeri.

Type: ANR.

Duration: 2016-2020 - extended into 2021.

Web page: Aster.

Fast-Big

Title: Efficient Statistical Testing for high-dimensional Models: application to Brain Imaging and Genomics.

Coordinator: Bertrand Thirion.

ERABLE participant(s): Laurent Jacob, Antoine Villié.

Type: ANR.

Duration: 2018-2022.

Web page: Fast-Big.

GrR

Title: Graph Reconfiguration.

Coordinator: Nicolas Bousquet.

ERABLE participant(s): Arnaud Mary.

Type: ANR JCJC.

Duration: 2019-2021.

Web page: Not available.

Green

Title: Deciphering host immune gene regulation and function to target symbiosis disturbance and endosymbiont control in insect pests.

Coordinator: Abdelaziz Heddi.

ERABLE participant(s): Marie-France Sagot, Cristina Vieira.

Type: ANR.

Duration: 2018-2021.

Web page: Green.

PIECES

Title: Statistical learning for genome-wide on endless collections of patterns of sequences.

Coordinator: Laurent Jacob.

ERABLE participant(s): Laurent Jacob, Luca Nesterenko, Johanna Trost, Antoine Villié.

Type: ANR JCJC.

Duration: 2021-2024.

Web page: PIECES.

U4atac-brain

Title: Rôle de l'épissage mineur dans le développement cérébral

Coordinator: Patrick Edery, Centre de Recherche en Neurosciences de Lyon.

ERABLE participants: Vincent Lacroix (ERABLE coordinator), Audric Cologne.

Type: ANR.

Duration: 2018-2021.

Web page: Not available.

10.3.2 Idex

Fapesp-UdL project

Title: Graph/Hypergraph (spectral) analysis to compare metabolic networks of pathogenic Trypanosoma sp.

Coordinators: Marie-France Sagot (ERABLE), André Fujita (University of São Paulo (USP), São Paulo, Brazil).

ERABLE participant: Mariana Ferrarini, Vincent Lacroix, Arnaud Mary, Marie-France Sagot, Blerina Sinaimeri.

Type: Fapesp-UcL.

Duration: 2020-2021.

Web page: Not available.

RESPOND

Coordinators: Xavier Charpentier, CIRI Lyon; Samuel Venner, LBBE.

ERABLE participant(s): Laurent Jacob.

Type: Idex.

Duration: 2021-2023.

Web page: Not available.

10.3.3 Others

MITOTIC

Title: Ressources Balances Analyses pour découvrir la vulnérabilité métabolique dans le cancer et identifier de nouvelles thérapies.

Coordinator: Sabine Peres.

ERABLE participant(s): Sabine Peres.

Type: Program "Mathématiques et Informatique" 2021 of ITMO Cancer.

Duration: 2021-2024.

Web page: Not available.

Notice that, besides the project above, were included here also national projects of our members from Italy and the Netherlands when these have no other partners than researchers from the same country. These concern the following:

AHeAD

Title: efficient Algorithms for HArnessing networked Data.

Coordinator: Giuseppe Italiano.

ERABLE participant(s): Roberto Grossi, Giuseppe Italiano.

Type: MUIR PRIN, Italian Ministry of Education, University and Research.

Duration: 2019-2022.

Web page: AHeAD.

Networks

Title: Networks.

Coordinator: Michel Mandjes, University of Amsterdam.

ERABLE participant(s): Solon Pissis, Leen Stougie.

Type: NWO Gravity Program.

Duration: 2014-2024.

Web page: Networks.

Optimal

Title: Optimization for and with Machine Learning.

Coordinator: Dick den Hertog.

ERABLE participant(s): Leen Stougie.

Type: NWO ENW-Groot Program.

Web page: Not available.

11 Dissemination

11.1 Promoting Scientific Activities

11.1.1 Scientific Events: Organisation

General Chair, Scientific Chair

- Giuseppe Italiano is member of the Steering Committee of the *Workshop on Algorithm Engineering and Experimentation (ALENEX)*, of the *International Colloquium on Automata, Languages and Programming (ICALP)*, and of the *Workshop Symposium on Experimental Algorithms (SEA)*.
- Laurent Jacob was Chair of the Track Systems biology and Networks of ISMB-ECCB 2021
- Alberto Marchetti-Spaccamela is a member of the Steering committee of *Workshop on Graph Theoretic Concepts in Computer Science (WG)*), and of *Workshop on Algorithmic Approaches for Transportation Modeling, Optimization, and Systems (ATMOS)*.
- Arnaud Mary is member of the Steering Committee of *Workshop on Enumeration Problems and Applications (WEPA)*.
- Marie-France Sagot is member of the Steering Committee of *European Conference on Computational Biology (ECCB), International Symposium on Bioinformatics Research and Applications (ISBRA)*, and *Workshop on Enumeration Problems and Applications (WEPA)*. In 2021, she was also co-chair (and member of the organizing committee) of ISMB-ECCB 2021.

Member of Organizing Committees

- Marie-France Sagot was (co-chair and) member of the organising committee of ISMB-ECCB 2021 (July 25-30).
- Leen Stougie was co-organiser of the KNAW (Royal Dutch Academy of Sciences)-webinar "Mathematics in times of Corona: Network models and the spread of infectious diseases" (April 19). He was also co-organizer of the 2nd Networks Conference, CWI, Amsterdam (June 7-8).

11.1.2 Scientific Events: Selection

Member of Conference Program Committees

- Giuseppe Italiano was a member of the Program Committee of *FCT*, *IWOCA*, *SODA*, and *SPIRE*.
- Nadia Pisanti was a member of the Program Committee of *ALENEX*, *ISBRA*, *ICCS*, *MFCS*, *SOFSEM*, *SPIRE*, and *WABI*.
- Solon Pissis was a member of the Program Committee of *ALENEX*, *SOFSEM*, *SPIRE*, *WABI*, and *WALCOM*.
- Marie-France Sagot was a member of the Program Committee of *LAGOS*, *RecombCG*, *PSC*, and *WABI*.
- Blerina Sinaimeri was a member of the Program Committee of *CIAC*, and *SPIRE*.

11.1.3 Journal

Member of Editorial Boards

- Roberto Grossi is member of the Editorial Board of *Theory of Computing Systems (TOCS)* and of *RAIRO – Theoretical Informatics and Applications*.
- Giuseppe Italiano is member of the Editorial Board of *Algorithmica* and *Theoretical Computer Science*.

- Vincent Lacroix is recommender for *Peer Community in Genomics*.
- Alberto Marchetti-Spaccamela is member of the Editorial Board of *Theoretical Computer Science*.
- Nadia Pisanti is since 2017 of *Network Modeling Analysis in Health Informatics and Bioinformatics*.
- Marie-France Sagot is member of the Editorial Board of *BMC Bioinformatics, Algorithms for Molecular Biology, Lecture Notes in BioInformatics*, and since 2021 of *Computer Science Review*.
- Leen Stougie is member of the Editorial Board of *AIMS Journal of Industrial and Management Optimization*.
- Cristina Vieira is Executive Editor of *Gene*, and since 2014 member of the Editorial Board of *Mobile DNA*.

Reviewer - Reviewing Activities Members of ERABLE have reviewed papers for a number of journals including: *Theoretical Computer Science, Algorithmica, SIAM Journal on Computing, Algorithms for Molecular Biology, Bioinformatics, BMC Bioinformatics, Genome Biology, Genome Research, IEEE/ACM Transactions in Computational Biology and Bioinformatics (TCBB), Molecular Biology and Evolution, Nucleic Acid Research.*

11.1.4 Invited Talks

Marie-France Sagot was Keynote Speaker at the 25th International Conference on Research in Computational Molecular Biology - RECOMB, initially Venice but actually virtual, Aug. 29 - Sept. 1.

Leen Stougie was Keynote Speaker at the 31st Conference on Operational Research - EURO, Athens, July 11-14. He also gave an invited talk at the Workshop on Stochastic Programming; Honoring Maarten van der Vlerk that took place in Groningen, The Netherlands, August 10-11.

11.1.5 Scientific Expertise

Giuseppe Italiano is member of the Advisory Board of MADALGO - Center for MAssive Data ALGOrithmics, Aarhus, Denmark.

Laurent Jacob is member of the CNU 26 and of the Pedagogical and Orientation Council (CPO) of the Doctoral School E2M2.

Alberto Marchetti-Spaccamela is since 2021, Vice Rector (Prorettore) for "Digital Technologies" at Sapienza University of Rome.

Nadia Pisanti is since November 1st 2017 member of the Board of the PhD School in Data Science (University of Pisa jointly with Scuola Normale Superiore Pisa, Scuola S. Anna Pisa, IMT Lucca).

Marie-France Sagot is member of the Advisory Board of CWI, Amsterdam, the Netherlands. She is since 2020 a member of the Research Grant Review Committee for the Human Frontier Science Program.

Leen Stougie is since April 2017 Leader of the Life Science Group at CWI. He is member of the General Board of the Dutch Network on the Mathematics of Operations Research (Landelijk Netwerk Mathematische Besliskunde (LNMB)), and member of the Management Team of the Gravity project Networks. In 2021, he was a member of the Gijs de Leve Award committee.

Alain Viari is member of a number of scientific advisory boards (IRT (Institut de Recherche Technologique) BioAster; Centre Léon Bérard). He also coordinates together with J.-F. Deleuze (CNRGH-Evry) the Research & Development part (CRefIX) of the “Plan France Médecine Génomique 2025”.

Cristina Vieira is member of the “Conseil National des Universités” (CNU) 67 (“Biologie des Populations et Écologie”), and since 2017 member of the “Conseil de la Faculté des Sciences et Technologies (FST)” of the University Lyon 1.

11.1.6 Research Administration

Since 2021, Marie-France Sagot is member of the COS of the future Inria Center at Lyon, and member of the COMI.

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

France The members of ERABLE teach both at the Department of Biology of the University of Lyon (in particular within the BISM (BioInformatics, Statistics and Modelling) specialty, and at the department of Bioinformatics of the Insa (National Institute of Applied Sciences). Cristina Vieira is responsible for the Master Biodiversity, Ecology and Evolution. She teaches genetics 192 hours per year at the University and at the ENS-Lyon. Hubert Charles teaches 192 hours per year in statistics and biology. Laurent Jacob is responsible for the EU "high dimensional statistics for genomics data" of the Master 2 "maths in action" at UCBL (12 hours in 2021), for the "Advanced machine learning theory" of the Master 2 "advanced maths" of the ENS de Lyon (6 hours in 2021), and taught at the Master 1 bioinformatics at UCBL (6 hours in 2021). He also gave a course (3h) and a tutorial (3h) "Learning with sequences" at the thematic school of the labex digicosme (Paris University Saclay). Vincent Lacroix is responsible for the M1 master in bioinformatics and of the following courses (L3: Advanced Bioinformatics, M1: Methods for Data Analysis in Genomics, M1: Methods for Data Analysis in Transcriptomics, M1: Bioinformatics Project, M2: Ethics). He taught 192 hours in 2021. Arnaud Mary is responsible for three courses of the Bioinformatics Curriculum at the University (L2: Introduction to Bioinformatics and Biostatistics, M1: Object Oriented Programming, M2: new course on Advanced Algorithms for Bioinformatics). He taught 198 hours in 2021. Blerina Sinaimeri taught 12H hours in 2021 on graph algorithms for the M1 students of the Master in Bioinformatics.

The ERABLE team regularly welcomes M1 and M2 interns from the bioinformatics Master.

All French members of the ERABLE team are affiliated to the doctoral school E2M2, Ecology-Evolution-Microbiology-Modelling.

Italy & The Netherlands Italian researchers teach between 90 and 140 hours per year, at both the undergraduate and at the Master levels. The teaching involves pure computer science courses (such as Programming foundations, Programming in C or in Java, Computing Models, Distributed Algorithms) and computational biology (such as Algorithms for Bioinformatics).

Dutch researchers teach between 60 and 100 hours per year, again at the undergraduate and Master levels, in applied mathematics (*e.g.*Operational Research, Advanced Linear Programming), machine learning (Deep Learning) and computational biology (*e.g.* Biological Network Analysis, Algorithms for Genomics).

In 2021, Leen Stougie also gave an online mini-course on Approximation in stochastic integer programming in collaboration with Ward Romeijnders (Univ. Groningen) at the Networks Training Week, April 19-23.

11.2.2 Supervision

The following PhDs were defended in ERABLE in 2021:

- Giulia Bernardini, University Milan-Bicocca (co-supervisor Nadia Pisanti), January 2021
- Marianne Borderes, University Lyon 1 (funded by ANR Technology Spock, co-supervisors: Susana Vinga – Instituto Superior Técnico at Lisbon; Marie-France Sagot), July 2021
- Camille Sessegolo, University of Lyon 1 (funded by ANR Aster; co-supervisors: Vincent Lacroix, Arnaud Mary), November 2021
- Luca Versari, University of Pisa (supervisor: Roberto Grossi, Software Engineer at Google Research, Zürich, since 2019), April 2021
- Yishu Wang, University Lyon 1 (funded by Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, co-supervisors: Mário Figueiredo – Instituto Superior Técnico at Lisbon; Marie-France Sagot; Blerina Sinaimeri), October 2021

In February 2021, there was also the PhD defence of André Verissímo of the Instituto Superior Técnico (IST) of Lisbon, who was supervised by Susana Vinga from IST and by Marie-France Sagot.

The following are the PhDs in progress:

- Esteban Gabory, CWI (supervisor: Solon Pissis)
- Nicolas Homberg, Inra, Inria & University of Lyon 1 (funded by Inra & Inria, co-supervisors: Christine Gaspin at Inra; Marie-France Sagot)
- Francesca Lizzi, Scuola Normale Superiore (co-supervisor: Nadia Pisanti)
- Luca Nesterenko, University of Lyon 1 (co-supervisors: Laurent Jacob; Bastien Boussau at the LBBE)
- Giulia Punzi, University of Pisa (supervisor: Roberto Grossi)
- Michelle Sweering, CWI (co-supervisors: Solon Pissis and Leen Stougie)
- Antoine Villie, University of Lyon 1 (supervisor: Laurent Jacob)

11.2.3 Juries

The following are the PhD or HDR juries to which members of ERABLE participated in 2021.

- Laurent Jacob: External reviewer of the PhD of Olga Permakovia, supervised by University of Grenoble Alpes, France.
- Arnaud Mary: External reviewer of the PhD of Simon Vilmon, supervised by Lhouari Nourine at University of Clermont-Ferrand, France.
- Marie-France Sagot: Reviewer of the HDR of Vincent Limouzy, University of Clermont-Ferrand, France, January 2021; the HDR of Yann Strozecki, University of Versailles Saint-Quentin-en-Yveline, November 2021; and the HDR of Christelle Gonindard, University Grenoble Alpes; Member of the PhD committee of Gabriela P. Paludo, supervised by Henrique Ferreira from the Federal University of Rio Grande do Sul.

12 Scientific production

12.1 Publications of the year

International journals

[1] V. Acuña, L. I. Soares De Lima, G. F. Italiano, L. P. Sciarria, M.-F. Sagot and B. Sinaimeri. 'A family of tree-based generators for bubbles in directed graphs'. In: *Journal of Graph Algorithms and Applications* 25.1 (2021), pp. 549–562. DOI: 10.7155/jgaa.00571. URL: https://hal.inria.fr/hal-03504540.

[2] H. Alamro, M. Alzamel, C. S. Iliopoulos, S. P. Pissis and S. Watts. 'IUPACpal: efficient identification of inverted repeats in IUPAC-encoded DNA sequences'. In: *BMC Bioinformatics* 22 (Dec. 2021), pp. 1–12. DOI: 10.1186/s12859-021-03983-2. URL: https://hal.inria.fr/hal-03498463.

[3] L. A. K. Ayad, G. Badkobeh, G. Fici, A. Héliou and S. P. Pissis. 'Constructing Antidictionaries in Output-Sensitive Space'. In: *Theory of Computing Systems* 65.5 (July 2021), pp. 777–797. DOI: 10.1007/s00224-020-10018-5. URL: https://hal.inria.fr/hal-03498331.

[4] D. Bacciu, A. Conte, R. Grossi, F. Landolfi and A. Marino. 'K-plex cover pooling for graph neural networks'. In: *Data Mining and Knowledge Discovery* 35 (11th Aug. 2021), pp. 2200–2220. DOI: 10.1007/s10618-021-00779-z. URL: https://hal.inria.fr/hal-03498374.

[5] S. Benhamou, I. Rahioui, H. Henri, H. Charles, P. Da Silva, A. Heddi, F. Vavre, E. Desouhant, F. Calevro and L. Mouton. 'Cytotype Affects the Capability of the Whitefly Bemisia tabaci MED Species To Feed and Oviposit on an Unfavorable Host Plant'. In: *mBio* 12.6 (Nov. 2021), pp. 1–16. DOI: 10.1128/mBio.00730-21. URL: https://hal.archives-ouvertes.fr/hal-03432050.

[6] A. Bjelde, Y. Disser, J. Hackfeld, C. Hansknecht, M. Lipmann, J. Meißner, K. Schewior, M. Schlöter and L. Stougie. 'Tight Bounds for Online TSP on the Line'. In: *ACM Transactions on Algorithms* 17.1 (28th Jan. 2021), pp. 1–58. DOI: 10.1145/3422362. URL: https://hal.inria.fr/hal-03498401.

[7] V. Bonifaci, G. D'Angelo and A. Marchetti-Spaccamela. 'Algorithms for hierarchical and semi-partitioned parallel scheduling'. In: *Journal of Computer and System Sciences* 120 (Sept. 2021), pp. 116–136. DOI: 10.1016/j.jcss.2021.03.006. URL: https://hal.inria.fr/hal-03498319.

[8] M. Borderes, C. Gasc, E. Prestat, M. G. Ferrarini, S. Vinga, L. Boucinha and M.-F. Sagot. 'A comprehensive evaluation of binning methods to recover human gut microbial species from a non-redundant reference gene catalog'. In: *NAR Genomics and Bioinformatics* 3.1 (2021). DOI: 10.1093/nargab/lqab009. URL: https://hal.inria.fr/hal-03157241.

[9] R. Carapito, R. Li, J. Helms, C. Carapito, S. Gujja, V. Rolli, R. Guimaraes, J. Malagon-Lopez, P. Spinnhirny, A. Lederle et al. 'Identification of driver genes for critical forms of COVID-19 in a deeply phenotyped young patient cohort'. In: *Science Translational Medicine* (26th Oct. 2021), pp. 1–59. DOI: 10.1126/scitranslmed.abj7521. URL: https://hal.archives-ouvertes.fr/hal-03436053.

[10] P. Charalampopoulos, T. Kociumaka, S. P. Pissis, J. Radoszewski, W. Rytter, J. Straszyński, T. Waleń and W. Zuba. 'Circular pattern matching with k mismatches'. In: *Journal of Computer and System Sciences* 115 (Feb. 2021), pp. 73–85. DOI: 10.1016/j.jcss.2020.07.003. URL: https://hal.inria.fr/hal-03498339.

[11] H. Chen, G. Loukides, S. P. Pissis and H. Chan. 'Influence maximization in the presence of vulnerable nodes: A ratio perspective'. In: *Theoretical Computer Science* 852 (Jan. 2021), pp. 84–103. DOI: 10.1016/j.tcs.2020.11.020. URL: https://hal.inria.fr/hal-03498444.

[12] L. Chen, N. Megow, R. Rischke, L. Stougie and J. Verschae. 'Optimal algorithms for scheduling under time-of-use tariffs'. In: *Annals of Operations Research* 304.1-2 (Sept. 2021), pp. 85–107. DOI: 10.1007/s10479-021-04059-3. URL: https://hal.inria.fr/hal-03474019.

[13] K. M. J. De Bontridder, B. V. Halldórsson, M. M. Halldórsson, C. A. J. Hurkens, J. K. Lenstra, R. Ravi and L. Stougie. 'Local improvement algorithms for a path packing problem: A performance analysis based on linear programming'. In: *Operations Research Letters* 49.1 (Jan. 2021), pp. 62–68. DOI: 10.1016/j.orl.2020.11.005. URL: https://hal.inria.fr/hal-03498424.

[14] M. Dyer, C. Greenhill, P. Kleer, J. Ross and L. Stougie. 'Sampling hypergraphs with given degrees'. In: *Discrete Mathematics* 344.11 (Nov. 2021), pp. 1–14. DOI: 10.1016/j.disc.2021.112566. URL: https://hal.inria.fr/hal-03474025.

[15] M. G. Ferrarini, L. M. Nisimura, R. M. B. M. Girard, M. B. Alencar, M. S. I. Fragoso, C. A. Araújo-Silva, A. D. A. Veiga, A. P. R. Abud, S. C. Nardelli, R. C. Vommaro, A. M. Silber, M. France-Sagot and A. R. Ávila. 'Dichloroacetate and Pyruvate Metabolism: Pyruvate Dehydrogenase Kinases as Targets Worth Investigating for Effective Therapy of Toxoplasmosis'. In: *MSphere* 6.1 (24th Feb. 2021). DOI: 10.1128/mSphere.01002-20. URL: https://hal.inria.fr/hal-03104886.

[16] C. S. Iliopoulos, R. Kundu and S. P. Pissis. 'Efficient pattern matching in elastic-degenerate strings'. In: *Information and Computation* 279 (Aug. 2021), pp. 1–13. DOI: 10.1016/j.ic.2020.104616. URL: https://hal.inria.fr/hal-03498329.

[17] M. Jones, S. Kelk and L. Stougie. 'Maximum parsimony distance on phylogenetic trees: A linear kernel and constant factor approximation algorithm'. In: *Journal of Computer and System Sciences* 117 (May 2021), pp. 165–181. DOI: 10.1016/j.jcss.2020.10.003. URL: https://hal.inria.fr/hal-03498430.

[18] J. Liu, D. Ottaviani, M. Sefta, C. Desbrousses, E. Chapeaublanc, R. Aschero, N. Sirab, F. Lubieniecki, G. Lamas, L. Tonon et al. 'A high-risk retinoblastoma subtype with stemness features, dedifferentiated cone states and neuronal/ganglion cell gene expression'. In: *Nature Communications* 12.1 (Dec. 2021). DOI: 10.1038/s41467-021-25792-0. URL: https://hal.archives-ouvertes.fr/hal-03374490.

[19] P. Marin, A. Jaquet, J. Picarle, M. Fablet, V. Merel, M.-L. Delignette-Muller, M. G. Ferrarini, P. Gibert and C. Vieira. 'Phenotypic and Transcriptomic Responses to Stress Differ According to Population Geography in an Invasive Species'. In: *Genome Biology and Evolution* 13.9 (2021), evab208. DOI: 10.1093/gbe/evab208. URL: https://hal-cnrs.archives-ouvertes.fr/hal-03417240.

[20] V. Mérel, P. Gibert, I. Buch, V. Rodriguez Rada, A. Estoup, M. Gautier, M. Fablet, M. Boulesteix and C. Vieira. 'The Worldwide Invasion of Drosophila suzukii Is Accompanied by a Large Increase of Transposable Element Load and a Small Number of Putatively Adaptive Insertions'. In: *Molecular Biology and Evolution* 38.10 (Oct. 2021), pp. 4252–4267. DOI: 10.1093/molbev/msab155. URL: https://hal-cnrs.archives-ouvertes.fr/hal-03417234.

[21] N. Parisot, C. Vargas-Chávez, C. Goubert, P. Baa-Puyoulet, S. Balmand, L. Beranger, C. Blanc, A. Bonnamour, M. Boulesteix, N. Burlet, F. Calevro, P. Callaerts, T. Chancy, H. Charles, S. Colella, A. Da Silva Barbosa, E. Dell'Aglio, A. Di Genova, G. Febvay, T. Gabaldón, M. Galvão Ferrarini, A. Gerber, B. Gillet, R. Hubley, S. Hughes, E. Jacquin-Joly, J. Maire, M. Marcet-Houben, F. Masson, C. Meslin, N. Montagné, A. Moya, A. T. Ribeiro de Vasconcelos, G. Richard, J. Rosen, M.-F. Sagot, A. Smit, J. Storer, C. Vincent-Monegat, A. Vallier, A. Vigneron, A. Zaidman-Rémy, W. Zamoum, C. Vieira, R. Rebollo, A. Latorre and A. Heddi. 'The transposable element-rich genome of the cereal pest Sitophilus oryzae'. In: *BMC Biology* 19.1 (Dec. 2021), pp. 1–29. DOI: 10.1186/s12915-021-01158-2. URL: https://hal.archives-ouvertes.fr/hal-03432029.

[22] R. Wallin, L. Van Iersel, S. Kelk and L. Stougie. 'Applicability of several rooted phylogenetic network algorithms for representing the evolutionary history of SARS-CoV-2'. In: *BMC Ecology and Evolution* 21.1 (Dec. 2021), pp. 1–14. DOI: 10.1186/s12862-021-01946-y. URL: https://hal.inria.fr/hal-03501817.

International peer-reviewed conferences

[23] G. Amanatidis, F. Fusco, P. Lazos, S. Leonardi, A. Marchetti-Spaccamela and R. Reiffenhäuser. 'Submodular Maximization subject to a Knapsack Constraint: Combinatorial Algorithms with Near-optimal Adaptive Complexity *'. In: ICML 2021 - 38th International Conference on Machine Learning. Vol. 139. Proceedings of Machine Learning Research (PMLR). Lugano, Italy, 18th July 2021, pp. 1–25. URL: https://hal.inria.fr/hal-03474045.

[24] G. Badkobeh, P. Charalampopoulos, D. Kosolobov and S. P. Pissis. 'Internal Shortest Absent Word Queries in Constant Time and Linear Space'. In: CPM 2021 - 32nd Annual Symposium on Combinatorial Pattern Matching. Schloss Dagstuhl - Leibniz-Zentrum für Informatik. Wroclaw, Poland, 5th July 2021, pp. 1–13. URL: https://hal.inria.fr/hal-03498358.

[25] S. Baruah and A. Marchetti-Spaccamela. 'Feasibility Analysis of Conditional DAG Tasks'. In: ECRTS 2021 - 33rd Euromicro Conference on Real-Time Systems. Vol. 196. LIPIcs 12. Modena, Italy, 7th July 2021, pp. 1–17. DOI: 10.4230/LIPIcs.ECRTS.2021.12. URL: https://hal.inria.fr/hal-03395589.

[26] G. Bernardini, A. Marchetti-Spaccamela, S. P. Pissis, L. Stougie and M. Sweering. 'Constructing Strings Avoiding Forbidden Substrings'. In: CPM 2021 - 32nd Annual Symposium on Combinatorial Pattern Matching. Vol. 191. LIPIcs 9. Wroclaw, Poland, 5th July 2021, pp. 1–18. URL: https://hal.inria.fr/hal-03395386.

[27] S. Chakraborty, R. Grossi, K. Sadakane and S. R. Satti. 'Succinct Representation for (Non)Deterministic Finite Automata'. In: LATA 2021 - 15th International Conference on Language and Automata Theory and Applications. Vol. 12638. Lecture Notes in Computer Science. Milan, Italy: Springer, 20th Sept. 2021, pp. 55–67. DOI: 10.1007/978-3-030-68195-1_5. URL: https://hal.inria.fr/hal-03498364.

[28] P. Charalampopoulos, T. Kociumaka, S. P. Pissis and J. Radoszewski. 'Faster Algorithms for Longest Common Substring'. In: ESA 2021 - 29th Annual European Symposium on Algorithms. Schloss Dagstuhl - Leibniz-Zentrum für Informatik. Virtual, Portugal, 6th Sept. 2021. URL: https://hal.inria.fr/hal-03498345.

[29] H. Chen, A. Conte, R. Grossi, G. Loukides, S. P. Pissis and M. Sweering. 'On Breaking Truss-Based Communities'. In: KDD 2021 - 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. ACM. Virtual Event Singapore, Singapore: ACM, 14th Aug. 2021, pp. 117–126. DOI: 10.1145/3447548.3467365. URL: https://hal.inria.fr/hal-03498386.

[30] A. Conte, R. Grossi, G. Loukides, N. Pisanti, S. Pissis and G. Punzi. 'Beyond the BEST Theorem: Fast Assessment of Eulerian Trails'. In: FCT 2021 - 23rd International Symposium on Fundamentals of Computation Theory. Vol. 12867. Lecture Notes in Computer Science. Athens, Greece, 12th Sept. 2021, pp. 162–175. DOI: 10.1007/978-3-030-86593-1_11. URL: https://hal.inria.fr/hal-03498416.

[31] L. Georgiadis, K. Giannis, G. F. Italiano and E. Kosinas. 'Computing Vertex-Edge Cut-Pairs and 2-Edge Cuts in Practice'. In: SEA 2021 - 19th International Symposium on Experimental Algorithms. Vol. 190. LIPIcs. Nice, France, 7th June 2021, pp. 1–19. DOI: 10.4230/LIPIcs.SEA.2021.20. URL: https://hal.inria.fr/hal-03395500.

[32] L. Georgiadis, G. F. Italiano and E. Kosinas. 'Computing the 4-Edge-Connected Components of a Graph in Linear Time'. In: ESA 2021 - 29th Annual European Symposium on Algorithms. Lisbon, Portugal, 6th Sept. 2021, pp. 1–41. URL: https://hal.inria.fr/hal-03474050.

[33] F. Grandoni, G. F. Italiano, A. Łukasiewicz, N. Parotsidis and P. Uznański. 'All-Pairs LCA in DAGs: Breaking through the $O(n2.5)$ barrier'. In: SODA 2021 - ACM-SIAM Symposium on Discrete Algorithms. Alexandria, United States: Society for Industrial and Applied Mathematics (SIAM), 7th Jan. 2021, pp. 273–289. DOI: 10.1137/1.9781611976465.18. URL: https://hal.inria.fr/hal-03474036.

[34] G. F. Italiano, A. Karczmarz and N. Parotsidis. 'Planar Reachability Under Single Vertex or Edge Failures'. In: SODA 2021 - ACM-SIAM Symposium on Discrete Algorithms. Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA). Alexandria, United States: Society for Industrial and Applied Mathematics (SIAM), 7th Jan. 2021, pp. 2739–2758. DOI: 10.1137/1.9781611976465.163. URL: https://hal.inria.fr/hal-03474039.

[35] G. F. Italiano, N. Prezza, B. Sinaimeri and R. Venturini. 'Compressed Weighted de Bruijn Graphs'. In: CPM 2021 - 32nd Annual Symposium on Combinatorial Pattern Matching. Vol. 191. LIPIcs 16. Wroclaw, Poland, 5th July 2021, pp. 1–16. DOI: 10.4230/LIPIcs.CPM.2021.16. URL: https://hal.inria.fr/hal-03395413.

[36] G. Loukides and S. P. Pissis. 'Bidirectional String Anchors: A New String Sampling Mechanism'. In: ESA 2021 - 29th Annual European Symposium on Algorithms. Vol. 204. LIPIcs 64. Lisbon, Portugal, 6th Sept. 2021, pp. 1–21. DOI: 10.4230/LIPIcs.ESA.2021.64. URL: https://hal.inria.fr/hal-03395425.

[37] T. Mieno, S. P. Pissis, L. Stougie and M. Sweering. 'String Sanitization Under Edit Distance: Improved and Generalized'. In: CPM 2021 - 32nd Annual Symposium on Combinatorial Pattern Matching. Wrocław, Poland, 5th July 2021, pp. 1–21. URL: https://hal.inria.fr/hal-03474030.

[38] S. Min, S. G. Park, K. Park, D. Giammarresi, G. Italiano and W.-S. Han. 'Symmetric continuous subgraph matching with bidirectional dynamic programming'. In: VLDB 201 - 47th International Conference on Very Large Data Bases. Vol. 14. Proceedings of the VLDB Endowment 8. Copenhagen, Denmark, Apr. 2021, pp. 1298–1310. DOI: 10.14778/3457390.3457395. URL: https://hal.inria.fr/hal-03498315.

[39] Y. Wang, A. Mary, M.-F. Sagot and B. Sinaimeri. 'A General Framework for Enumerating Equivalence Classes of Solutions'. In: ESA 2021 - 29th Annual European Symposium on Algorithms. Leibniz International Proceedings in Informatics (LIPIcs). Virtual, Portugal, 6th Sept. 2021, pp. 1–14. DOI: 10.4230/LIPIcs.ESA.2021.80. URL: https://hal.inria.fr/hal-03333503.

[40] Y. Wang, A. Mary, M.-F. Sagot and B. Sinaimeri. 'Making Sense of a Cophylogeny Output: Efficient Listing of Representative Reconciliations'. In: WABI 2021 - 21st International Workshop on Algorithms in Bioinformatics. Leibniz International Proceedings in Informatics (LIPIcs). Chicago, United States, 2nd Aug. 2021, pp. 1–18. DOI: 10.4230/LIPIcs.WABI.2021.3. URL: https://hal.inria.fr/hal-03295799.

Doctoral dissertations and habilitation theses

[41] C. Sessegolo. 'Developpement of bioinformatic methods for the study of alternative splicing in non model species : Complex splicing and contribution of third generation sequencing technologies.' ERABLE - Equipe de recherche européenne en algorithmique et biologie formelle et expérimentale; LBBE - Laboratoire de Biométrie et Biologie Evolutive - UMR 5558; Univ Lyon, Université Claude-Bernard Lyon 1, 22nd Oct. 2021. URL: https://tel.archives-ouvertes.fr/tel-03520227.

[42] Y. Wang. 'Algorithmic investigations of the dynamics of species interactions'. ERABLE - Equipe de recherche européenne en algorithmique et biologie formelle et expérimentale; LBBE - Laboratoire de Biométrie et Biologie Evolutive - UMR 5558; Univ Lyon, Université Claude-Bernard Lyon 1, 5th Oct. 2021. URL: https://tel.archives-ouvertes.fr/tel-03499342.

Reports & preprints

[43] C. Genestet, E. Hodille, F. Massol, G. Refrégier, A. Barbry, E. Westeel, G. Lina, F. Ader, L. Jacob, S. Dray, J.-L. Berland, S. Venner and O. Dumitrescu. *Within-host genetic micro-diversity of Mycobacterium tuberculosis and the link with tuberculosis disease features*. 17th Nov. 2021. DOI: 10.1101/2021.04.07.438754. URL: https://hal.archives-ouvertes.fr/hal-03433572.

[44] H. Roux de Bézieux, L. Lima, F. PERRAUDEAU, A. Mary, S. Dudoit and L. Jacob. *CALDERA: Finding all significant de Bruijn subgraphs for bacterial GWAS*. 17th Nov. 2021. DOI: 10.1101/2021.11.05.467462. URL: https://hal.archives-ouvertes.fr/hal-03433563.