RESEARCH CENTRE

Inria Lyon Centre

IN PARTNERSHIP WITH:

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

2023

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 1

1 Team members, visitors, external collaborators 2

2 Overall objectives 3

3 Research program 4
3.1 Two main goals . . . 4
3.2 Different research axes . . . 4

4 Application domains 6
4.1 Biology and Health . . . 6

5 Social and environmental responsibility 7
5.1 Footprint of research activities . . . 7
5.2 Expected impact of research results . . . 7

6 Highlights of the year 8

7 New software, platforms, open data 8
7.1 New software . . . 8
7.1.1 AmoCoala . . . 8
7.1.2 BrumiR . . . 8
7.1.3 Caldera . . . 9
7.1.4 Capybara . . . 9
7.1.5 C3Part/Isofun . . . 9
7.1.6 Cassis . . . 10
7.1.7 Coala . . . 10
7.1.8 CSC . . . 10
7.1.9 Cycads . . . 10
7.1.10 DBGWAS . . . 11
7.1.11 Eucalypt . . . 11
7.1.12 Fast-SG . . . 11
7.1.13 Gobbolino-Touché . . . 11
7.1.14 HapCol . . . 12
7.1.15 HgLib . . . 12
7.1.16 KissDE . . . 12
7.1.17 KisSplice . . . 12
7.1.18 KisSplice2RefGenome . . . 13
7.1.19 KisSplice2RefTranscriptome . . . 13
7.1.20 MetExplore . . . 13
7.1.21 Mirinho . . . 14
7.1.22 Momo . . . 14
7.1.23 Moomin . . . 14
7.1.24 MultiPus . . . 15
7.1.25 paSAmcs . . . 15
7.1.26 Pitufolandia . . . 15
7.1.27 Sasita . . . 15
7.1.28 Smile . . . 16
7.1.29 Totoro . . . 16
7.1.30 Wengan . . . 16
7.1.31 WhatsHap . . . 16

8 New results **17**
8.1 General comments . . . 17
8.2 General theoretical result: Efficient enumeration of all solutions to a problem . . . 17
8.3 Axis 1: (Pan)Genomics and transcriptomics in general . . . 17
8.3.1 Identification and quantification of transposable element transcripts using Long-Read RNA-seq . . . 17
8.3.2 Comparing elastic-degenerate strings with an application to pangenomes . . . 18
8.4 Axis 2: Metabolism and (post)transcriptional regulation . . . 18
8.4.1 Metabolism: Hybrid modelling to Solve Optimal Concentrations of Metabolites and Enzymes in Constraint-based modelling . . . 18
8.4.2 Metabolism: Predicting the active reactions in a transient state between two conditions 19
8.4.3 Metabolism: Taking into account toxicity in a synthetic biology context . . . 19
8.4.4 Metabolism and tropical diseases . . . 20
8.4.5 Post-transcriptional regulation: MicroRNA Target Identification: Revisiting Accessibility and Seed Anchoring . . . 20
8.5 Axis 3: (Co)Evolution . . . 20
8.5.1 Phylogenetic networks: Constructing such via cherry picking and machine learning 20
8.5.2 Cophylogeny: Revisiting event probabilities allowing for species invasions (also termed spread) . . . 21
8.6 Axis 4: Health in general . . . 21

9 Partnerships and cooperations **22**
9.1 International initiatives . . . 22
9.1.1 Inria associate team not involved in an IIL or an international program . . . 22
9.1.2 Participation in other International Programs . . . 22
9.2 International research visitors . . . 22
9.2.1 Visits of international scientists . . . 22
9.2.2 Visits to international teams . . . 24
9.3 European initiatives . . . 25
9.3.1 H2020 projects . . . 25
9.4 National initiatives . . . 25
9.4.1 ANR . . . 25
9.4.2 Others . . . 26

10 Dissemination **27**
10.1 Promoting scientific activities . . . 27
10.1.1 Scientific events: organisation . . . 27
10.1.2 Journal . . . 28
10.1.3 Invited talks . . . 28
10.1.4 Scientific expertise . . . 28
10.1.5 Research administration . . . 29
10.1.6 International school organisation . . . 29
10.2 Teaching - Supervision - Juries . . . 29
10.2.1 Teaching . . . 29
10.2.2 Supervision . . . 30
10.2.3 Juries . . . 31

11 Scientific production **31**
11.1 Publications of the year . . . 31

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]
- Mariana Ferrarini [INRIA, Advanced Research Position]
- Laurent Jacob [CNRS, Researcher, until Jun 2023, HDR]
- Solon Pissis [CWI, Senior Researcher]
- Leen Stougie [CWI, Senior Researcher]
- Alain Viari [INRIA, Senior Researcher]

Faculty Members

- Roberto Grossi [UNIV PISA, Professor]
- Giuseppe Italiano [UNIV LUISS, Professor]
- Vincent Lacroix [UNIV LYON I, Associate Professor, HDR]
- Alberto Marchetti Spaccamela [SAPIENZA ROME, Professor]
- Arnaud Mary [UNIV LYON I, Associate Professor]
- Sabine Peres [UNIV LYON I, Professor, HDR]
- Nadia Pisanti [UNIV PISA, Associate Professor]
- Blerina Sinaimeri [LUISS University Rome, in detachment from INRIA, Associate Professor]
- Cristina Vieira [UNIV LYON I, Associate Professor, HDR]

PhD Students

- Emma Crisci [INRIA, from Oct 2023]
- Sasha Darmon [UNIV LYON I, from Oct 2023]
- Nicolas Homberg [INRIA, until Apr 2023]
- Maxime Mahout [INRIA, from Oct 2023 until Nov 2023]
- Maxime Mahout [UNIV PARIS SACLAY, until Sep 2023]
- Luca Nesterenko [CNRS, until Jun 2023]
- Camille Siharath [UNIV LYON I, from Oct 2023]
- Antoine Villié [CNRS, until Apr 2023]

Technical Staff

- François Gindraud [INRIA, Engineer]

Interns and Apprentices

- Pierre Gerenton [CNRS, from Feb 2023 until Nov 2023]
- Jeremie Muller-Prokob [AVIESAN, from Feb 2023 until Jul 2023]
- Pierre-Antoine Navarro [INRIA, Intern, from Apr 2023 until Jul 2023]
- Camille Siharath [AVIESAN, from Feb 2023 until Jul 2023]
- Johanna Trost [CNRS, until Mar 2023]

Administrative Assistant

- Anouchka Ronceray [INRIA]

External Collaborators

- Laurent Jacob [CNRS, from Jul 2023, Laurent having had to move to Paris for family reasons, he is now an external collaborator of ERABLE.]
- Susana Vinga [ULISBOA]

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 may be artificially split into two different Axes.

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 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) were *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.

We do however wish to call attention to the fact that in 2023, two members of the team defended their HDR ("Habilitation à Diriger des Recherches"). The first was Laurent Jacob, who defended in April 4. L. Jacob had already co-supervised 4 PhD students since 2016, 2 as actually main supervisor, and he is currently co-supervising a fifth PhD student. At the time of his HDR defense, L. Jacob was still full member of ERABLE. The second HDR was defended by Vincent Lacroix on July 5. Previous to this, V. Lacroix had already officially co-supervised 6 PhD students since 2010 who have already defended, 5 of which as actually main supervisor. He is currently main supervisor of a PhD which just started, namely of Sasha Darmon.

7 New software, platforms, open data

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://team.inria.fr/erable/en/software/amocoala/

Contact: Blerina Sinaimeri

Participants: Laura Urbini, Marie-France Sagot, Catherine Matias, 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

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 paSAmcs

Keyword: Metabolism

Functional Description: Computation of Minimal Cut Sets using Answer Set Programming (ASP), and more precisely aspefm.

URL: https://github.com/maxm4/paSAmcs

Contact: Sabine Peres

Participants: Sabine Peres, Maxime Mahout

7.1.26 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

7.1.27 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

7.1.28 Smile

Keywords: Bioinformatics, Genomic sequence

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

URL: https://gitlab.inria.fr/nhomberg/smile

Contact: Marie-France Sagot

Participants: Marie-France Sagot, Nicolas Homberg

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

No open data in the case of ERABLE.

8 New results

8.1 General comments

We present in this section the main results obtained in 2023.

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.

We would like also to call attention to two main facts.

The first one was already pointed out in our reports for the previous years. It concerns the fact that we choose in general not detail the results on more theoretical aspects of computer science when these are initially addressed in contexts not directly related to computational biology 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. Examples of these are [2, 4, 15, 6, 17]. We also chose not to detail the results concerning a Python package for the statistical analysis of networks, including biological ones, and more specifically in the case of this paper, of the REACTOME [8], as well as results related to text algorithms even though these may, or have already more direct applications in biology [1, 13, 14, 5, 18].

This year, there is an exception to that in the sense that we obtained a result – theoretical – that provides a general framework for enumerating equivalence classes of solutions. Enumeration of all solutions to a problem has since a very long time been one of the major theoretical and applied interests of the team. This result has already been shown to be important in different aspects of computational biology that are of the team's interest. Because of this, we chose to provide more details on the paper [12] that was accepted this year in *Algorithmica* in a special section that in a way concerns all our main four axes of research and that is presented before the sections devoted to such.

The second fact we want to call attention to is that 2023 represents a transition period for the ERABLE team. Indeed, due to the fact that in the next couple of years, various of the more senior members will retire (namely, Alberto Marchetti-Spaccamela, Leen Stougie, Alain Viari, and the team's leader Marie-France Sagot), there will be many changes in the overall composition of the team and in the scientific topics it continues to address. Already this year although for another reason, we saw the departure of one member of the team, Laurent Jacob, who for family matters moved to Paris at the end of June 2023, which implied also in the full move of one of his PhD students, Luca Nesterenko, who had been a member of ERABLE to another team.

8.2 General theoretical result: Efficient enumeration of all solutions to a problem

When a problem has more than one solution, it is often important, depending on the underlying context, to enumerate (*i.e.*, to list) them all. Even when the enumeration can be done in polynomial delay, that is, spending no more than polynomial time to go from one solution to the next, this can be costly as the number of solutions themselves may be huge, including sometimes exponential. Furthermore, depending on the application, many of these solutions can be considered equivalent. The problem of an efficient enumeration of the 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 the paper [12], we provided a general framework that solves this problem in polynomial delay for a wide variety of contexts, including optimization ones that can be addressed by dynamic programming algorithms, and for certain types of equivalence relations between solutions. In order to reach this goal, we went through an intermediate problem, namely the enumeration of coloured subtrees in acyclic decomposable AND/OR graphs (ad-AND/OR graph).

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

8.3.1 Identification and quantification of transposable element transcripts using Long-Read RNA-seq

Participants: Vincent Lacroix, Arnaud Mary, Cristina Vieira.

Transposable elements (TEs) are repeated DNA sequences potentially able to move throughout the genome. In addition to their inherent mutagenic effects, TEs can disrupt nearby genes by donating their intrinsic regulatory sequences, for instance, promoting the ectopic expression of a cellular gene. TE transcription is therefore not only necessary for TE transposition per se but can also be associated with TE-gene fusion transcripts, and in some cases, be the product of pervasive transcription. Hence, correctly determining the transcription state of a TE copy is essential to apprehend the impact of the TE in the host genome. Methods to identify and quantify TE transcription have mostly relied on short RNA-seq reads to estimate TE expression at the family level while using specific algorithms to discriminate copy-specific transcription. However, assigning short reads to their correct genomic location, and genomic feature is not trivial. In a paper submitted in 2023 which is under revision (see the bioRxiv version the here), we retrieved full-length cDNA (TeloPrime, Lexogen) of *Drosophila melanogaster* gonads and sequenced them using Oxford Nanopore Technologies. We showed that long-read RNA-seq can be used to identify and quantify transcribed TEs at the copy level. In particular, TE insertions overlapping annotated genes are better estimated using long reads than short reads. Nevertheless, long TE transcripts (> 4.5 kb) are not well captured. Most expressed TE insertions correspond to copies that have lost their ability to transpose, and within a family, only a few copies are indeed expressed. Long-read sequencing also allowed the identification of spliced transcripts for around 105 TE copies. Overall, this first comparison of TEs between testes and ovaries uncovers differences in their transcriptional landscape, at the subclass and insertion level.

8.3.2 Comparing elastic-degenerate strings with an application to pangenomes

Participants: Nadia Pisanti, Solon Pissis.

Sequence (or string) comparison is a fundamental task in computer science, with numerous applications notably in computational biology. Given two or more sequences and a distance function, the task is to compare the sequences in order to infer or visualise their (dis)similarities. Many sequence representations have been introduced over the years to account for unknown or uncertain letters, a phenomenon that often occurs in data that come from experiments. In the context of computational biology, for example, the IUPAC notation is used to represent locations of a DNA sequence for which several alternative nucleotides are possible. This gives rise to the notion of degenerate string (or indeterminate string): a sequence of finite sets of letters. When all sets are of size 1, we are in the special case of a standard string (or deterministic string). Degenerate strings can encode the consensus of a population of DNA sequences in a gapless multiple sequence alignment (MSA). Iliopoulos *et al.* (*Information and Computation*, 279:104616, 2021. doi:10.1016/j.ic.2020. 104616) generalised this notion to also encode insertions and deletions (gaps) occurring in MSAs by introducing the notion of elastic-degenerate string: a sequence of finite sets of strings. The main motivation to consider elastic-degenerate (ED) strings is that they can be used to represent a pangenome: a collection of closely-related genomic sequences that are meant to be analysed together. In the paper [16], we showed different results related to the comparison of pangenomes represented as ED strings.

8.4 Axis 2: Metabolism and (post)transcriptional regulation

8.4.1 Metabolism: Hybrid modelling to Solve Optimal Concentrations of Metabolites and Enzymes in Constraint-based modelling

Participants: Sabine Peres.

Constraint-based modelling is a widely used approach to analyse genotype-phenotype relationships. The main key concepts are stoichiometric analysis such as flux balance analysis (FBA), Resource Balance Analysis (RBA) or elementary flux mode (EFM) analysis. While FBA identifies optimal flux distribution with respect to a given objective, EFMs characterize all the solution space in terms of minimal pathways

but their number leads to a combinatorial explosion for large networks. RBA predicts for a specific environment, the set of possible cell configurations compatible with the available resources and extends very significantly the predictive power of FBA. However, when stoichiometric and kinetic constraints are considered together, the set of possible flux configurations is in general not convex since the kinetic functions are not linear. The problem resolution has thus multiple local maxima. Recent works showed that the optimal solution of constraint enzyme allocation problems with general kinetics is an EFM analysis. Based on this recent outcome, we decided to write the resource allocation constraint on the kinetic optimization problem into a geometric problem in an EFM analysis, *i.e.* a convex optimal problem that is easily solved. To predict optimal flux modes, we thus compute constrained EFMs with our tool ASPEFM based on Answer Set Programming to save time and space computation. ASPEFM allows the integration of Boolean and linear constraints such as thermodynamic, environment, transcriptomic regulatory rules, and resource operating cost (that identify the most efficient EFMs for converting substrate into biomass) using the solver CLINGOLP which combines logic and linear programming. The convex optimisation problem is then resolved on each constrained EFM which provides for this mode, the optimal repartition of resources among enzymes and the associated metabolite concentrations. We applied our method to the central carbon metabolism of *Escherichia coli*, with a detailed model of the respiration chains, ATPase (including explicitly the proton motive force). The optimal flux mode is the overflow of acetate which is in agreement with known experimental results. This approach allowed us to explore whether certain experimental properties observed on *E. coli* are consistent and what are the consequences of an optimal repartition of bacterial resources. Our method is very promising in synthetic biology and increased the ability to efficiently design biological systems. It was presented at BIOSTEC [19]. A paper is in preparation.

8.4.2 Metabolism: Predicting the active reactions in a transient state between two conditions

Participants: Mariana Galvão Ferrarini, François Gindraud, Arnaud Mary, Marie-France Sagot.

We are currently working on a method that would enable to take into account at the same time metabolomic and transcriptomic data in order to predict the reactions that were active during a transient state between two conditions instead of each type of data separately as was the case of two method previously developed in the team, namely TOTORO and MOOMIN. The first indeed integrates only concentrations of internal metabolites and the second only differential expression, in both cases measured before and after a perturbation, into a genome-scale metabolic reconstruction. We wish now to be able to consider both types of data simultaneously, a non-trivial modelling problem. 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. The members of ERABLE involved are M. Galvão Ferrarini, A. Mary and M.-F. Sagot.

8.4.3 Metabolism: Taking into account toxicity in a synthetic biology context

Participants: Mariana Galvão Ferrarini, François Gindraud, Arnaud Mary, Marie-France Sagot, Susana Vinga.

In parallel to the above, we 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 should happened within the context of a sabbatical of Nuno Mira, a professor from Instituto Superior Técnico in Lisbon, within ERABLE due to take place from October 2022 to September 2023 but which had to be cancelled by Nuno because of family problems. We did pick it up with again Henri Taneli Pusa and also with Susana Vinga, and intend to pursue it in 2024, hopefully with N. Mira even if he cannot have a sabbatical anymore.

All the methods developed in the past related to metabolism are currently been adapted, notably with the help of a permanent Inria engineer, François Gindraud, to become more user-friendly and integrated

within a same framework.

8.4.4 Metabolism and tropical diseases

Participants: Mariana Galvão Ferrarini, Arnaud Mary, Gabriela Torres Montanaro, Marie-France Sagot, Ariel Silber.

Finally, in the context of both the Inria Associated Team Capoeira, and of a PhD by Gabriela T. Montanaro, co-supervised between Ariel M. Silber, Professor at the University of São Paulo, Brazil, and M.-F. Sagot, ERABLE is working on problems related with metabolism and tropical diseases, in the case linked to *Trypanosoma cruzi*. In 2022, both A. M. Silber and G. T. Montanaro made regular more or less long visits to Lyon. In 2023, G. T. Montanaro stayed in Brazil to conduct experiments in the laboratory of A. M. Silber. She will renew her visits to Lyon later in 2024.

8.4.5 Post-transcriptional regulation: MicroRNA Target Identification: Revisiting Accessibility and Seed Anchoring

Participants: Mariana Galvão Ferrarini, Nicolas Homberg, Marie-France Sagot.

By pairing to messenger RNAs (mRNAs for short), microRNAs (miRNAs) regulate gene expression in animals and plants. Accurately identifying which mRNAs interact with a given miRNA and the precise location of the interaction sites is crucial to reaching a more complete view of the regulatory network of an organism. Only a few experimental approaches, however, allow the identification of both within a single experiment. Computational predictions of miRNA-mRNA interactions thus remain generally the first step used, despite their drawback of a high rate of false-positive predictions. The major computational approaches available rely on a diversity of features, among which anchoring the miRNA seed and measuring mRNA accessibility are the key ones, with the first being universally used, while the use of the second remains controversial. Revisiting the importance of each was the aim of our paper [7], which used Cross-Linking, Ligation, And Sequencing of Hybrids (CLASH) datasets to achieve this goal. Contrary to what might be expected, the results were more ambiguous regarding the use of the seed match as a feature, while accessibility appeared to be a feature worth considering, indicating that, at least under some conditions, it may favour anchoring by miRNAs.

This work was part also of the PhD defense of N. Homberg [20] which took place on June 15.

8.5 Axis 3: (Co)Evolution

8.5.1 Phylogenetic networks: Constructing such via cherry picking and machine learning

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. Existing methods are computationally expensive and can either handle only small numbers of phylogenetic trees or are limited to severely restricted classes of networks. In the paper [3], we applied the recently-introduced theoretical framework of cherry picking to design a class of efficient heuristics that are guaranteed to produce a network containing each of the input trees, for practical-size datasets consisting of binary trees. Some of the heuristics in this framework are based on 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. We also proposed simple and fast randomised heuristics that proved to be very effective when run multiple times. Unlike the existing exact methods, our heuristics are applicable to datasets of practical size, and the experimental study we conducted on both simulated and real data shows that these solutions are qualitatively good,

always within some small constant factor from the optimum. Moreover, our machine-learned heuristics are one of the first applications of machine learning to phylogenetics and show its promise.

8.5.2 Cophylogeny: Revisiting event probabilities allowing for species invasions (also termed spread)

Participants: Marie-France Sagot, Blerina Sinaimeri.

Phylogenetic tree reconciliation is extensively employed for the examination of coevolution between host and symbiont species. An important concern is the requirement for dependable cost values when selecting event-based parsimonious reconciliation. Although certain approaches deduce event probabilities unique to each pair of host and symbiont trees, which can subsequently be converted into cost values, a significant limitation lies in their inability to model the *invasion* of diverse host species by the same symbiont species (termed as a spread event), which is believed to occur in symbiotic relationships. Invasions lead to the observation of multiple associations between symbionts and their hosts (indicating that a symbiont is no longer exclusive to a single host), which are incompatible with the existing methods of coevolution. In the paper [10], we presented a method called AMOCOALA (an enhanced version of the tool COALA) that provides a more realistic estimation of cophylogeny event probabilities for a given pair of host and symbiont trees, even in the presence of spread events. We expanded the classical 4-event coevolutionary model to include 2 additional spread events (vertical and horizontal spreads) that lead to multiple associations. In the initial step, we estimated the probabilities of spread events using heuristic frequencies. Subsequently, in the second step, we employed an approximate Bayesian computation (ABC) approach to infer the probabilities of the remaining 4 classical events (cospeciation, duplication, host switch, and loss) based on these values. By incorporating spread events, our reconciliation model enables a more accurate consideration of multiple associations. This improvement enhances the precision of estimated cost sets, paving the way to a more reliable reconciliation of host and symbiont trees. To validate our method, we conducted experiments on synthetic datasets and demonstrated its efficacy using real-world examples. Our results showcase that AMOCOALA produces biologically plausible reconciliation scenarios, further emphasizing its effectiveness.

8.6 Axis 4: Health in general

Tropical diseases

Participants: Mariana G. Ferrarini, Arnaud Mary, Marie-France Sagot.

One of the main works in the area of health is related 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 2023, such as [9] but also others. We highlight the results of two main ones below.

In the paper that may be found here, results using genomic, transcriptomic and epigenetic data are presented on Gynecologic CarcinoSarcoma (CS), a rare cancer composed of both carcinomatous and sarcomatous malignant components. Reconstructions of the evolutionary history of these tumours revealed that each component is composed of both ancestral cell populations and component-specific

subclones, supporting a common origin followed by distinct evolutionary trajectories. Epithelial-to-Mesenchymal Transition (EMT) appears as a common mechanism associated with this phenotypic divergence, linking CS heterogeneity to genetic, transcriptomic but also epigenetic influences. This work represents the latest contribution of the Gilles Thomas Platform at the Centre Léon Bérard to the International Cancer Genome Consortium (ICGC) program which started in 2008. Previous contributions included studies of: (1) HER2+ Breast Cancers (Ferrari et al. 2016), (2) Prostate Cancer (Tonon et al. 2019), and finally, (3) Retinoblastoma (Liu et al. 2021).

On the other hand, the paper that may be found here results from a long lasting collaboration with the team of Véronique Maguer-Satta at CRCL/CLB Lyon. It aimed at defining a gene expression signature based on immunological markers of stem cell properties in order to predict patient outcome and drug efficiency, regardless of the tumour stage. The signature was trained on Breast Cancers but further successfully validated on a larger pan-cancer cohort (more than 10^4 samples).

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.

Webpage: 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 participants: M. Ferrarini, A. Mary, M.-F. Sagot, B. Sinaimeri.

Type: Capes-Cofecub (2020-2022, extended until 2023 and then possibly further to 2024 due to the pandemic).

Webpage: Ahimsa

9.2 International research visitors

9.2.1 Visits of international scientists

Alex di Genova and Carol Moraga Quinteros

Status: Both now Associate professors (at the time of the visit, Carol was still post-doc but has since obtained a permanent position)

Institution of origin: University O'Higgins

Country: Chile

Dates: Jan. 9 to 22

Context of the visit: Collaboration

Mobility program/type of mobility: Research stay

Ariel Mariano Silber

Status: Professor

Institution of origin: University of São Paulo

Country: Brazil

Dates: Two visits of approximately 2 months (Jan. 16 to Mar. 19) and 2 weeks (Sep. 23 to Oct. 7) respectively

Context of the visit: Collaboration

Mobility program/type of mobility: Research stay

Henri Taneli Pusa

Status: Postdoc

Institution of origin: Aalto University

Country: Finland

Dates: Three visits of approximately 1 to 2 weeks each time, from Feb. 20 to Mar. 1, then from Jun. 5 to 10, and finally from Sep. 26 to Oct. 5

Context of the visit: Collaboration

Mobility program/type of mobility: Research stay

Renata Wassermann

Status: Associate professor

Institution of origin: University of São Paulo

Country: Brazil

Dates: Sep. 3 to 10

Context of the visit: Collaboration

Mobility program/type of mobility: Research stay

André Fujita

Status: Associate professor

Institution of origin: University of São Paulo

Country: Brazil

Dates: Sep. 26 to Oct. 6

Context of the visit: Collaboration

Mobility program/type of mobility: Research stay

Andréa Ávila

Status: Senior researcher

Institution of origin: Instituto de Biologia Molecular do Paraná - Fiocruz-PR, Curitiba, Paraná

Country: Brazil

Dates: Sep. 26 to Oct. 6

Context of the visit: Collaboration

Mobility program/type of mobility: Research stay

Erida Gjini

Status: Researcher

Institution of origin: Instituto Superior Técnico, Lisbon

Country: Portugal

Dates: Oct. 9 to 13

Context of the visit: Initiation of collaboration

Mobility program/type of mobility: Research stay

Luís Felipe Ignácio Cunha

Status: Associate professor

Institution of origin: Federal University of Fluminense

Country: Brazil

Dates: Nov. 8 to 13

Context of the visit: Initiation of collaboration

Mobility program/type of mobility: Research stay

Besides the above, we had also in 2023 two visits to Lyon of Susana Vinga, one of our external collaborators and the coordinator of the European Twinning project Olissipo to which ERABLE also participates. The first visit of a few days (Mar. 28 to Apr. 1) was in the context of the PhD defence of Antoine Villié which took place on Mar. 31, while the second (Jul. 23 to 27) happened in the context of the ISMB/ECCB conference to which some of the members of ERABLE also participated, and notably M.-F. Sagot as co-organiser of the Special Session: Bioinformatics in France of the conference.

9.2.2 Visits to international teams

Maxime Mahout

Visited institution: University of São Paulo

Country: Brazil

Dates: Jun. 26 to Jul. 10

Context of the visit: Initiation of collaboration in view of applying for a postdoc at the University of São Paulo after his PhD defense which took place in November 2023

Mobility program/type of mobility: Research stay

Here again, in the context of the European Twinning project Olissipo, there were moreover two visits to Lisbon by some members of ERABLE, both of them linked to the schools we organised together with Susana Vinga and the Olissipo project manager, Sara Ramalho Tanqueiro, the first from Feb. 5 to 10 on Modelling and Analysis of Single Cell Multiple Biological Omics and the second from Jul. 2 to 7 on Computational phylogenetics to analyse the evolution of cells and communities - Tree for a Tango School. M.-F. Sagot thus visited the Instituto Superior Técnico (IST) from Feb. 1 to 11 and again later from Jun. 28 to Jul. 10 to discuss various organisational aspects of Olissipo as well as ideas for new scientific projects to submit in the future involving both IST and Inria. In February, the visit was done with also Ariel M. Silber from the Inria Associated Team Capoeira and Capes/Cofecub project Ahimsa, and in July with Blerina Sinaimeri and Mariana G. Ferrarini.

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. For the same reason, although it should have lasted until the end of 2023, it was extended until the end of June 2024.

Webpages: Olissipo-Erable and Olissipo

Besides Olissipo, three members of ERABLE, Nadia Pisanti in Italy, and Solon Pissis and Leen Stougie in the Netherlands, are partners of the EU MSCA-ITN-2020 project (2020-2024) called ALgorithms for PAngenome Computational Analysis (ALPACA) coordinated by Alexander Schoenhuth (University of Bielefeld, Germany). The webpage of ALPACA may be found here.

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.

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:

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 Conference on Algorithms and Complexity (CIAC)*.
- 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 co-organiser of the JGA (Journées Graphes el Algorithmes) 2023, held November 21-24, 2023, in Lyon.
- Solon Pissi was the chief organiser of the 2023 ALGO conference, held September 4-8, 2023, at CWI in Amsterdam.
- Marie-France Sagot was co-organiser of the Third Edition of the Workshop Metabolism and mathematical models: Two for a tango, held virtually, Nov 14-15, 2023. She is co-organiser of the recurrent Small non-coding RNA bioinformatics club since 2021.
- Leen Stougie was co-organiser of the 2023 ALGO conference, held September 4-8, 2023, at CWI in Amsterdam.

Member of the conference program committees

- Giuseppe Italiano was a member of the Program Committee of *ESA*, *LAGOS*, *SEA*, *SOSA*, and *STOC*.
- Nadia Pisanti was a member of the Program Committee of *RECOMB*.
- Solon Pissis was co-chair of the Program Committee of *PSC* and member of the Program committee of *WABI*.
- Marie-France Sagot was a member of the Program Committee of *ISMB/ECCB Special Session of Bioinformatics in France*, and of *PSC*.
- Blerina Sinaimeri was a member of the Program Committee of *CIAC*, and of *ICTCS*.
- Leen Stougie was member of the Program Committee of the COSI on Systems Biology and Networks at *ISMB/ECCB*.

10.1.2 Journal

Member of the editorial boards

- Roberto Grossi is member of the Editorial Board of *Theory of Computing Systems (TOCS) and of RAIRO - Theoretical Informatics and Applications*.
- Giuseppe Italiano is member of the Editorial Board of *ACM Transactions on Algorithms*, of *Algorithmica* and *Theoretical Computer Science*.
- Vincent Lacroix is recommender for *Peer Community in Genomics*.
- Alberto Marchetti-Spaccamela is member of the Editorial Board of *Theoretical Computer Science*.
- Arnaud Mary is guest editor of the special issue "WEPA22" for *Discrete Applied Mathematics*.
- 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*, *Computer Science Review*, 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*, *Annals of Operations Research*, *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*.

10.1.3 Invited talks

Vincent Lacroix gave an invited talk at the Laboratoire d'Écologie Alpine (LECA), University of Grenoble, on April 27.

Arnaud Mary gave an invited talk at the "Graphes et Bioinformatiques" day, Paris, November 8.

Leen Stougie gave the invited plenary lecture at the Fourth International Workshop on Dynamic Scheduling, June 5-6, 2023, Winterthur, Switzerland.

10.1.4 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.

Vincent Lacroix is responsible together with Arnaud Mary for the 1st year of the Master's degree in bioinformatics - University Lyon 1. He is also member of the Advisory committee section 67-68 of the University Lyon 1 and internal member of the E2M2 doctoral school of the University of Lyon 1

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). She 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 and until 2023 included, she was member of the Review Committee for the Human Frontier Science Program. She was member of the recruitment committee for Junior Researchers at Inria Lyon.

Leen Stougie was member of the General Board of the Dutch Network on the Mathematics of Operations Research (Landelijk Netwerk Mathematische Besliskunde (LNMB)). He is 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.

10.1.5 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.

10.1.6 International school organisation

In the context of the European Twinning project Olissipo coordinated by Susana Vinga, Marie-France Sagot was co-organiser of two international schools, one which took place from Feb. 5 to 10 on Modelling and Analysis of Single Cell Multiple Biological Omics and the second from Jul. 2 to 7 on Computational phylogenetics to analyse the evolution of cells and communities - Tree for a Tango School. Blerina Sinaimeri was also co-organiser of this second school.

In the context of the EU-projects ALPACA and PANGAIA to which members of ERABLE participate, Solon Pissis and Leen Stougie co-organised a Winterschool at CWI, Amsterdam, November 20-24, 2023.

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.

Vincent Lacroix is co-responsible for the M1 master in bioinformatics (together with Arnaud Mary) and responsible for 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 2023. Since 2021, he is also involved in the group who proposed a new course called Climate and Transitions, mandatory for L1 students in Science at University Lyon1 (1500 students). Most of the course is a MOOC, but there are also 4 occasions where teachers and students discuss the topics covered by the course with various group activities described briefly here Since 2023, the course is also proposed as an optional course for students at Université Lyon 2.

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 2023.

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). She was also invited to give tutorial classes at a thematic research school called "BioRegul: Modélisation formelle de réseaux de régulation biologique" that took place at Porquerolles in June 2023.

Notice that Laurent Jacob was responsible for different courses at the UCBL and the ENS Lyon until his departure for Paris for family reasons. He is now located at the located at the Laboratory of Computational and Quantitative Biology of the Sorbonne University in Paris.

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 at CWI teach at universities between 50 and 80 hours per year, again at the undergraduate and Master levels, in applied mathematics (*e.g.* Operations Research, Advanced Linear Programming), computer science (basic course in Python) and computational biology (*e.g.* Stringology).

10.2.2 Supervision

The following are the PhDs in progress or which ended in 2023:

- Emma Crisci, University of Lyon 1 (funded by Inria, co-supervisors: Sabine Peres and Arnaud Mary), started in October 2023.
- Sasha Darmon, University of Lyon 1 (co-supervisors: Vincent Lacroix and Arnaud Mary), started in October 2023.
- 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), PhD defended in June [20].
- Maxime Mahout, University Paris-Saclay (supervisor: Sabine Peres), PhD defended in November, manuscript available here.
- Moses Njagi Mwaniki, Università di Pisa (supervisor: Nadia Pisanti).
- Luca Nesterenko, University of Lyon 1 (co-supervisors: Laurent Jacob; Bastien Boussau at the LBBE), left ERABLE (although remaining in Lyon) when L. Jacob moved to Paris for family reasons.
- Luca Pepé Sciarria, University of Rome Tor Vergata (supervisor: Giuseppe F. Italiano), PhD defended in July.
- Camille Siharat, University of Lyon 1 (co-supervirors: Sabine Peres and Olivier Bondi, Université Évry Val-Essonne), started in October 2023.
- Michelle Sweering, CWI (co-supervisors: Solon Pissis and Leen Stougie).
- Antoine Villie, University of Lyon 1 (supervisor: Laurent Jacob), PhD defended in March, the PhD manuscript is not yet publicly available but part of the work it covered may be found in this paper [11].
- Hilde Verbeek, CWI (Supervisor: Solon Pissis, co-supervisor: Leen Stougie).

10.2.3 Juries

The following are the PhD and HDR juries to which members of ERABLE participated in 2023:

- Sabine Peres: Reviewer of the PhD of Marie Burel, Paris-Saclay University, June 2023; Reviewer of the PhD of Sahar Aghakhani, Paris-Saclay University, September 2023; Reviewer of the PhD of Bianca Buchner, Vienna University, October 2023; Reviewer of the PhD of Clémence Dupond Thibert, CEA Grenoble, December 2023; Reviewer of the PhD of Léon Faure, INRAe Jouy-en-Josas, December 2023; and member of the PhD of Sahar Aghakhani, Paris-Saclay University, September 2023.
- Vincent Lacroix: Reviewer of the PhD of Louison Fresnais, INRAe and Institut national polytechnique Toulouse, and L'Oréal, December 2023.
- Marie-France Sagot: Reviewer of the HDR of Sarah Djebali, IRSD-Inserm Toulouse, October 2023; Reviewer of the PhD of Bertrand Marchand, Institut Polytechnique of Paris, September 2023; Reviewer of the PhD of Darryl Ondoua, Sorbonne University, Paris, October 2023.
- Leen Stougie: Chair of the PhD-committee of Irving van Heuven van Staereling, Vrije Universiteit, Amsterdam, September 2023; member of the PhD-commmittee of Danny Blom, Technische Universiteit Eindhoven, December 2023.

11 Scientific production

11.1 Publications of the year

International journals

[1] L. a. K. Ayad, R. Chikhi and S. P. Pissis. 'Seedability: optimizing alignment parameters for sensitive sequence comparison'. In: *Bioinformatics Advances* 3.1 (1st Jan. 2023). DOI: 10.1093/bioadv/vbad108. URL: https://inria.hal.science/hal-04385612.

[2] S. Baruah and A. Marchetti-Spaccamela. 'The Computational Complexity of Feasibility Analysis for Conditional DAG Tasks'. In: *ACM Transactions on Parallel Computing* 10 (21st Sept. 2023), pp. 1–22. DOI: 10.1145/3606342. URL: https://inria.hal.science/hal-04365671.

[3] G. Bernardini, L. van Iersel, E. Julien and L. Stougie. 'Constructing phylogenetic networks via cherry picking and machine learning'. In: *Algorithms for Molecular Biology* 18 (16th Sept. 2023). DOI: 10.1186/s13015-023-00233-3. URL: https://inria.hal.science/hal-04365666.

[4] M. Bernaschi, A. Celestini, M. Cianfriglia, S. Guarino, G. F. Italiano, E. Mastrostefano and L. R. Zastrow. 'Seeking critical nodes in digraphs'. In: *Journal of computational science* 69 (31st Mar. 2023). DOI: 10.1016/j.jocs.2023.102012. URL: https://hal.science/hal-04365646.

[5] V. R. Carr, S. P. Pissis, P. Mullany, S. Shoaie, D. Gomez-Cabrero and D. L. Moyes. 'Palidis: fast discovery of novel insertion sequences'. In: *Microbial Genomics* 9.3 (14th Mar. 2023). DOI: 10.1099/mgen.0.000917. URL: https://inria.hal.science/hal-04392744.

[6] 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://inria.hal.science/hal-03913681.

[7] N. Homberg, M. Galvão Ferrarini, C. Gaspin and M.-F. Sagot. 'MicroRNA Target Identification: Revisiting Accessibility and Seed Anchoring'. In: *Genes* 14.3 (7th Mar. 2023), p. 664. DOI: 10.3390/genes14030664. URL: https://inria.hal.science/hal-04365469.

[8] A. Marino, B. Sinaimeri, E. Tronci and T. Calamoneri. 'STARGATE-X: a Python package for statistical analysis on the REACTOME network'. In: *Journal of Integrative Bioinformatics* (21st Sept. 2023). DOI: 10.1515/jib-2022-0029. URL: https://inria.hal.science/hal-04365656.

[9] H. Paraqindes, N.-E.-H. Mourksi, S. Ballesta, J. Hedjam, F. Bourdelais, T. Fenouil, T. Picart, F. Catez, T. Combe, A. Ferrari, J. Kielbassa, E. Thomas, L. Tonon, A. Viari, V. Attignon, M. Carrere, J. Perrossier, S. Giraud, C. Vanbelle, M. Gabut, D. Bergeron, M. Scott, L. Castro Vega, N. Magne, E. Huillard, M. Sanson, D. Meyronet, J.-J. Diaz, F. Ducray, V. Marcel and S. Durand. 'Isocitrate dehydrogenase wt and IDHmut adult-type diffuse gliomas display distinct alterations in ribosome biogenesis and 2'O-methylation of ribosomal RNA'. In: *Neuro-Oncology* (8th Dec. 2023). DOI: 10.1093/neuonc/noad140. URL: https://hal.science/hal-04203242.

[10] B. Sinaimeri, L. Urbini, M.-F. Sagot and C. Matias. 'Cophylogeny Reconstruction Allowing for Multiple Associations Through Approximate Bayesian Computation'. In: *Systematic Biology* (13th Sept. 2023), syad058. DOI: 10.1093/sysbio/syad058. URL: https://hal.science/hal-03673256.

[11] A. Villié, P. Veber, Y. de Castro and L. Jacob. 'Neural Networks beyond explainability: Selective inference for sequence motifs'. In: *Transactions on Machine Learning Research Journal* (4th July 2023). URL: https://hal.science/hal-03895446.

[12] Y. Wang, A. Mary, M.-F. Sagot and B. Sinaimeri. 'A General Framework for Enumerating Equivalence Classes of Solutions'. In: *Algorithmica* 85.10 (4th May 2023), pp. 3003–3023. DOI: 10.1007/s00453-023-01131-1. URL: https://inria.hal.science/hal-04365403.

International peer-reviewed conferences

[13] L. a. K. Ayad, G. Loukides and S. P. Pissis. 'Text Indexing for Long Patterns: Anchors are All you Need'. In: *Proceedings of the VLDB Endowment*. VLDB 2023 - 49th International Conference on Very Large Data Bases. Vol. 16. 9. Vancouver, Canada, May 2023, pp. 2117–2131. DOI: 10.14778/3598581.3598586. URL: https://inria.hal.science/hal-04385571.

[14] G. Bernardini, G. Fici, P. Gawrychowski and S. P. Pissis. 'Substring Complexity in Sublinear Space'. In: ISAAC 2023 - 34th International Symposium on Algorithms and Computation. Kyoto, Japan: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023. DOI: 10.4230/LIPIcs.ISAAC.2023.12. URL: https://inria.hal.science/hal-04385532.

[15] T. Bosman, M. van Ee, E. Ergen, C. Imreh, A. Marchetti-Spaccamela, M. Skutella and L. Stougie. 'Total Completion Time Scheduling Under Scenarios'. In: WAOA 2023 - International Workshop on Approximation and Online Algorithms. Vol. 14297. Lecture Notes in Computer Science. Amsterdam, Netherlands: Springer Nature Switzerland, 22nd Dec. 2023, pp. 104–118. DOI: 10.1007/978-3-031-49815-2_8. URL: https://inria.hal.science/hal-04385325.

[16] E. Gabory, M. N. Mwaniki, N. Pisanti, S. P. Pissis, J. Radoszewski, M. Sweering and W. Zuba. 'Comparing Elastic-Degenerate Strings: Algorithms, Lower Bounds, and Applications'. In: 34th Annual Symposium on Combinatorial Pattern Matching (CPM 2023). Marne-la-Vallée, France, 2023. DOI: 10.4230/LIPIcs.CPM.2023.11. URL: https://inria.hal.science/hal-04365687.

[17] G. Italiano, A. Konstantinidis and C. Papadopoulos. 'Structural Parameterization of Cluster Deletion'. In: WALCOM 2023 - International Conference and Workshops on Algorithms and Computation. Vol. 13973. Lecture Notes in Computer Science. Hsinchu, Taiwan: Springer Nature Switzerland, 13th Mar. 2023, pp. 371–383. DOI: 10.1007/978-3-031-27051-2_31. URL: https://inria.hal.science/hal-04385361.

[18] G. Loukides, S. P. Pissis, S. V. Thankachan and W. Zuba. 'Suffix-Prefix Queries on a Dictionary'. In: *Leibniz International Proceedings in Informatics (LIPIcs)*. CPM 2023 - 34th Annual Symposium on Combinatorial Pattern Matching. Vol. 259. Marne-la-Vallée, France: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023, 21:1–21:20. DOI: 10.4230/LIPIcs.CPM.2023.21. URL: https://inria.hal.science/hal-04385499.

Conferences without proceedings

[19] S. Peres. 'Hybrid modelling to Solve Optimal Concentrations of Metabolites and Enzymes in Constraint-based modelling'. In: BIOSTEC 2023. Lisbon (Portugal), Portugal, 16th Feb. 2023. URL: https://hal.science/hal-04036239.

Doctoral dissertations and habilitation theses

[20] N. Homberg. 'New models and algorithms for the identification of sncRNA-(snc)RNAinteractions intra and across-species/kingdoms'. Université Claude Bernard Lyon 1, 15th June 2023. URL: https://inria.hal.science/tel-04366914.