RESEARCH CENTRE

Inria Lyon Center

IN PARTNERSHIP WITH:

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

2022

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

Inria

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]
- Alain Viari [INRIA, Senior Researcher, HDR]

Faculty Members

- Roberto Grossi [UNIV PISE, Professor]
- Giuseppe Italiano [UNIV LUISS, Professor]
- Vincent Lacroix [UNIV LYON I, Associate Professor]
- Alberto Marchetti Spaccamela [SAPIENZA ROME, Professor]
- Arnaud Mary [UNIV LYON I, Associate Professor]
- Sabine Peres [UNIV LYON I, Professor, HDR]
- Nadia Pisanti [UNIV PISE, Associate Professor]
- Leen Stougie [CWI]
- Cristina Vieira [UNIV LYON I, Associate Professor, HDR]

Post-Doctoral Fellows

- Mariana Galvão Ferrarini [ICC - BRESIL, from Jun 2022]
- Scheila Gabriele Mucha [INRIA, until May 2022]

PhD Students

- Nicolas Homberg [INRAE]
- Maxime Mahout [UNIV PARIS SACLAY]
- Luca Nesterenko [CNRS]
- Antoine Villie [CNRS]

Technical Staff

- François Gindraud [INRIA, Engineer]
- Johanna Trost [CNRS, Engineer, (Temporary)]

Administrative Assistant

- Anouchka Ronceray [INRIA]

Visiting Scientists

- Nuno Pereira Mira [IST Lisbon, from Nov 2022, Assistant Professor]
- Ariel Mariano Silber [University of São Paulo, Brazil, Professor]
- Gabriela Torres Montanaro [University of São Paulo, Brazil, PhD student]

External Collaborator

- Susana Vinga [IST Lisbon, Associate Professor]

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

7.1 New software

7.1.1 AmoCoala

Name: Associations get Multiple for Our COALA

Keyword: Evolution

Functional Description: Despite an increasingly vaster literature on cophylogenetic reconstructions for studying host-parasite associations, understanding the common evolutionary history of such systems remains a problem that is far from being solved. Many of the most used algorithms do the host-parasite reconciliation analysis using an event-based model, where the events include in general (a subset of) cospeciation, duplication, loss, and host-switch. All known event-based methods then assign a cost to each type of event in order to find a reconstruction of minimum cost. The main problem with this approach is that the cost of the events strongly influence the reconciliation obtained. To deal with this problem, we developed an algorithm, called AMOCOALA, for estimating the frequency of the events based on an approximate Bayesian computation approach in presence of multiple associations.

URL: https://sites.google.com/view/blerinasinaimeri/software/amocoala

Contact: Blerina Sinaimeri

Participants: Laura Urbini, Blerina Sinaimeri

7.1.2 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

7.1.3 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

7.1.4 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

7.1.5 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

7.1.6 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

7.1.7 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

7.1.8 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

7.1.9 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

7.1.10 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

7.1.11 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

7.1.12 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

7.1.13 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

7.1.14 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

7.1.15 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

7.1.16 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

7.1.17 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

7.1.18 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

7.1.19 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

7.1.20 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

7.1.21 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

7.1.22 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

7.1.23 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

7.1.24 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

7.1.25 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

7.1.26 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

7.1.27 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

7.1.28 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

7.1.29 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

7.1.30 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

7.1.31 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 2022.

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 [7, 8, 9, 10, 28, 20, 21] 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 [2, 14, 22, 23, 26, 27].

A few other results of 2022 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 replenishment problems [6]. 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 [4, 11, 32].

The work above has involved the following members of Erable:

Participants: Roberto Grossi, Giuseppe Francesco Italiano, Alberto Marchetti-Spaccamela, Nadia Pisanti, Solon Pissis, Blerina Sinaimeri, Leen Stougie, Cristina Vieira.

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

Alternative splicing and genetic diseases associated with microcephaly and developmental defects

Participants: Audric Cologne, Vincent Lacroix.

Various genetic diseases associated with microcephaly and developmental defects are due to pathogenic variants in the U4atac small nuclear RNA (snRNA), a component of the minor spliceosome essential for the removal of U12-type introns from eukaryotic mRNAs. While it has been shown that a few RNU4ATAC mutations result in impaired binding of essential protein components, the molecular defects of the vast majority of variants remain unknown. In the paper [1], in collaboration notably with Patrick Edery and Sylvie Mazoyer following the ANR project U4ATAC-Brain Vincent Lacroix had with them, lymphoblastoid cells derived from RNU4ATAC compound heterozygous twin patients with MOPD1 phenotypes were used to analyse the molecular consequences of the mutations on small nuclear ribonucleoproteins (snRNPs) formation and on splicing. It was found that the $U4atac108-126del$ mutant is unstable and that the $U4atac111G>A$ mutant as well as the minor di- and tri-snRNPs are present at reduced levels. The results presented in the paper also reveal the existence of 3'-extended snRNA transcripts

in patients' cells. Moreover, we showed that the mutant cells have alterations in splicing of INTS7 and INTS10 minor introns, contain lower levels of the INTS7 and INTS10 proteins and display changes in the assembly of Integrator subunits. Altogether, the results obtained show that the compound heterozygous $g.108_{1}26del; g.111G>A$ mutations induce splicing defects and affect the homeostasis and function of the Integrator complex.

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 was 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. The paper was now accepted in *Bioinformatics* [18].

8.3 Axis 2: Metabolism and (post)transcriptional regulation

Metabolism

Participants: Mariana Galvão Ferrarini, François Gindraud, Maxime Mahout, Arnaud Mary, Nuno Mira, Gabriela Torres Montanaro, Sabine Peres, Marie-France Sagot, Ariel Silber, Susana Vinga.

In 2021, 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"). This paper has now been published in *Frontiers in Genetics* [13]. 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. It was applied to real data of three different growth experiments (pulses of glucose, pyruvate, succinate) from the bacterium *Escherichia coli* and we were able to predict known active pathways and gather new insights on the different metabolisms related to each substrate. We used both the *E. coli* metabolic core and the iJO1366 full metabolic network models to demonstrate that our approach is applicable to both smaller and larger networks. An implementation in C++ is freely available

here. It depends on IBM CPLEX which is freely available for academic purposes. As indicated, 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.

We are currently working on a method that would enable to take into account at the same time metabolomic data as is the case TOTORO and transcriptomic data as is the case of another method developed in the team called MOOMIN. This work and the discussions around it are being conducted with Henri Taneli Pusa, who was PhD student in the team having defended in early 2019, and with whom we have continued collaborating, M. Galvão Ferrarini, A. Mary and M.-F. Sagot.

In parallel to the above, the same plus two external collaborators, Nuno Mira and Susana Vinga, are working on extending two other previous works of the team related to synthetic biology, namely MULTIPUS and MOMO, to be able to address the issue of a potentially toxic character of the compound(s) of interest synthetically produced. This work had started already within the context of Irene Ziska's PhD, and is been taken up again within the context of the Sabbatical of N. Mira within Erable from October 2022 to September 2023.

Another work in progress from this time Arnaud Mary concerns using some of the ideas presented in a theoretical paper [25] that was accepted at ICALP this year to address the problem of cuts in metabolic networks. In the conference paper, the problem that was considered was the one of listing all the minimal chordal completions of a graph, which itself was motivated by the one of enumerating all tree decompositions of a graph. A. Mary and his two co-authors provided a polynomial delay algorithm to solve these problems which, moreover, uses polynomial space, thus improving on the current literature on this topic where the best method available was an incremental polynomial time one, which moreover required exponential space. The algorithm proposed in [25] relies on Proximity Search, a framework introduced in 2019 by Alessio Conte and Takeaki Uno which was shown powerful to obtain polynomial delay algorithms but generally requires exponential space. In order to obtain a polynomial space algorithm for the problem of minimal chordal completions, a new general method was introduced called canonical path reconstruction to design polynomial delay and polynomial space algorithms based on proximity search.

All the methods developed in the past related to metabolism are currently been adapted to become more user-friendly and integrated within a same framework. This is been made possible by the arrival in the team in 2022 of a permanent Inria engineer, François Gindraud.

Finally, in the context of both the Inria Associated Team Capoeira, and of a PhD student, Gabriela T. Montanaro, co-supervised between Ariel M. Silber, Professor at the University of São Paulo, Brazil, and M.-F. Sagot, Erable has started working on problems related with metabolism and tropical diseases, in the case linked to *Trypanosoma cruzi*. Both A. M. Silber and G. T. Montanaro have started making regular more or less long visits to Lyon in 2022, visits which will continue in the next years.

Post-transcriptional regulation

Participants: Mariana Galvão Ferrarini, Nicolas Homberg, 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 now four tools; 1) a new discovery miRNA tool (BRUMIR-CORE) a specific genome mapper (BRUMIR2REFERENCE), 3) a sRNA-seq read simulator (MIRSIM), and 4) a machine learning classifier based on a random forest model that evaluates the sequence-derived features to

further refine the prediction obtained from the (BRUMIR-CORE. 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. The paper has however now been published in *GigaScience* [15] and the code is in GitHub here.

Besides the above, the team has also been investigating the problem of the identification of the targets of miRNAs, a topic not only difficult *per se* but that has also led to intense discussions in the literature, with some (notably the group of Hervé Seitz) arguing that only a small number of the targets currently predicted in silico by the existing methods are actually functional. Independent of whether the latter assessment is correct, clearly there is no strict consensus on which features are important for the identification of targets, whether of miRNAs or of other small non coding RNAs. This work is ongoing, with one paper currently in revision and another that will be submitted in early 2023. It is part of the PhD of Nicolas Homberg.

8.4 Axis 3: (Co)Evolution

Phylogeny

Participants: Roberto Grossi, Laurent Jacob, Luca Nesterenko.

An important problem in molecular evolution is that of phylogenetic reconstruction, that is, given a set of sequences descending from a common ancestor, the reconstruction of the binary tree describing their evolution from the latter. State-of-the-art methods for the task, namely Maximum likelihood and Bayesian inference, have a high computational cost, which limits their usability on large datasets. Recently researchers have begun investigating deep learning approaches to the problem but so far these attempts have been limited to the reconstruction of quartet tree topologies, addressing phylogenetic reconstruction as a classification problem. We presented in this paper [30] that is currently submitted and was already presented at the ISMB Conference this year a radically different approach with a transformer-based network architecture that, given a multiple sequence alignment, predicts all the pairwise evolutionary distances between the sequences, which in turn allow us to accurately reconstruct the tree topology with standard distance-based algorithms. The architecture and its high degree of parameter sharing allow us to apply the same network to alignments of arbitrary size, both in the number of sequences and in their length. We evaluate our network PHYLOFORMER on two types of simulations and find that its accuracy matches that of a Maximum Likelihood method on datasets that resemble training data, while being significantly faster.

In a second paper [29], a method was developed, called PHYBWT, that uses the extended BurrowsWheeler Transform (eBWT) for a collection of DNA sequences to directly reconstruct phylogeny, bypassing the alignment against a reference genome or *de novo* assembly. This method hinges on the combinatorial properties of the eBWT positional clustering framework. The eBWT is employed to detect relevant blocks of the longest shared substrings of varying length (unlike the k-mer-based approaches that need to fix the length k *a priori*), and build a suitable decomposition leading to a phylogenetic tree, step by step. As a result, PHYBWT is a new alignment-, assembly-, and reference-free method that builds a partition tree without relying on the pairwise comparison of sequences, thus avoiding to use a distance matrix to infer a phylogeny. The preliminary experimental results on sequencing data show that our method can handle datasets of diferent types (short reads, contigs, or entire genomes), producing trees of quality comparable to that found in the benchmark phylogeny.

Maximum Agreement Forest

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 a previous paper of the team where we showed that the problem is fixed parameter tractable. This year we worked on the problem of the Maximum Agreement Forest between two rooted binary trees. This NP-hard problem has been studied extensively in the past two decades, since it can be used to compute one type of distance between two phylogenetic trees, in this case rooted. This is the so-called rooted Subtree Prune-and-Regraft (rSPR) distance. We presented in [16] a 2-approximation algorithm that is combinatorial and whose complexity is quadratic in the input size. To prove the approximation guarantee, we constructed a feasible dual solution for a novel exponential-size linear programming formulation. In addition, we showed that this linear program has a smaller integrality gap than previously known formulations, and we gave an equivalent compact formulation, showing that it can be solved in polynomial time.

Phylogenetic networks

Participants: Leen Stougie.

Combining a set of phylogenetic trees into a single phylogenetic network that explains all of them is a fundamental challenge in evolutionary studies. In the paper[24], a recently-introduced theoretical framework of cherry picking was applied to design a class of heuristics that are guaranteed to produce a network containing each of the input trees, for practical-size datasets. The main contribution of the paper was the design and training of a machine learning model that captures essential information on the structure of the input trees and guides the algorithms towards better solutions. This is one of the first applications of machine learning to phylogenetic studies, and we showed its promise with a proof-of-concept experimental study conducted on both simulated and real data consisting of binary trees with no missing taxa.

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 phylogenetic 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.

The theoretical paper was accepted and presented at ESA 2021. An extended journal version is currently submitted.

In the case of phylogenetic tree reconciliation, the paper was accepted and presented at WABI 2021. An extended journal version was then submitted and has been published in 2022 in *Algorithms Mol. Biol.* [19]. There, 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 paper presented at ESA 2021 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.

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

There is however new work that was in preparation last year and is currently in revision in a journal. This work is also related to cophylogeny but 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. Indeed, a major limitation of the current cophylogeny approaches is their inability to model the invasion of different host species by a same symbiont species (the spread events), which is thought to happen in symbiotic relations. To mention one example, the same species of insects may pollinate different species of plants. This results in multiple associations observed between the symbionts and their hosts (meaning that a symbiont is no longer specific to a host), that are not taken into account in most the current methods. In the case of AMOCOALA, we rely on an approximate Bayesian computation (ABC) approach and we had done in 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 preprint of the paper is available in arXiv [31] and the software here. This is work done with Catherine Matias, as was already the work on COALA.

8.5 Axis 4: Health in general

Rare and tropical diseases

Participants: Audric Cologne, Mariana G. Ferrarini, Vincent Lacroix, Arnaud Mary, Marie-France Lacroix.

One main work in the area of health is related to rare diseases, and notably to genetic diseases associated with microcephaly and developmental defects. The work developed this year on this topic has already been described in Axis 1 above. Another work that started more recently relates to tropical diseases and is being conducted in collaboration with Ariel M. Silber, Professor at the University of São Paulo in Brazil together with a PhD student co-supervised by him and M.-F. Sagot. This was mentioned already in the Axis 2 above.

Cancer

Participants: Alain Viari.

What will be mentioned below concerns then mostly cancer, and notably the work of Alain Viari who indeed has continued to be very active in the area of human cancer research. A number of papers have thus been published in 2022 [3, 5, 12, 17].

Amongst the main three groups of Breast Cancers (ER/PR+, HER2+ and TNBC), the TNBC group currently has the poorest prognosis due to the lack of targeted therapies. In the context of the Profiler-1 study conducted at the Centre Leon Bérard [5], we focused on a specific subset of metastatic TNBC, characterised by homologous recombination deficiency (HRD+), leading to enhanced sensitivity to platinum-based chemotherapy. We have shown that mutations in HRR genes and epimutations in

RAD51C were associated with disease control through platinum-based chemotherapy. Paraneoplastic cerebellar degeneration (PCD) with anti-Yo antibodies is a cancer-related autoimmune disease directed against neural antigens expressed by tumor cells. A putative trigger of the immune tolerance breakdown is genetic alteration of Yo antigens. These data confirm the role of genetic alterations of Yo antigens but also outline a specific biomolecular profile in Yo-PCD BCs, suggesting a cancer-specific pathogenesis [17].

Despite decades of research, PDAC is still a highly devastating cancer with very poor prognosis and growing incidence. One of the direction of research focuses on the TGF-β signaling and SMAD4 that is lost in about 50% of PDACs. Using a PDAC patient cohort, we showed that SMAD4-negative tumors with high levels of phospho-SMAD2 are more aggressive and have a poorer prognosis [3]. Thus, loss of SMAD4 tumor suppressive activity in PDAC leads to an oncogenic gain-of-function of SMAD2/3, and to the onset of associated deleterious effects.

Beside purely somatic analyses such as described above, we are also interested in germline genetic variations (SNPs) that may contribute to cancer suceptibility. We are thus participating in several GWAS studies in Breast Cancer (in the context of the MyProbe project) and Lung Cancer (LC) in collaboration with J. D. Mckay at IARC/WHO in Lyon.

9 Partnerships and cooperations

9.1 International initiatives

9.1.1 Inria associate 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: 2020-2022, extended to 2024 due to the pandemic.

Coordinators: Marie-France Sagot (ERABLE) and André Fujita (Instituto de Matemática e Estatistíca, Universidade de São Paulo, Brazil).

ERABLE participants: G. Italiano, V. Lacroix, A. Marchetti-Spaccamela, A. Mary, M.-F. Sagot, B. Sinaimeri, L. Stougie.

Web page: Capoeira.

9.1.2 Participation in other International Programs

Ahimsa

Title: Alternative approacH to Investigating and Modelling Sickness and health.

Coordinators: M.-F. Sagot (ERABLE), A. Ávila (Instituto de Biologia Molecular do Paraná – Fiocruz-PR, Curitiba, Paraná, Brazil).

ERABLE participant(s): M. Ferrarini, A. Mary, S. Mucha, M.-F. Sagot, B. Sinaimeri.

Type: Capes-Cofecub (2020-2022, possibly extended until 2023 due to the pandemic).

Web page: Ahimsa.

9.2 International research visitors

9.2.1 Visits of international scientists

Nuno Mira

Status Assistant Professor

Institution of origin: IST Lisbon

Country: Portugal

Dates: Approximately one week per month from October 2022 to September 2023

Context of the visit: Collaboration

Mobility program/type of mobility: Sabbatical

Ariel Mariano Silber

Status Professor

Institution of origin: University of São Paulo

Country: Brazil

Dates: Regular visits of a few weeks each time

Context of the visit: Collaboration

Mobility program/type of mobility: Research stay

Gabriela Torres Montanaro

Status PhD student in co-supervision

Institution of origin: University of São Paulo

Country: Brazil

Dates: Regular visits of a few weeks each time

Context of the visit: Collaboration

Mobility program/type of mobility: Research stay

9.3 European initiatives

9.3.1 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.

Comments: 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 and some of the second may not be realisable, once again because of the Covid-19.

Web pages: Olissipo-Erable and Olissipo.

9.4 National initiatives

9.4.1 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.

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.

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.

9.4.2 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.

10 Dissemination

10.1 Promoting scientific activities

10.1.1 Scientific events: organisation

General chair, scientific chair

- Giuseppe Italiano is member of the Steering Committee of the International Colloquium on Automata, Languages and Programming (ICALP).
- 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)*.

Member of the organizing committees

- Arnaud Mary was member of the organizing committee of *WEPA* 2022.
- Marie-France Sagot was co-organiser of the Second Edition of the Workshop Metabolism and mathematical models: Two for a tango, held virtually, Oct 25-26, 2022. She is co-organiser of the recurrent Small non-coding RNA bioinformatics club since 2021.

Member of the conference program committees

- Roberto Grossi was member of the Program Committee of *WEPA*.
- Giuseppe Italiano was a member of the Program Committee of *ESA*, *HALG*, and *SODA*.
- Arnaud Mary was a member of the Program Committee of *WEPA*.
- Sabine Peres was a member of the Program Committee of *MPA*.
- Nadia Pisanti was a member of the Program Committee of *CiE*, *CPM*, *ISBRA*, *SPIRE*, and *WABI*.
- Solon Pissis was a member of the Program Committee of *CiE*, *CPM*, *ECCB*, *SPIRE*, and *WABI*.
- Marie-France Sagot was a member of the Program Committee of *RecombCG*, and of *WEPA*.
- Blerina Sinaimeri was a member of the Program Committee of *IWOCA*.

Member of the editorial boards

- Roberto Grossi is member of the Editorial Board of *Theory of Computing Systems (TOCS)* and pf *RAIRO – Theoretical Informatics and Applications*.
- Giuseppe Italiano is member of the Editorial Board of *ACM Transactions on Algorithms* and of *Algorithmica* and *Theoretical Computer Science*.
- Vincent Lacroix is recommender for *Peer Community in Genomics*, see `https://genomics.peercommunityin.org/`.
- 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*, and *Lecture Notes in BioInformatics*.
- Blerina Sinaimeri is member of the Editorial Board of *Information Processing Letters* and of *Theoretical Computer Science*.
- 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, PLoS Computational Biology.

10.1.2 Invited talks

Sabine Peres gave an invited talk at the Institut Curie, and another at the Meetochondrie 2022 colloquium at Lège-Cap-Ferret. She also gave a talk at ICSB 2022 at Berlin, Germany.

10.1.3 Scientific expertise

Giuseppe Italiano is since 2020 Vice-President of the European Association for Theoretical Computer Science (EATCS). He is Director of the Master of Science in Data Science and Management, LUISS University, Rome, besides having a number of other responsabilities at LUISS. He is also member of the Advisory Board of MADALGO - Center for MAssive Data ALGOrithmics, Aarhus, Denmark.

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

Sabine Peres is since 2022 Head of the Master's degree in bioinformatics - University Lyon 1, member of the Advisory committee section 67-68 University Lyon 1, and internal member of the E2M2 doctoral school of the University of Lyon 1. She is also member of the coordination committee of DigitBioMed (Digital Sciences for Biology and Health) of the SFRI (Structuration de la Formation par la Recherche dans les Initiatives d'excellence). was member of the recruitment committee for a Professor position at Sorbonne University of Paris, and for an Associate Professor at Polytech, Nice.

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 since 2014 member of the Scientific Advisory Board of CWI, and since 2022 member of the Scientific Advisory Board of the Dept. of Computational Biology at the Univ. of Lausanne, Switzerland. Since 2022 also, she is member of the Scientific Advisory Board of the MATOMIC project funded by the Novo Nordisk Foundation, Denmark, and coordinated by Prof. Daniel Merkle, Univ. of South Denmark. Since 2020, she is member of the Review Committee for the Human Frontier Science Program. She was member was member of the recruitment committee for a Professor at LIRMM / University of Montpellier, and of Junior Researchers at Inria Lyon.

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.

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.

10.1.4 Research administration

Marie-France Sagot is since 2021, member of the "Conseil Scientifique (COS)" and of the "COmité des Moyens Incitatifs (COMI)" for Inria Lyon.

10.2 Teaching - Supervision - Juries

10.2.1 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.

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 2022), for the "Advanced machine learning theory" of the Master 2 "advanced maths" of the ENS de Lyon (6 hours in 2022), and taught at the Master 1 bioinformatics at

UCBL (6 hours in 2022). 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 2022.

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 2022.

Sabine Peres is responsible for four courses at the University, one at the Licence level and three at the Master level (L2: Mathematics life science, Python programming, M2 Bioinformatics: Modelling of metabolic networks; M2 Integrative Biology and Physiology: Modelling in Physiology, M2 Biodiversity, ecology and evolution: Python programming - simulation of population genetics).

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).

10.2.2 Supervision

The following are the PhDs in progress in 2022:

- 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)
- Njagi Mwaniki, Università di Pisa (supervisor: Nadia Pisanti)
- Luca Nesterenko, University of Lyon 1 (co-supervisors: Laurent Jacob; Bastien Boussau at the LBBE)
- Michelle Sweering, CWI (co-supervisors: Solon Pissis and Leen Stougie)
- Antoine Villie, University of Lyon 1 (supervisor: Laurent Jacob)

10.2.3 Juries

The following are the PhD and HDR juries to which members of ERABLE participated in 2022.

- Sabine Peres: Reviewer of the HDR pf Caroline Baroukh, University of Toulouse and INRAe Auzeville, March 2022; reviewer of the PhD of Laetitia Gibart, University of Sophia-Antipolis, Nov 2022; president of the PhD committee of Hugo Menet, University pf Lyon 1, July 2022; and member of the PhD committee of Samuel Buchet, École Normale de Nantes, March 2022.
- Marie-France Sagot: Reviewer of the HDR of Matthias Zytnicki, INRAe Toulouse, April 2022; Reviewer of the PhD od Marta Lucchetta, Università Degli Studi di Siena, Italy, January 2022; President of the PhD committee of Florian Ingels, École Normale Supérieure de Lyon, October 2022.

11 Scientific production

11.1 Publications of the year

International journals

[1] F. Almentina Ramos Shidi, A. Cologne, M. Delous, A. Besson, A. Putoux, A.-L. Leutenegger, V. Lacroix, P. Edery, S. Mazoyer and R. Bordonné. 'Mutations in the non-coding RNU4ATAC gene affect the homeostasis and function of the Integrator complex'. In: *Nucleic Acids Research* (20th Dec. 2022). DOI: 10.1093/nar/gkac1182. URL: https://hal.inria.fr/hal-03913654.

[2] G. Bernardini, P. Gawrychowski, N. Pisanti, S. Pissis and G. Rosone. 'Elastic-Degenerate String Matching via Fast Matrix Multiplication'. In: *SIAM Journal on Computing* 51.3 (June 2022), pp. 549–576. DOI: 10.1137/20M1368033. URL: https://hal.inria.fr/hal-03676475.

[3] A. Bertrand-Chapel, C. Caligaris, T. Fenouil, C. Savary, S. Aires, S. Martel, P. Huchedé, C. Chassot, V. Chauvet, V. Cardot-Ruffino, A.-P. Morel, F. Subtil, K. Mohkam, J.-Y. Mabrut, L. Tonon, A. Viari, P. Cassier, V. Hervieu, M. Castets, A. Mauviel, S. Sentis and L. Bartholin. 'SMAD2/3 mediate oncogenic effects of TGF-β in the absence of SMAD4'. In: *Communications Biology* 5.1 (7th Oct. 2022), p. 1068. DOI: 10.1038/s42003-022-03994-6. URL: https://hal.inria.fr/hal-03915470.

[4] A. Bodelón, M. Fablet, P. Veber, C. Vieira and M. P. García Guerreiro. 'High Stability of the Epigenome in Drosophila Interspecific Hybrids'. In: *Genome Biology and Evolution* 14.2 (10th Feb. 2022). DOI: 10.1093/gbe/evac024. URL: https://hal.archives-ouvertes.fr/hal-03594035.

[5] E. Bonnet, V. Haddad, S. Quesada, K.-A. Baffert, A. Lardy-Cléaud, I. Treilleux, D. Pissaloux, V. Attignon, Q. Wang, A. Buisson, P.-E. Heudel, T. Bachelot, A. Dufresne, L. Eberst, P. Toussaint, V. Bonadona, C. Lasset, A. Viari, E. Sohier, S. Paindavoine, V. Combaret, D. Pérol, I. Ray-Coquard, J.-Y. Blay and O. Trédan. 'Alterations in Homologous Recombination-Related Genes and Distinct Platinum Response in Metastatic Triple-Negative Breast Cancers: A Subgroup Analysis of the ProfiLER-01 Trial'. In: *Journal of Personalized Medicine* 12 (27th Sept. 2022). DOI: 10.3390/jpm12101595. URL: https://hal.inria.fr/hal-03915443.

[6] T. Bosman, M. van Ee, Y. Jiao, A. Marchetti-Spaccamela, R. Ravi and L. Stougie. 'Approximation Algorithms for Replenishment Problems with Fixed Turnover Times'. In: *Algorithmica* 84.9 (Sept. 2022), pp. 2597–2621. DOI: 10.1007/s00453-022-00974-4. URL: https://hal.inria.fr/hal-03832879.

[7] T. Calamoneri, A. Monti and B. Sinaimeri. 'On the domination number of t-constrained de Bruijn graphs'. In: *Discrete Mathematics and Theoretical Computer Science* (14th Aug. 2022). URL: https://hal.inria.fr/hal-03832873.

[8] S. Chakraborty, R. Grossi, K. Sadakane and S. R. Satti. 'Succinct representation for (non)deterministic finite automata'. In: *Journal of Computer and System Sciences* 131 (Feb. 2023), pp. 1–12. DOI: 10.1016/j.jcss.2022.07.002. URL: https://hal.inria.fr/hal-03913681.

[9] P. Charalampopoulos, C. Iliopoulos, T. Kociumaka, S. Pissis, J. Radoszewski and J. Straszyński. 'Efficient Computation of Sequence Mappability'. In: *Algorithmica* 84.5 (May 2022), pp. 1418–1440. DOI: 10.1007/s00453-022-00934-y. URL: https://hal.inria.fr/hal-03832866.

[10] A. Conte, R. Grossi, A. Marino, T. Uno and L. Versari. 'Proximity Search for Maximal Subgraph Enumeration'. In: *SIAM Journal on Computing* 51.5 (31st Oct. 2022), pp. 1580–1625. DOI: 10.1137/20M1375048. URL: https://hal.inria.fr/hal-03913673.

[11] M. G. Ferrarini, E. Dell'aglio, A. Vallier, S. Balmand, C. Vincent-Monégat, S. Hughes, B. Gillet, N. Parisot, A. Zaidman-Rémy, C. Vieira, A. Heddi and R. Rebollo. 'Efficient compartmentalization in insect bacteriomes protects symbiotic bacteria from host immune system'. In: *Microbiome* 10.1 (Dec. 2022), p. 156. DOI: 10.1186/s40168-022-01334-8. URL: https://hal.archives-ouvertes.fr/hal-03913752.

[12] A. a. G. Gabriel, J. R. Atkins, R. C. C. Penha, K. Smith-Byrne, V. Gaborieau, C. Voegele, B. Abedi-Ardekani, M. Milojevic, R. Olaso, V. Meyer, A. Boland, J. F. Deleuze, D. Zaridze, A. Mukeriya, B. Swiatkowska, V. Janout, M. Schejbalová, D. Mates, J. Stojšić, M. Ognjanovic, J. S. Witte, S. R. Rashkin, L. Kachuri, R. J. Hung, S. Kar, P. Brennan, A.-S. Sertier, A. Ferrari, A. Viari, M. Johansson, C. I. Amos, M. Foll and J. D. Mckay. 'Genetic Analysis of Lung Cancer and the Germline Impact on Somatic Mutation Burden'. In: *JNCI: Journal of the National Cancer Institute* 114.8 (1st Aug. 2022), pp. 1159–1166. DOI: 10.1093/jnci/djac087. URL: https://hal.inria.fr/hal-03914530.

[13] M. Galvão Ferrarini, I. Ziska, R. Andrade, A. Julien-Laferrière, L. Duchemin, R. M. César, A. Mary, S. Vinga and M.-F. Sagot. 'Totoro: Identifying Active Reactions During the Transient State for Metabolic Perturbations'. In: *Frontiers in Genetics* 13 (21st Feb. 2022), pp. 1–12. DOI: 10.3389/fgene.2022.815476. URL: https://hal.inria.fr/hal-03584295.

[14] G. Loukides and S. P. Pissis. 'All-pairs suffix/prefix in optimal time using Aho-Corasick space'. In: *Information Processing Letters* 178 (Nov. 2022), p. 106275. DOI: 10.1016/j.ipl.2022.106275. URL: https://hal.inria.fr/hal-03832860.

[15] C. Moraga, E. Sanchez, M. G. Ferrarini, R. A. Gutierrez, E. A. Vidal and M.-F. Sagot. 'BrumiR: A toolkit for de novo discovery of microRNAs from sRNA-seq data'. In: *GigaScience* 11 (25th Oct. 2022). DOI: 10.1093/gigascience/giac093. URL: https://hal.inria.fr/hal-03831360.

[16] N. Olver, F. Schalekamp, S. van Der Ster, L. Stougie and A. van Zuylen. 'A duality based 2-approximation algorithm for maximum agreement forest'. In: *Mathematical Programming* (21st Mar. 2022). DOI: 10.1007/s10107-022-01790-y. URL: https://hal.inria.fr/hal-03671089.

[17] E. Peter, I. Treilleux, V. Wucher, E. Jougla, A. Vogrig, D. Pissaloux, S. Paindavoine, J. Berthet, G. Picard, V. Rogemond, M. Villard, C. Vincent, L. Tonon, A. Viari, J. Honnorat, B. Dubois and V. Desestret. 'Immune and Genetic Signatures of Breast Carcinomas Triggering Anti-Yo–Associated Paraneoplastic Cerebellar Degeneration'. In: *Neurology Neuroimmunology & Neuroinflammation* 9.5 (12th July 2022), e200015. DOI: 10.1212/nxi.0000000000200015. URL: https://hal.inria.fr/hal-03915422.

[18] 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'. In: *Bioinformatics* (27th June 2022). DOI: 10.1093/bioinformatics/btac238. URL: https://hal.archives-ouvertes.fr/hal-03433563.

[19] Y. Wang, A. Mary, M.-F. Sagot and B. Sinaimeri. 'Efficiently sparse listing of classes of optimal cophylogeny reconciliations'. In: *Algorithms for Molecular Biology* 17.1 (Dec. 2022), pp. 1–16. DOI: 10.1186/s13015-022-00206-y. URL: https://hal.inria.fr/hal-03576371.

[20] H. Zhong, G. Loukides and S. P. Pissis. 'Clustering sequence graphs'. In: *Data and Knowledge Engineering* 138 (Mar. 2022), p. 101981. DOI: 10.1016/j.datak.2022.101981. URL: https://hal.inria.fr/hal-03832863.

International peer-reviewed conferences

[21] G. Bernardini, H. Chen, G. Loukides, S. P. Pissis, L. Stougie and M. Sweering. 'Making de Bruijn Graphs Eulerian'. In: CPM 2022 - 33rd Annual Symposium on Combinatorial Pattern Matching. Prague, Czech Republic, 27th June 2022. DOI: 10.4230/LIPIcs.CPM.2022.12. URL: https://hal.inria.fr/hal-03832887.

[22] G. Bernardini, A. Conte, E. Gabory, R. Grossi, G. Loukides, S. P. Pissis, G. Punzi and M. Sweering. 'On Strings Having the Same Length-k Substrings'. In: 33rd Annual Symposium on Combinatorial Pattern Matching (CPM). Prague, Czech Republic, 2022. DOI: 10.4230/LIPIcs.CPM.2022.16. URL: https://hal.inria.fr/hal-03829979.

[23] G. Bernardini, E. Gabory, S. Pissis, L. Stougie, M. Sweering and W. Zuba. 'Elastic-Degenerate String Matching with 1 Error'. In: LATIN 2022: Theoretical Informatics. Vol. LNCS - 13568. Lecture Notes in Computer Science. Guanajuato, Mexico: Springer International Publishing, 29th Oct. 2022, pp. 20–37. DOI: 10.1007/978-3-031-20624-5_2. URL: https://hal.inria.fr/hal-03913707.

[24] G. Bernardini, L. van Iersel, E. Julien and L. Stougie. 'Reconstructing Phylogenetic Networks via Cherry Picking and Machine Learning'. In: WABI 2022 - 2nd International Workshop on Algorithms in Bioinformatics. Potsdam, Germany, 5th Sept. 2022. DOI: 10.4230/LIPIcs.WABI.2022.16. URL: https://hal.inria.fr/hal-03832882.

[25] C. Brosse, V. Limouzy and A. Mary. 'Polynomial Delay Algorithm for Minimal Chordal Completions'. In: 49th International Colloquium on Automata, Languages, and Programming (ICALP). Paris, France, 2022. DOI: 10.4230/LIPIcs.ICALP.2022.33. URL: https://hal.inria.fr/hal-03829955.

[26] P. Charalampopoulos, T. Kociumaka, J. Radoszewski, S. P. Pissis, W. Rytter, T. Waleń and W. Zuba. 'Approximate Circular Pattern Matching'. In: ESA 2022 - 30th Annual European Symposium on Algorithms. Berlin/Potsdam, Germany, 2022. DOI: 10.4230/LIPIcs.ESA.2022.35. URL: https://hal.inria.fr/hal-03829963.

[27] P. Charalampopoulos, S. P. Pissis and J. Radoszewski. 'Longest Palindromic Substring in Sublinear Time'. In: 33rd Annual Symposium on Combinatorial Pattern Matching (CPM). Prague, Czech Republic, 2022. DOI: 10.4230/LIPIcs.CPM.2022.20. URL: https://hal.inria.fr/hal-03829971.

[28] L. Georgiadis, G. F. Italiano and E. Kosinas. 'Computing the 4-Edge-Connected Components of a Graph: An Experimental Study'. In: ESA 2022 - 30th Annual European Symposium on Algorithms. Berlin/Potsdam, Germany, 2022. DOI: 10.4230/LIPIcs.ESA.2022.60. URL: https://hal.inria.fr/hal-03829988.

[29] V. Guerrini, A. Conte, R. Grossi, G. Liti, G. Rosone and L. Tattini. 'phyBWT: Alignment-Free Phylogeny via eBWT Positional Clustering'. In: WABI 2022 - 22nd International Workshop on Algorithms in Bioinformatics. Berlin/Postdam, Germany, 2022. DOI: 10.4230/LIPIcs.WABI.2022.23. URL: https://hal.inria.fr/hal-03829984.

Reports & preprints

[30] L. Nesterenko, B. Boussau and L. Jacob. *Phyloformer: towards fast and accurate phylogeny estimation with self-attention networks*. 21st Nov. 2022. DOI: 10.1101/2022.06.24.496975. URL: https://hal-cnrs.archives-ouvertes.fr/hal-03756990.

[31] B. Sinaimeri, L. Urbini, M.-F. Sagot and C. Matias. *Cophylogeny Reconstruction Allowing for Multiple Associations Through Approximate Bayesian Computation*. 12th Oct. 2022. URL: https://hal.archives-ouvertes.fr/hal-03673256.

Other scientific publications

[32] N. Parisot, M. G. Ferrarini, C. Goubert, C. Vargas-Chávez, A. Vallier, B. Gillet, S. Hughes, P. Baa-Puyoulet, H. Charles, F. Calevro, R. Gil, A. Latorre, C. Vieira, R. Rebollo and A. Heddi. 'High-throughput genomics and transcriptomics to decipher host-symbiont molecular dialogue in the cereal weevil Sitophilus oryzae andSodalis pierantonius endosymbiotic association'. In: ISS 2022 - 10th Congress of the International Symbiosis Society. Lyon, France, 25th July 2022. URL: https://hal.inrae.fr/hal-03810657.