IN PARTNERSHIP WITH:
Centrum Wiskunde & Informatica

Institut national des sciences appliquées de Lyon

Université Claude Bernard (Lyon 1)

Université de Rome la Sapienza

Activity Report 2017

Project-Team ERABLE

European Research team in Algorithms and Biology, formaL and Experimental

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

RESEARCH CENTER
Grenoble - Rhône-Alpes

THEME
Computational Biology

Table of contents

Project-Team ERABLE

Creation of the Team: 2015 January 01, updated into Project-Team: 2015 July 01

ERABLE is a European Inria team gathering French researchers together with researchers in Italy under the banner of the Sapienza University of Rome and researchers in the Netherlands under the banner of the CWI.

Keywords:

Computer Science and Digital Science:

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 biology
B1.1.5. - Genetics
B1.1.6. - Genomics
B1.1.8. - Evolutionnary biology
B1.1.9. - Bioinformatics
B1.1.11. - Systems biology
B1.1.12. - Synthetic biology
B2. - Health
B2.2. - Physiology and diseases
B2.2.3. - Cancer
B2.2.4. - Infectious diseases, Virology
B2.3. - Epidemiology

1. Personnel

Research Scientists

Marie-France Sagot [Team leader, Inria, Senior Researcher, HDR]
Blerina Sinaimeri [Inria, Researcher]

Fabrice Vavre [CNRS, Senior Researcher]
Alain Viari [Inria, Senior Researcher]
Alexander Schönhuth [CWI, The Netherlands, Senior Researcher, from Jul 2017]

Faculty Members
Hubert Charles [INSA Lyon, Professor, HDR]
Vincent Lacroix [Univ Claude Bernard Lyon, Associate Professor]
Arnaud Mary [Univ Claude Bernard Lyon, Associate Professor]
Cristina Vieira [Univ Claude Bernard Lyon, Professor, HDR]
Roberto Grossi [Univ Pisa, Italy, Professor]
Giuseppe Francesco Italiano [Univ Tor Vergata, Rome, Italy, Professor, from Jul 2017]
Pierluigi Crescenzi [Univ Florence, Italy, Professor]
Alberto Marchetti Spaccamela [Sapienza Univ Rome, Italy, Professor]
Nadia Pisanti [Univ Pisa, Italy, Associate Professor]
Leen Stougie [CWI & Free Univ Amsterdam, The Netherlands, Professor]

Post-Doctoral Fellows
Ricardo de Andrade Abrantes [Univ of São Paulo, Brazil, & Inria, from Jun 2017]
Alex Di Genova [Inria, from Dec 2017]

PhD Students
Audric Cologne [Inria & INSERM]
Mattia Gastaldello [Sapienza Univ Rome & Univ Claude Bernard Lyon]
Carol Moraga Quinteros [Univ Claude Bernard Lyon]
Henri Pusa [Inria]
Laura Urbini [Inria, until Sep 2017]
Irene Ziska [Inria, from Oct 2017]
Camille Sessegolo [Univ Claude Bernard Lyon, from Sept 2017]
Leandro Ishi Soares de Lima [Univ Claude Bernard Lyon]

Technical staff
Clara Benoît-Pilven [INSERM, until Sep 2017]
Martin Wannagat [Inria]

Interns
Camille Sessegolo [Univ Claude Bernard Lyon, until Jul 2017]
Irene Ziska [Inria, until May 2017]

Administrative Assistants
Marina Da Graca [Inria, until Aug 2017]
Claire Sauer [Inria]

Visiting Scientist
Alex Di Genova [Univ Adolfo Ibañez, Chile, until Feb 2017]

External Collaborators
Laurent Jacob [LBBE UMR5558, Researcher, external collaborator]
Susana Vinga [IST Lisbon, Researcher, external collaborator]

2. Overall Objectives

2.1. 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 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 other living systems, or systems that have been described as being at the edge of life such as viruses, or else genetic or inorganic material such as, respectively, transposable elements and chemical compounds.

The application goal of ERABLE is, through the use of mathematical models and algorithms, to better understand such close and often persistent interactions, with a longer term objective 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 goal 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 goal 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 goals, one related to biology and the other to methodology (algorithms, combinatorics, statistics). In relation to biology, the main goal of ERABLE is to contribute, through the use of mathematical models and algorithms, to a better understanding of close and often persistent interactions between "collections of genetically identical or distinct self-replicating cells" which will correspond to organisms/species or to actual cells. The first will cover the case of what has been called symbiosis, meaning when the interaction involves different species, while the second will cover the case of a (cancerous) tumour which may be seen as a collection of cells which suddenly disrupts its interaction with the other (collections of) cells in an organism by starting to grow uncontrollably.

Such interactions are being explored initially at the molecular level. Although we rely as much as possible on already available data, we intend to also continue contributing to the identification and analysis of the main genomic and systemic (regulatory, metabolic, signalling) elements involved or impacted by an interaction, and how they are impacted. We started going to the populational and ecological levels by modelling and analysing the way such interactions influence, and are or can be influenced by the ecosystem of which the "collections of cells" are a part. The key steps are:

- identifying the molecular elements based on so-called omics data (genomics, transcriptomics, metabolomics, proteomics, etc.): such elements may be gene/proteins, genetic variations, (DNA/RNA/protein) binding sites, (small and long non coding) RNAs, etc.
- simultaneously inferring and analysing the network that models how these molecular elements are physically and functionally linked together for a given goal, or find themselves associated in a response to some change in the environment;
- modelling and analysing the populational and ecological network formed by the "collections of cells in interaction", meaning modelling a network of networks (previously inferred or as already available in the literature);
- analysing how the behaviour and dynamics of such a network of networks might be controlled by modifying it, including by subtracting some of its components from the network or by adding new ones.

In relation to methodology, the main goal is to provide those enabling to address our main biological objective as stated above that lead to the best possible interpretation of the results within a given pre-established model and a well defined question. Ideally, given such a model and question, the method is exact and also exhaustive if more than one answer is possible. Three aspects are thus involved here: establishing the model within which questions can and will be put; clearly defining such questions; exactly answering to them or providing some guarantee on the proximity of the answer given to the "correct" one. We intend to continue contributing to these three aspects:

- at the modelling level, by exploring better models that at a same time are richer in terms of the information they contain (as an example, in the case of metabolism, using hypergraphs as models for it instead of graphs) and are susceptible to an easier treatment:
  - these two objectives (rich models that are at the same time easy to treat) might in many cases be contradictory and our intention is then to contribute to a fuller characterisation of the frontiers between the two;
  - even when feasible, the richer models may lack a full formal characterisation (this is for instance the case of hypergraphs) and our intention is then to contribute to such a characterisation;
- at the question level, by providing clear formalisations of those that will be raised by our biological concerns;
- at the answer level:
  - to extend the area of application of exact algorithms by: (i) a better exploration of the combinatorial properties of the models, (ii) the development of more efficient data structures, (iii) a smarter traversal of the space of solutions when more than one solution exists;
  - when exact algorithms are not possible, or when there is uncertainty in the input data to an algorithm, to improve the quality of the results given by a deeper exploration of the links between different algorithmic approaches: combinatorial, randomised, stochastic.

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. Our choice is based more on the biological questions as these are a main (but not unique) driver for the methodological developments. However, since another main objective is to contribute to the fields of exact enumeration algorithms and of combinatorics, we also defined an axis that is exclusively oriented towards some of the more theoretical aspects of such objective in as much as these can be abstracted from the biological motivation. This will concern improving theory and deeply exploring the links between different algorithmic approaches: combinatorial, randomised, stochastic.

Initially, when ERABLE was created, five axes were defined. The first four fell in the first category above, and the fifth one in the second.

More recently however, as was indicated in the evaluation report for the period 2013-2017, a new biological axis was added to the four that existed already. This axis is specifically oriented towards health in general, human or animal. It was numbered as Axis 4 as the last biological one which existed already may be seen as a generalisation of the first four (three old ones and new fourth). Indeed, one overall objective of ERABLE for the next four years will be to try to establish the links between non infectious diseases such as cancer or rare ones on one hand, and infectious diseases (related to symbiosis understood in its more general sense) on the other.

As concerns symbiosis, the model organisms or systems chosen include the following cases:

- Arthropods, notably insects, and their parasites;
- Symbiont-harbouring trypanosomatids and trypanosomas more in general;
- The bacterial communities inside the respiratory tract of mammals (swine, bovine);
- Human in general, and the human microbiota in particular also for its possible relation to cancer.

Notice however that: (1) new model organisms or systems may be considered as the opportunity for new collaborations appears, indeed such collaborations will be actively searched for; and (2) we will always attempt to explore mathematical and computational models and to develop algorithmic methods that are as much as possible generic.

Axis 1: Identifying the molecular elements
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: Inferring and analysing the networks of molecular elements

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 genetic, metabolic, protein-protein interaction and signalling networks. This raises two main classes of problems. The first is to accurately infer such networks. Reconstructing, by analogy, the metabolic network of an organism is often considered, rightly or wrongly, to be easier than inferring a gene regulatory network, also because in the latter case, identifying all the elements participating in the network is in itself a complex and far from solved issue, as we saw in Axis 1. Moreover, the difficulty varies depending on whether only the structure or also the dynamics of the network is of interest, assuming that the latter may be studied (kinetics data are often missing even with the increasingly more sophisticated and performing technologies we have nowadays). A more complete picture of the functioning of a cell would further require that ever more layers of network and molecular profile data, when available, are integrated together, which raises the problem of how to model together information that is heterogeneous at different levels. Modelling together metabolic and gene regulation for instance is already a hard problem given that the two happen at very different time-scales: fast for metabolic regulation, slow for gene regulation.

Even 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 difficulty of this differs of course again depending on whether only the structure of the network is of interest, or also its dynamics. We are addressing various questions related to one or the other of the above aspects – inference and analysis.

Axis 3: Modelling and analysing a network of individuals, or a network of individuals' networks

As mentioned, at its extreme, life can be seen 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 a same or with distinct functional objectives. One striking example is human, who is composed of cells which are both native and extraneous; in fact, a surprising 90% is believed to belong to the second category, mostly bacteria, including one which lost its identity to become a "mere" human organelle, the mitochondrion. Bacteria on the other hand group into colonies of genetically identical individuals which may sometimes acquire the ability to become specialised for different tasks. Which is the "individual", a single bacterium or a group thereof is difficult to say. To understand human or bacteria, or to understand any other organism, it appears therefore essential to better comprehend the interactions in which they are involved. Methodologically speaking, we must therefore move towards modelling and analysing not a single individual anymore but a network of individuals. Ultimately, we should move towards investigating a network of individuals' networks. Moreover, since organisms interact not only with others but also with their abiotic environment, there is a need to model full ecosystems, at a static but also at a dynamic level, that is by taking into account the fact that individuals or populations move in space. Our intention at a longer term is to address all such different levels. We started with the molecular and static one that we are treating from different perspectives for a large number of species at the genomic level (Baudet *et al.*, Syst Biol, 64(3), 2015) and for a small number at the network level (Cottret *et al.*, PLoS Comput Biol, 6(9), 2010). We intend in a near future to slowly move towards a populational and ecological approach that is dynamic in both time and space.

Axis 4: Human and animal health

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.

Infectiology is the oldest one and is covered also by Axis 5 below. It will thus be described there. It started by a collaboration with Arnaldo Zaha from the Federal University of Rio Grande do Sul in Brazil. 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 the Taybi-Linder Syndrome (TALS) and on abnormal splicing of U12 introns, while Cancer rests on a collaboration with the Centre Léon Bérard (CLB) and Centre de Recherche en Cancérologie of Lyon (CRCL) which is focused on Breast and Prostate carcinomas and Gynaecological carcinosarcomas.

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

Within Inria and beyond, the first two applications (Infectiology and Rare Diseases) may be seen as unique because of their specific focus (resp. respiratory tract of swines and TALS). In the first case, such uniqueness is also related to the fact that the work done involves a strong computational part but also experiments *performed within ERABLE itself*.

Axis 5: Going towards control

What was described in the Axes 2 and 3 above concerned modelling and analysing a molecular network, or network of networks, but not attempting to control the network at either level for bio-technological, environmental or health purposes.

In the bio-technological case, the objective can be briefly described as involving the manipulation of a species, in general a bacterium, in order for it to produce more of a given chemical compound it already synthetises (for instance, ethanol) but not in enough quantity, or to produce a metabolite it normally is not able to synthetise. The motivation for transplanting its production in a bacterium is, again, to be able to make it more effective.

As concerns control for environmental or health purposes, this could be achieved at least in some cases by manipulating the symbionts with which an organism, insect pest for instance, or humans leave. In the environmental case, this has gone under the name of "biological control" (see for instance Flint & Dreistat, "Natural Enemies Handbook: The Illustrated Guide to Biological Pest Control", University of California Press, 1998) and involves the use of "natural enemies" of a pest organism. This idea has a long history: the ancient Chinese, observing that ants were effective predators of many citrus pests, decided to increase the ants population by displacing their nests from the surrounding habitats and placing them inside their orchards to protect them. More recently, there has been growing evidence that some endosymbiotic bacteria, that is bacteria that live within the cells of their hosts, could become efficient biocontrol agents. This is in particular the case of *Wolbachia*, a bacterium much studied in ERABLE (Ahantarig & Kittayapong, J Appl Entomology, 135(7):479-486, 2011).

The connection between disease and the disruption of homeostatic interactions between the host and its microbiota is on the other hand now well established. Microbiota-targeted therapies involve altering the community composition by eliminating individual strains of a single species (for example, with antibiotics) or replacing the entire community with a new intact microbiota. Secondary infections linked to antibiotic use provide however a cautionary tale of the possible consequences of perturbing a microbial species network.

Besides the biotechnological aspects on which we are already working in the context of two European projects (BacHBerry, and to a lesser extent, MicroWine), our main goal in this case is to try to formalise such type of control. There are two objectives here. One is methodological and concerns attempting to provide a single formal framework for the diverse ways of controlling a network, or a network of networks. Our attention has concentrated initially on metabolism, and will at a mid to longer term include regulation. Our intention notably as concerns the incorporation of regulation is to collaborate with other Inria teams, most notably IBIS with whom we are already in discussion. The second objective is biological and concerns control for environmental and health purposes. The originality we are seeking in this case is to attempt such control not by eliminating species, which is done mainly through the use of antibiotics that may then create resistance, a phenomenon that is becoming a major clinical and public health problem, but by manipulating the species or their environment, or by changing the composition of the community by adding or displacing some other species in such a way that new equilibria may be reached which enable all the species living in a same niche to survive. The idea is not new: the areas of prebiotics (non-digestible food ingredients that stimulate the growth and/or activity of bacteria in the digestive system in beneficial ways) and probiotics (micro-organisms claimed to provide benefits when consumed) indeed cover similar concerns in relation to health. Other novel approaches propose to work at the level of bacterial communication (quorum sensing) to control for pathogenicity (Rutherford & Bassler, Cold Spring Harbor Perspectives in Medicine, 2012). Small RNAs in particular are believed to play an important role in quorum sensing.

Axis 6: Cross-fertilising different computational approaches
In computer science and in optimisation, different approaches and techniques have been proposed to cope with hardness results. It is clear that none of them is dominant: there are classes of problems for which approach A is better than approach B, and vice-versa. Moreover, there is no satisfactory understanding of the conditions that favour one approach with respect to another one.

As an example, the team that gave birth to ERABLE, BAMBOO, had expertise more in the area of combinatorial algorithms for strings (sequences), trees and graphs. Many such algorithms addressed an enumeration problem: given a certain description of the object(s) searched for or definition of a function to be optimised, the method was supposed to list all the solutions. In many real life situations, notably in biology, a majority of the problems treated, of whatever kind, enumeration or else, are however hard. Although combinatorics remains crucial to better understand the structure of such problems and delimit the conditions that could render them easy or at least tractable in practice, often other types of approaches have to be attempted.

Although all approaches may be valid and valuable, in many cases one only is explored. More in general, there appears to be relatively little cross-talk and cross-fertilisation being attempted between these different approaches. Guided by problems from computational biology, the goal of this axis is to add to the growing insights on how well such problems can be solved theoretically.

4. Application Domains

4.1. Biology & 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, and cancer.

5. New Software and Platforms

5.1. 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.

- Participants: Alain Viari, Anne Morgat, Frédéric Boyer, Marie-France Sagot and Yves-Pol Deniélou
- Contact: Alain Viari
- URL: http://www.inrialpes.fr/helix/people/viari/lxgraph/index.html

5.2. Cassis

KEYWORDS: Bioinformatics - Genomics
FUNCTIONAL DESCRIPTION: Implements methods for the precise detection of genomic rearrangement breakpoints.

- Participants: Christian Baudet, Christian Gautier, Claire Lemaitre, Eric Tannier and Marie-France Sagot
- Contact: Marie-France Sagot
- URL: http://pbil.univ-lyon1.fr/software/Cassis/

5.3. Coala

CO-evolution Assessment by a Likelihood-free Approach
KEYWORDS: Bioinformatics - Evolution
SCIENTIFIC 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 nd a reconstruction of minimum cost. The main problem with this approach is that the cost of the events strongly inuence the reconciliation obtained.

To deal with this problem, we developed an algorithm, called Coala, for estimating the frequency of the events based on an approximate Bayesian computation approach.
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.

- Participants: Beatrice Donati, Blerina Sinaimeri, Catherine Matias, Christian Baudet, Christian Gautier, Marie-France Sagot and Pierluigi Crescenzi
- Contact: Blerina Sinaimeri
- URL: http://coala.gforge.inria.fr/

5.4. 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.

- Contact: Nadia Pisanti
- URL: https://github.com/solonas13/csc

5.5. 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...).

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

5.6. 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.

- Participants: Beatrice Donati, Blerina Sinaimeri, Christian Baudet, Marie-France Sagot and Pierluigi Crescenzi
- Contact: Blerina Sinaimeri
- URL: http://eucalypt.gforge.inria.fr/

5.7. 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.

- Contact: Alex Di Genova
- URL: https://github.com/adigenova/fast-sg

5.8. 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.

- Participants: Etienne Birmelé, Fabien Jourdan, Ludovic Cottret, Marie-France Sagot, Paulo Vieira Milreu, Pierluigi Crescenzi, Vicente Acuna Aguayo and Vincent Lacroix
- Contact: Marie-France Sagot
- URL: http://gforge.inria.fr/projects/gobbolino

5.9. 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.

- Contact: Nadia Pisanti
- URL: http://hapcol.algolab.eu/

5.10. HgLib

KEYWORD: Graph algorithmics
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.

- Contact: Arnaud Mary
- URL: https://gitlab.inria.fr/kirikomics/hglib

5.11. 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 FUNCTIONAL DESCRIPTION: This new version improved the recall and made more precise the size of the effect computation.

- Participants: Camille Marchet, Aurélie Siberchicot, Audric Cologne, Clara Benoît-Pilven, Janice Kielbassa, Lilia Brinza and Vincent Lacroix
- Contact: Vincent Lacroix
- URL: http://kissplice.prabi.fr/tools/kissDE/

5.12. KisSplice

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

RELEASE FUNCTIONAL DESCRIPTION: Improvements : 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+.

- Participants: Alice Julien-Laferrière, Leandro Ishi Soares De Lima, Vincent Miele, Rayan Chikhi, Pierre Peterlongo, Camille Marchet, Gustavo Akio Tominaga Sacomoto, Marie-France Sagot and Vincent Lacroix
- Contact: Vincent Lacroix
- URL: http://kissplice.prabi.fr/

5.13. 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.

- Participants: Audric Cologne, Camille Marchet, Camille Sessegolo, Alice Julien-Laferrière and Vincent Lacroix
- Contact: Vincent Lacroix
- URL: http://kissplice.prabi.fr/tools/kiss2refgenome/

5.14. 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.

- Participants: Helene Lopez Maestre, Mathilde Boutigny and Vincent Lacroix
- Contact: Vincent Lacroix
- URL: http://kissplice.prabi.fr/tools/kiss2rt/

5.15. MetExplore

KEYWORDS: Systems Biology - Bioinformatics
SCIENTIFIC DESCRIPTION: MetExplore stores metabolic networks of 160 organisms into a relational database. Information about metabolic networks mainly come from BioCyc-like databases. Two BioCyc-like databases contain information about several organisms: PlantCyc and MetaCyc. MetExplore contains also the information about metabolites stored in Metabolome.jp. Note that there is no information about reactions in this database and is only useful to identify compounds from masses. Several genome-scale models designed for Flux Balance Analysis have also been imported into MetExplore. The table below gives details about the sources of the metabolic networks present in MetExplore.
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.

- Participants: Fabien Jourdan, Hubert Charles, Ludovic Cottret and Marie-France Sagot
- Contact: Fabien Jourdan
- URL: http://metexplore.toulouse.inra.fr/metexplore/

5.16. Mirinho

KEYWORDS: Bioinformatics - Computational biology - Genomics - Structural Biology
FUNCTIONAL DESCRIPTION: Predicts, at a genome-wide scale, microRNA candidates.

- Participants: Christian Gautier, Christine Gaspin, Cyril Fournier, Marie-France Sagot and Susan Higashi
- Contact: Marie-France Sagot
- URL: http://mirinho.gforge.inria.fr/

5.17. MultiPus

KEYWORDS: Systems Biology - Algorithm - Graph algorithmics - Metabolic networks - Computational biology
SCIENTIFIC DESCRIPTION: Synthetic biology has boomed since the early 2000s when it started being shown that it was possible to efficiently synthetise compounds of interest in a much more rapid and effective way by using other organisms than those naturally producing them. However, to thus engineer a single organism, often a microbe, to optimise one or a collection of metabolic tasks may lead to difficulties when attempting to obtain a production system that is efficient, or to avoid toxic effects for the recruited microorganism. The idea of using instead a microbial consortium has thus started being developed in the last decade. This was motivated by the fact that such consortia may perform more complicated functions than could single populations and be more robust to environmental fluctuations. Success is however not always guaranteed. In particular, establishing which consortium is best for the production of a given compound or set thereof remains a great challenge. The algorithm MultiPus is based on an initial model that enables to propose a consortium to synthetically produce compounds that are either exogenous to it, or are endogenous but where interaction among the species in the consortium could improve the production line.
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.

- Participants: Alberto Marchetti-Spaccamela, Alice Julien-Laferrière, Arnaud Mary, Delphine Parrot, Laurent Bulteau, Leen Stougie, Marie-France Sagot and Susana Vinga
- Contact: Marie-France Sagot
- URL: http://multipus.gforge.inria.fr/

5.18. 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.

- Contact: Marie-France Sagot
- URL: http://gforge.inria.fr/projects/pitufo/

5.19. Sasita

KEYWORDS: Bioinformatics - Graph algorithmics - Systems Biology
FUNCTIONAL DESCRIPTION:SASITA is a software for the exhaustive enumeration of minimal precursor sets in metabolic networks.

- Contact: Marie-France Sagot
- URL: http://sasita.gforge.inria.fr/

5.20. Savage

KEYWORDS: Algorithm - Genomics
FUNCTIONAL DESCRIPTION: Reconstruction of viral quasi species without using a reference genome.

- Contact: Alexander Schonhuth
- URL: https://bitbucket.org/jbaaijens/savage

5.21. Smile

KEYWORDS: Bioinformatics - Genomic sequence
FUNCTIONAL DESCRIPTION: Motif inference algorithm taking as input a set of biological sequences.

- Participant: Marie-France Sagot
- Contact: Marie-France Sagot

5.22. Rime

KEYWORDS: Bioinformatics - Genomics - Sequence alignment
FUNCTIONAL DESCRIPTION: Detects long similar fragments occurring at least twice in a set of biological sequences.

- Contact: Nadia Pisanti

5.23. Totoro & Kotoura

KEYWORDS: Bioinformatics - Graph algorithmics - Systems Biology
FUNCTIONAL DESCRIPTION: Both TOTORO and KOTOURA decipher the reaction changes during a metabolic transient state, using measurements of metabolic concentrations. These are called metabolic hyperstories. TOTORO (for TOpological analysis of Transient metabOlic RespOnse) is based on a qualitative measurement of the concentrations in two steady-states to infer the reaction changes that lead to the observed differences in metabolite pools in both conditions. In the currently available release, a pre-processing and a post-processing steps are included. After the post-processing step, the solutions can be visualised using DINGHY (http://dinghy.gforge.inria.fr). KOTOURA (for Kantitative analysis Of Transient metabOlic and regUlatory Response And control) infers quantitative changes of the reactions using information on measurement of the metabolite concentrations in two steady-states.

- Contact: Marie-France Sagot
- URL: http://hyperstories.gforge.inria.fr/

5.24. 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.

- Contact: Nadia Pisanti
- URL: https://bitbucket.org/whatshap/whatshap

6. New Results

6.1. General comments

We present in this section the main results obtained in 2017.

We tried to organise these following the six main axes of research of the team. 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 did not indicate here the results on more theoretical aspects of computer science if it did not seem for now that they could be relevant in contexts related to computational biology. Actually, we do believe those on scheduling [25], [24], and on text [37], graph [4], [32], [34], [5], [36], [35] or general algorithmic problems notably related to performance issues [23], [28] could in the future become more specifically relevant for life sciences (biology or ecology). We did not indicate either work that was done a few years ago by members who were in ERABLE but whose associated publication appeared only this year [27].

Notice that the theoretical results related to problems closely resembling questions that have already been addressed by us in computational biology and that we present below concern not only cross-fertilising issues among different computational approaches, and we therefore extended the title of this axis for the purpose of presenting such results, for now purely theoretical.

A few other results of 2017 are not mentioned in this report, not because the corresponding work is not important, but because it was likewise more specialised, or the work represented a survey.

6.2. Identifying the molecular elements

Motif tries for pattern discovery. In [14], the motif trie data structure was introduced to improve the extraction of recurring patterns in sequences. Such extraction concerned maximal patterns with at most k don't care symbols and at least q occurrences, according to a given maximality notion. The motif trie was applied to this problem, also showing how to build it efficiently. This led to the first algorithm that attains a stronger notion of output-sensitivity, where the cost for an input sequence of n symbols is proportional to the actual number of occurrences of each pattern, which is at most n (much smaller in practice). This avoids the best-known cost of $O(nc)O(nc)$ per pattern, for a constant $c > 1$, which is otherwise impractical for massive sequences with a large value of n.

Identification of genome and alternative splicing variants in RNA-seq data. The team's work on identifying alternative splicing and other genome variants such as SNPs (Single Nucleotide Polymorphism), indels, etc., started around 2010. This has concerned mostly RNA-seq data also for the variants investigated.

Both DNA and RNA-seq data analysis using so-called NGS (Next Generation Sequencing) is a domain of research that has been active for decades now, with many open questions remaining despite such long and intense activity. One is the case of non-model organisms, but actually there is another major problem that has not been solved, at least in any really satisfying way since the premises of genome sequencing. This is the problem of repeats. Notice however that repeats are not just "problems to be avoided", but have a strong biological interest in themselves, notably those related to transposable elements. Various papers of the team in 2017, notably [13], [19], [22], [1], were concerned with the study of such elements.

As concerns non-model organisms, the team extended a method it had previously developed, called KISSPLICE, to identify, quantify and annotate SNPs without any reference genome, using RNA-seq data only. The paper (Lopez-Maestre *et al.*, *Nucleic Acids Research*, 44(19):e148, 2016) appeared at the end of 2016. There we showed that individuals can be pooled prior to sequencing if not enough material is available from one individual. Using pooled human RNA-seq data, we clarified the precision and the recall of our method and discussed them with respect to others which use a reference genome or an assembled transcriptome. We then validated experimentally the predictions of our method using RNA-seq data from two non-model species. The method can be used for any species to annotate SNPs and predict their impact on the protein sequence. It enables to test for the association of the identified SNPs with a phenotype of interest. One of the phenotypes explored was related to the dependence of the insect *Asobara tabida* on its endosymbiont *Wolbachia*.

The methodological part of the work above relied in part on a number of more theoretical results, related to algorithmics and more specifically focused on the problem of repeats [21]. The most theoretical recent work of the team, accepted at the 43rd International Workshop on Graph-Theoretic Concepts in Computer Science (WG) in 2017 [30], proposed the notion of a bubble generator set, *i.e.* of a polynomial-sized subset of bubbles from which all the others can be obtained, also in polynomial time, through the application of a specific symmetric difference operator. This is further described in the last axis (Axis 6).

Genome and haplotype assembly. Fully assembling the genome sequence of an organism remains an important and challenging task. Genome scaffolding (*i.e.* the process of ordering and orientating contigs) of *de novo* assemblies usually represents the first step in most genome finishing pipelines. The team started by developing an algorithm (called MEDUSA) for such task (Bosi *et al.*, *Bioinformatics*, 31(15):2443-2451, 2015). It exploited information obtained from a set of (draft or closed) genomes from related organisms to determine the correct order and orientation of the contigs. It formalised the scaffolding problem by means of a combinatorial optimisation formulation on graphs and implements an efficient constant factor approximation algorithm to solve it. In contrast to the majority of the scaffolders, it did not require either prior knowledge on the input dataset (usually of micro-organisms) or the availability of paired-end read libraries. MEDUSA however presented limitations both in the construction of the scaffolding graph for large genomes, and in the subsequent assembly. The first aspect has been recently greatly improved by a method developed in collaboration with researchers (among which Alex di Genova) from Chile. This work led to the software FAST-SG already publicly available, and to a first publication that is in revision.

6.3. Inferring and analysing the networks of molecular elements

Metabolic impact of a change of conditions. The increasing availability of metabolomics data enables to better understand the metabolic processes involved in the immediate response of an organism to environmental changes, where the latter can be related to the presence of other species. The data usually come in the form of a list of metabolites whose concentrations significantly changed under some conditions, and are thus not easy to interpret without being able to precisely infer how such metabolites are interconnected. The team introduced a method that enables to organise the data from any metabolomics experiment into what we initially called *metabolic stories* when we were working with a simpler, graph representation of metabolism, and which have now become *metabolic hyperstories* as more accurate directed hypergraphs representations are considered. Each (hyper)story corresponds to a possible scenario explaining the flow of matter between the metabolites of interest. The initial work on a graph representation led to the GOBBOLINO + TOUCHÉ software (Milre *et al.*, *Bioinformatics*, 30(1):61-70, 2014). Two newer works working with directed hypergraphs were presented in the PhD of Alice Julien-Laferrière (defended in 2016). Two papers are currently in preparation. They led to the software TOTORO (which uses a qualitative measurement of concentrations in two steady-states) and KOTOURA (which infers quantitative changes of the reactions) which are both already publicly available.

Metabolic network reconstruction and comparison for understanding virulence. The respiratory tract of swines is colonised by several bacteria among which are three Mycoplasma species: *Mycoplasma flocculare*, *Mycoplasma hyopneumoniae* and *Mycoplasma hyorhinis*. While colonisation by *M. flocculare* is virtually asymptomatic, *M. hyopneumoniae* is the causative agent of enzootic pneumonia and *M. hyorhinis* is present in cases of pneumonia, polyserositis and arthritis. The genomic resemblance among these three Mycoplasma species combined with their different levels of pathogenicity is an indication that they have unknown mechanisms of virulence and differential expression, as for most mycoplasmas. We performed whole-genome metabolic network reconstructions for the three mycoplasmas, as well as cultivation tests and metabolomic experiments through nuclear magnetic resonance spectroscopy (NMR) (Ferrarini *et al.*, *BMC Genomics*, 17(1):353, 2016). We were able to infer from such reconstructed networks that the lack of pathogenicity of *M. flocculare* if compared to the highly pathogenic *M. hyopneumoniae* may be related to its incapacity to produce cytotoxic hydrogen peroxide. A second, more experimentally oriented-paper is currently under revision.

6.4. Modelling and analysing a network of individuals, or a network of individuals' networks

On unrooted and root-uncertain variants of several well-known phylogenetic network problems Genetic hybridisation is the process individuals from genetically distinct populations that are able to interbreed and this produce a hybrid.

The hybridisation number problem refers to finding the minimum number of hybridisation events necessary to explain conflicts among several evolutionary trees. It requires to embed a set of binary rooted phylogenetic trees into a binary rooted phylogenetic network such that the number of nodes with in-degree two is minimised. However, from a biological point of view accurately inferring the root location in a phylogenetic tree is notoriously difficult and poor root placement can artificially inflate the hybridisation number. To this end, a number of relaxed variants of this problem were studied in [29]. We started by showing that the fundamental problem of determining whether an unrooted phylogenetic network displays (*i.e.* embeds) an unrooted phylogenetic tree, is NP-hard. On the positive side, we showed that this problem is FPT in reticulation number. In the rooted case, the corresponding FPT result is trivial, but here a more subtle argumentation was required. Next, we showed that the hybridisation number problem for unrooted networks (when given two unrooted trees) is equivalent to the problem of computing the tree bisection and reconnect distance of the two unrooted trees. We then considered the "root uncertain" variant of the hybridisation number. Here we are free to choose the root location in each of a set of unrooted input trees such that the hybridisation number of the resulting rooted trees is minimised. On the negative side, we showed that this problem is APX-hard. On the positive side, we showed that it is FPT in the hybridisation number, via kernelisation, for any number of input trees.

Phylogenetic tree reconciliation. Phylogenetic tree reconciliation consists in a mapping of one tree (usually the symbiont tree) to the other (the host tree) using event-based maximum parsimony. Given a cost model for the events, many optimal reconciliations are however possible. Any further biological interpretation of them must therefore take this into account, making the capacity to enumerate all optimal solutions a crucial point. Indeed, the problem is not just that if we proposed a single solution, there is a good chance we would miss the "true" answer, but also that we would lose the capacity to verify whether there exist some characteristics that are common to enough of the solutions to increase our confidence in the "story" such reconciliation tells of the past.

When the ERABLE team started addressing this issue, only two algorithms existed that attempted such enumeration; in one case (software CoRe-Pa) not all possible solutions were produced while in the other (software Notung) not all cost vectors were handled. We then introduced a polynomial-delay algorithm, called Eucalypt, for enumerating all optimal reconciliations, and showed that in general many solutions exist (Donati *et al.*, *Algorithms for Molecular Biology*, 10(1):11, 2015). Some might not be time-feasible. However, we further showed that, among the many solutions that are usually found, in the majority of the cases, at least some will be time-feasible, and we provided a polynomial algorithm to test for time-feasibility.

We also considered a restricted version of the model where host switches are allowed to happen only between species that are within some fixed distance along the host tree. This restriction allows to reduce the number of time-feasible solutions while preserving the same optimal cost, as well as to find time-feasible solutions with a cost close to the optimal in the cases where no time-feasible optimal solution is found.

More recently, we defined two equivalence relations that enable to identify many reconciliations with a single one, thereby reducing their number. These results were published in a paper which was accepted at CIBB 2017 and will appear in the *LNCI-LNCS* proceedings of the conference (published after CIBB). Extensive experiments indicated that the number of output solutions greatly decreases in general. By how much clearly depends on the constraints that are given as input. An extended journal version of this work that includes its theoretical part will be submitted at the beginning of 2018. Other forms of grouping (or clustering) solutions are also being explored that rely instead on defining a distance between two different reconciliations. Two approaches are being investigated, one in collaboration with a researcher in Italy (paper in preparation), and the other with researchers in the UK (one paper submitted and one in preparation).

Improving the biological realism of coevolutionary models. The host-symbiont coevolutionary models developed so far needed also to be improved. The realism we wished to add to such models was for now the possibility to handle the case of multiple associations of a symbiont. Among the few previous works that allowed for this, all presented some limitation either in terms of the model or of the algorithm developed. Handling such multiple associations requires to introduce an event that was little or not formally considered in the literature. This is the event of *spread*, which precisely corresponds to the invasion of different hosts by a same symbiont. In this case, as when spreads are not considered, the optimal reconciliations obtained will depend on the choice made for the costs of the events. The need to develop statistical methods to assign the most appropriate ones therefore remained also of actuality. This is one of the problems we addressed in the PhD of Laura Urbini that was defended in October 2017. Two types of spread were in fact introduced: vertical and horizontal. The first corresponds to the case where the evolution of the symbiont "freezes" while the symbiont continues to be associated with a host and with the new species that descend from this host. The second includes both an invasion, of the symbiont which remains with the initial host but at the same time gets associated with ("invades") another one incomparable with the first, and a double freeze (in relation to the evolution of the host with which it was initially associated and in relation to the evolution of the second one it "invaded"). Two papers addressing distinct aspects related to the spread problem with different approaches are in preparation and will be submitted before the end of 2017 or beginning of 2018.

Estimating the frequency and expansion process of an infection We addressed the question of how often an infection occurs and of whether its expansion reached an equilibrium using as model *Wolbachia*. *Wolbachia* is a bacterial genus that infects about half of all arthropods, with diverse and extreme consequences ranging from sex-ratio distortion and mating incompatibilities to protection against viruses. These phenotypic effects, combined with efficient vertical transmission from mothers to offspring, satisfactorily explain the invasion dynamics of *Wolbachia* within species. However, beyond the species level, the lack of congruence between the host and symbiont phylogenetic trees indicates that *Wolbachia* horizontal transfers and extinctions do happen and underlie its global distribution.

In [3], we inferred recent acquisition/loss events from the distribution of *Wolbachia* lineages across the mitochondrial DNA tree of 3600 arthropod specimens, spanning 1100 species from Tahiti and the surrounding islands. We showed that most events occurred within the last million years, but are likely attributable to individual level variation (*e.g.*, imperfect maternal transmission) rather than to population level variation (*e.g.*, *Wolbachia* extinction). At the population level, we estimated that mitochondria typically accumulate 4.7% substitutions per site during an infected episode, and 7.1% substitutions per site during the uninfected phase. Using a Bayesian time calibration of the mitochondrial tree, these numbers translate into infected and uninfected phases of approximately 7 and 9 million years. Infected species thus lose *Wolbachia* slightly more often than uninfected species acquire it, supporting the view that its present incidence, estimated here slightly below 0.5, represents an epidemiological equilibrium.

6.5. Going towards control

Quantitative synthetic biology. Synthetic biology has boomed since the early 2000s when it started being shown that it was possible to efficiently synthetise compounds of interest in a much more rapid and effective way by using other organisms than those naturally producing them. However, to thus engineer a single organism, often a microbe, to optimise one or a collection of metabolic tasks may lead to difficulties when attempting to obtain a production system that is efficient, or to avoid toxic effects for the recruited microorganism. The idea of using instead a microbial consortium has thus started being developed in the last decade. Establishing which consortium is best for the production of a given compound or set thereof remains however a great challenge. The team introduced an initial model and a method, called MULTIPUS, that enable to propose a consortium to synthetically produce compounds that are either exogenous to it, or are endogenous but where interaction among the species in the consortium could improve the production line (Julien-Laferrière *et al.*, *Scientific Reports*, 6, 2016).

Since the work on MULTIPUS, the team has been considering quantitative approaches for synthetic biology. We thus explored the concept of multi-objective optimisation in the field of metabolic engineering when both continuous and integer decision variables are involved in the model. In particular, we proposed multi-objective models, initially for a single species, to suggest reaction deletion strategies, and also to deal with situations where several functions must be optimised simultaneously, such as the maximisation of bioproducts while minimising toxicity (Hartmann *et al.*, *BMC Systems Biology*, see https://www.ncbi.nlm.nih.gov/pubmed/29268790, just accepted and not yet visible in Hal-Inria). We compared our results with those obtained by using the well-known bi-level optimisation model of OPTKNOCK, and studied two multi-objective optimisation problems arising from the metabolic engineering of microorganisms. One of them, using Yeast, has been validated experimentally. The work is submitted. The team has then started expanding it to communities (Master Thesis of Irene Ziska who is continuing into a PhD).

6.6. Health

Rare Diseases. Splicing is an essential step in the process leading to gene expression because it not only removes the introns from the primary transcripts, but also generates a combination of mature transcripts through the differential inclusion/exclusion of exons and sometimes retention of introns. Some pathologies are associated to such abnormal splicing. This is the case of the Taybi-Linder Syndrome (TALS), a very rare malformative syndrome with autosomal recessive transmission, belonging to the group of microcephalic dwarfism and responsible for death usually before the age of 2 years. This pathology was recently found to be caused by mutations in RNU4ATAC, a small nuclear RNA, which is an essential component of the minor spliceosome. We started a collaboration with the group of Pr. P. Edery (who first identified this alteration in 2012) with the objective to establish a comprehensive catalog of splice alterations in several cohorts of TALS patients. In this work, we take advantage of our reference-free assembly approach of transcripts (KISSPLICE) in order to detect new splicing alterations and to identify the associated deregulated signalling genes and pathways.

Cancer. Alain Viari has continued to develop a strong interaction with clinicians concerned with cancer, notably of the breast and in the early human embryo. A number of papers have appeared in 2017 that describe this work [7], [6], [10], [11], [12], [16], [17], [18]. We highlight here just two.

The first [7] refers to breast cancer. Mismatch repair (MMR)-deficient cancers have been discovered to be highly responsive to immune therapies such as PD-1 checkpoint blockade, making their definition in patients, where they may be relatively rare, paramount for treatment decisions. In the study published in [7], we utilised patterns of mutagenesis known as mutational signatures, which are imprints of the mutagenic processes associated with MMR deficiency, to identify MMR-deficient breast tumours from a whole-genome sequencing dataset comprising a cohort of 640 patients. We identified 11 of 640 tumours as MMR deficient, but only 2 of 11 exhibited germline mutations in MMR genes or Lynch Syndrome. Two additional tumours had a substantially reduced proportion of mutations attributed to MMR deficiency, where the predominant mutational signatures were related to APOBEC enzymatic activity. Overall, 6 of 11 of the MMR-deficient

cases in this cohort were confirmed genetically or epigenetically as having abrogation of MMR genes. However, IHC analysis of MMR-related proteins revealed all but one of 10 samples available for testing as MMR deficient. Thus, the mutational signatures more faithfully reported MMR deficiency than sequencing of MMR genes, because they represent a direct pathophysiologic readout of repair pathway abnormalities. As whole-genome sequencing continues to become more affordable, it could be used to expose individually abnormal tumours in tissue types where MMR deficiency has been rarely detected, but also rarely sought.

The second [18] concerns early human embryo. Somatic cells acquire mutations throughout the course of an individual's life. Mutations occurring early in embryogenesis are often present in a substantial proportion of, but not all, cells in postnatal humans and thus have particular characteristics and effects. Depending on their location in the genome and the proportion of cells they are present in, these mosaic mutations can cause a wide range of genetic disease syndromes and predispose carriers to cancer. They have a high chance of being transmitted to offspring as de novo germline mutations and, in principle, can provide insights into early human embryonic cell lineages and their contributions to adult tissues. Although it is known that gross chromosomal abnormalities are remarkably common in early human embryos, our understanding of early embryonic somatic mutations is very limited. In this work, whole-genome sequences of normal blood from 241 adults was used to identify 163 early embryonic mutations. It was estimated that approximately three base substitution mutations occur per cell per cell-doubling event in early human embryogenesis and these are mainly attributable to two known mutational signatures. The mutations were then used to reconstruct developmental lineages of adult cells and demonstrate that the two daughter cells of many early embryonic cell-doubling events contribute asymmetrically to adult blood at an approximately 2:1 ratio. This study provided insights into the mutation rates, mutational processes and developmental outcomes of cell dynamics that operate during early human embryogenesis.

6.7. Cross-fertilising different computational approaches and other theoretical results

Bubble generator.
As mentioned earlier, a theoretical recent work of the team related to NGS analysis was accepted at the 43rd International Workshop on Graph-Theoretic Concepts in Computer Science (WG) in 2017 [30]. It introduced what was called a bubble generator.

Bubbles are pairs of internally vertex-disjoint (s, t)-paths with applications in the processing of DNA and RNA data. For example, enumerating alternative splicing events in a reference-free context can be done by enumerating all bubbles in a de Bruijn graph built from RNA-seq reads. However, listing and analysing all bubbles in a given graph is usually unfeasible in practice, due to the exponential number of bubbles present in real data graphs. In [30], we proposed a notion of a bubble generator set, *i.e.* a polynomial-sized subset of bubbles from which all the others can be obtained through the application of a specific symmetric difference operator. This set provides a compact representation of the bubble space of a graph, which can be useful in practice since some pertinent information about all the bubbles can be more conveniently extracted from this compact set. Furthermore, we provide a polynomial-time algorithm to decompose any bubble of a graph into the bubbles of such a generator in a tree-like fashion.

7. Bilateral Contracts and Grants with Industry

7.1. Bilateral Grants with Industry

ERABLE was awarded a PhD grant by the ANRt together with the Maat Pharma company. The PhD scholarship was granted to Marianne Borderes, who will be co-supervised starting from January 2018 by Marie-France Sagot and Susana Vinga (IST, Lisbon, Portugal) together with Lilia Boucinha from Maat Pharma.

8. Partnerships and Cooperations

8.1. National Initiatives

8.1.1. ANR

8.1.1.1. Aster

- Title: Algorithms and Software for Third gEneration Rna sequencing
- Coordinator: Hélène Touzet, University of Lille and Inria EPI Bonsai.
- ERABLE participants: Vincent Lacroix (ERABLE coordinator), Clara Benoît-Pilven, Audric Cologne, Alex di Genova, Leandro I. S. de Lima, Arnaud Mary, Marie-France Sagot, Camille Sessegolo, Blerina Sinaimeri.
- Type: ANR (2016-2020).
- Web page: http://bioinfo.cristal.univ-lille.fr/aster/.

8.1.1.2. ExHyb

- Title: Exploring genomic stability in hybrids
- Coordinator: C. Vieira
- ERABLE participant(s): C. Vieira
- Type: ANR (2014-2018)
- Web page: Not available

8.1.1.3. GraphEn

- Title: Enumération dans les graphes et les hypergraphes : Algorithmes et complexité
- Coordinator: D. Kratsch
- ERABLE participant(s): A. Mary
- Type: ANR (2015-2019)
- Web page: http://graphen.isima.fr/

8.1.1.4. Green

- Title: Deciphering host immune gene regulation and function to target symbiosis disturbance and endosymbiont control in insect pests
- Coordinator: A. Heddi
- ERABLE participant(s): M.-F. Sagot, C. Vieira
- Type: ANR (2018-2021)
- Web page: Not yet available

8.1.1.5. Hmicmac

- Title: Host-microbiota co-adaptations: mechanisms and consequences
- Coordinator: F. Vavre
- ERABLE participant(s): F. Vavre
- Type: ANR PRC (2017-2020)
- Web page: Not available

8.1.1.6. IMetSym

- Title: Immune and Metabolic Control in Intracellular Symbiosis of Insects
- Coordinator: A. Heddi
- ERABLE participant(s): H. Charles, S. Colella

- Type: ANR Blanc (2014-2017)
- Web page: Not available

8.1.1.7. Resist

- Title: Rapid Evolution of Symbiotic Interactions in response to STress: processes and mechanisms
- Coordinator: N. Kremer
- ERABLE participant(s): F. Vavre
- Type: ANR JCJC (2017-2020)
- Web page: Not available

8.1.1.8. Suzukill

- Title: Managing cold tolerance and quality of mass-produced *Drosophila suzukii* flies to facilitate the application of biocontrol through incompatible and sterile insect techniques
- Coordinator: H. Colinet
- ERABLE participant(s): F. Vavre
- Type: ANR PCRI (2015-2018)
- Web page: Not available

8.1.1.9. Swing

- Title: Worldwide invasion of the Spotted WING Drosophila: Genetics, plasticity and evolutionary potential
- Coordinator: P. Gibert
- ERABLE participant(s): C. Vieira
- Type: ANR PCR (2016-2020)
- Web page: Not available

8.1.2. ADT Inria

8.1.2.1. ADT Inria Kirikomics

- Main objective: Development of a portal to increase the visibility of the tools and resources elaborated by Erable around the analysis – using omics data – of metabolic networks modelled by hypergraphs, and enable to visualise the results. (the web page is for now private, it will be made public later in the project).
- Duration: 2016-2017, renewable one more year.
- Person responsible for ADT: Arnaud Mary with David Parsons (Inria).
- Beneficiary of ADT: Martin Wannagat.
- Funds received: Salary for engineer.

8.1.3. Others

Notice that were included here national projects of our members from Italy and the Netherlands when these have no other partners than researchers from the same country.

8.1.3.1. Advanced computational methodologies for the analysis of biomedical data

- Title: Advanced computational methodologies for the analysis of biomedical data
- Coordinator: P. Milazzo
- ERABLE participant(s): R. Grossi, N. Pisanti
- Type: PRA, MIUR PRIN, Italian Ministery of Research National Projects (2017-2018)
- Web page: Not available

8.1.3.2. Advanced Tools and Techniques for the analysis of criminal networks

- Title: Advanced Tools and Techniques for the analysis of criminal networks
- Coordinator: G. Italiano
- ERABLE participant(s): G. Italiano
- Type: LEONARDO SpA (2015-2018)
- Web page: Not available

8.1.3.3. Amanda

- Title: Algorithmics for MAssive and Networked DAta
- Coordinator: G. Di Battista (University of Roma 3)
- ERABLE participant(s): R. Grossi, G. Italiano, N. Pisanti
- Type: MIUR PRIN, Italian Ministery of Research National Projects (2014-2017)
- Web page: http://www.dia.uniroma3.it/~amanda/

8.1.3.4. CMACBioSeq

- Title: Combinatorial Methods for analysis and compression of biological sequences
- Coordinator: G. Rosone
- ERABLE participant(s): N. Pisanti
- Type: SIR, MIUR PRIN, Italian Ministery of Research National Projects (2015-2019)
- Web page: http://pages.di.unipi.it/rosone/CMACBioSeq.html

8.1.3.5. Statistical Models for Structural Genetic Variants in the Genome of the Netherlands

- Title: Statistical Models for Structural Genetic Variants in the Genome of the Netherlands
- Coordinator: A. Schönhuth
- ERABLE participant(s): A. Schönhuth
- Type: Nederlandse Wetenschappelijke Organisatie (NWO) (2013-2018)
- Web page: Not available

8.1.3.6. TALS and splicing

- Title: Development of bioinformatic methods for the analysis of splicing events in patients with the Taybi-Linder Syndrome (TALS)
- Coordinator: P. Edery
- ERABLE participant(s): C. Benoît-Pilven, Audric Cologne, V. Lacroix
- Type: INSERM
- Web page: Not available

8.2. European Initiatives

8.2.1. FP7 & H2020 Projects

8.2.1.1. MicroWine

- Title: Microbial metagenomics and the modern wine industry

- Duration: January 2015 - January 2019
- Coordinator: Lars Hestbjerg Hansen, University of Copenhagen
- ERABLE participant(s): A. Marchetti-Spaccamela, A. Mary, H. T. Pusa, M.-F. Sagot, L. Stougie
- Type: H2020-MSCA-ETN-2014
- Web page: https://team.inria.fr/erable/en/microwine/ and http://www.microwine.eu/

8.2.2. Collaborations in European Programs, Except FP7 & H2020

8.2.2.1. Combinatorics of co-evolution

- Title: The combinatorics of co-evolution
- Duration: 2015 - 2018
- Coordinator: Katharina Huber, University of Warwick, UK
- ERABLE participant(s): M.-F. Sagot, B. Sinaimeri
- Type: The Royal Society
- Web page: not available

8.2.3. Collaborations with Major European Organisations

By itself, ERABLE is built from what initially were collaborations with some major European Organisations (CWI, Sapienza University of Rome, Universities of Florence and Pisa, Free University of Amsterdam) and now has become a European Inria Team.

8.3. International Initiatives

8.3.1. Inria International Labs

ERABLE participates in a project within the Inria-Chile CIRIC (Communication and Information Research and Innovation Center) titled "Omics Integrative Sciences". The main objectives of the project are the development and implementation of mathematical and computational methods and the associated computational platforms for the exploration and integration of large sets of heterogeneous omics data and their application to the production of biomarkers and bioidentification systems for important Chilean productive sectors. The project started in 2011 and is coordinated in Chile by Alejandro Maass, Mathomics, University of Chile, Santiago. It is in the context of this project that we hosted Alex di Genova in ERABLE as a PhD sandwich student (for 18 months in 2015-2017). Alex has now defended his PhD. He was co-supervised by Gonzalo Ruz from the University Adolfo Ibañez, Santiago, Chile. He now, since Dec 2017, joined again ERABLE as postdoc.

8.3.2. Inria Associate Teams Not Involved in an Inria International Lab

ALEGRIA

- Title: ALgorithms for ExplorinG the inteRactions Involving Apicomplexa and kinetoplastida
- Duration: 2015-2017
- Coordinator: On the Brazilian side, Andréa Rodrigues Ávila; on the French side, Marie-France Sagot
- ERABLE participant(s): M. Ferrarini, L. Ishi Soares de Lima, A. Mary, H. T. Pusa, M.-F. Sagot, M. Wannagat
- Web page: http://team.inria.fr/erable/en/alegria/

8.3.3. Participation in Other International Programs

ERABLE is coordinator of a CNRS-UCBL-Inria Laboratoire International Associé (LIA) with the Laboratório Nacional de Computação Científica (LNCC), Petrópolis, Brazil. The LIA has for acronym LIRIO ("Laboratoire International de Recherche en bIOinformatique") and is coordinated by Ana Tereza Vasconcelos from the LNCC and Marie-France Sagot from BAOBAB-ERABLE. The LIA was created in January 2012 for 4 years, renewable once. A web page for the LIA LIRIO is available at this address: http://team.inria.fr/erable/en/cnrs-lia-laboratoire-international-associe-lirio/.

ERABLE has a Stic AmSud project that started in 2016 for 2 years. The title of the project is "Methodological Approaches Investigated as Accurately as possible for applications to biology", and its acronym MAIA. This project involves the following partners: (France) Marie-France Sagot, ERABLE Team, Inria; (Brazil) Roberto Marcondes César Jr, Instituto de Matemática e Estatística, Universidade de São Paulo; and Paulo Vieira Milreu, TecSinapse; (Chile) Vicente Acuña, Centro de Modelamiento Matemático, Santiago; and Gonzalo Ruz, University Adolfo Ibañez, Santiago. One of them, TecSinapse, is an industrial partner. MAIA has two main goals: one methodological that aims to explore how accurately hard problems can be solved theoretically by different approaches – exact, approximate, randomised, heuristic – and combinations thereof, and a second that aims to better understand the extent and the role of interspecific interactions in all main life processes by using the methodological insights gained in the first goal and the algorithms developed as a consequence. A succint web page for MAIA is available at this address: http://team.inria.fr/erable/en/projects/maia/.

ERABLE also participated to the BASIS project. This was funded by the European Community Seventh Framework Programme (Grant 242006 - 2010-2015). It was led by Dr. Mike Stratton and involved six European countries. It was primarily focused on ER+/HER2- breast cancers, but during the course of the project, was merged with the HER2+ French-ICGC and triple negative UK-ICGC projects, resulting in the analysis of the whole spectrum of breast cancers. The French group was initiated by Dr. Gilles Thomas and was pursued by Alain Viari after the loss of Dr. Thomas in 2014. The project resulted in the sequencing and thorough analysis of 560 breast cancer whole genomes (Nik-Zainai *et al.*, *Nature*, 534:47-54, 2016), including 75 HER2+ performed by the French working group (Ferrari *et al.*, *Nature Communications*, 7, 2016) and funded by the Institut National du Cancer and by Inserm.

Finally, Marie-France Sagot participates in a Portuguese FCT project, Perseids for "Personalizing cancer therapy through integrated modeling and decision" (2016-2019), with Susana Vinga and a number of other Portuguese researchers. The budget of Perseids is managed exclusively by the other Portuguese partner.

8.4. International Research Visitors

8.4.1. Visits of International Scientists

In 2017, ERABLE greeted the following International scientists:

- In France: Katharina Huber and Vincent Moulton (University of Warwick, UK), Ifigeneia Kyrkou (Aarhus University, Danemark), three members of the LIA LIRIO (Arnaldo Zaha from the Federal University of Rio Grande do Sul, Maria Cristina Motta from the Federal University of Rio de Janeiro, and Ana Tereza Vasconcelos from the LNCC, both in Brazil), two members of the Inria Associated Team Alegria (Andréa Ávila and Helisson Faoro), Ariel Silber (University of São Paulo, Brazil), Susana Vinga and various members of her team (IST Portugal).
- In Italy: May Alzamel, Lorraine A. K. Ayad, Panagiotis Charalampopoulos, Costas Iliopoulos, and Solon Pissis (King's College, London, UK) visited the University of Pisa as did Luca Cardelli (Microsoft Research), Giulia Bernardini (University of Milano Bicocca), Anthony Cox (Illumina) and Raffaele Giancarlo (University of Palermo); Loukas Georgiadis (University of Ioannina, Greece), Shahbaz Khan (University of Vienna, Austria), and Adam Karczmarcz, (University of Warsaw, Poland) visited the University of Rome Tor Vergata.
- In the Netherlands: Martin Dyer (Leeds University, England), Frans Schalekamp(Cornell University, Ithaca, New York, USA), and Anke van Zuylen (College of William and Mary, Virginia, USA) visited the FU & CWI.

8.4.2. Internships

In 2017, ERABLE greeted the following internship students:

- In France: Irene Ziska, Master Free University Berlin (6 months).

8.4.3. Visits to International Teams

In 2017, members of ERABLE visited the following International teams:

- From France: Visit to members of the LIA LIRIO at the LNCC in Brazil, the Department of Computer Science of the University of São Paulo and to members of the TecSinapse company in Brazil, Susana Vinga and members of her team (IST Portugal).

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific events organisation

9.1.1.1. General chair, scientific chair

- Giuseppe Italiano is member of the Steering Committee Member of the Workshop on Algorithm Engineering and Experimentation (ALENEX), the International Conference on Algorithms and Complexity (CIAC), and the Workshop/Symposium on Experimental Algorithms (SEA).
- 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)*.
- Marie-France Sagot is member of the Steering Committee of *European Conference on Computational Biology (ECCB)* and *International Symposium on Bioinformatics Research and Applications (ISBRA)*.

9.1.1.2. Member of the Organising Committees

- Alexander Schönhuth was main organiser of "Data Structures in Bioinformatics", Amsterdam, 21-22/02/2017, and publicity chair of RECOMB 2017.
- Fabrice Vavre was member of the Organisation Committee of the Colloquium Immuninv ("Immunité des invertébrés") in 2017.
- Cristina Vieira co-organised the Symposium "Evolutionary implications of transposable elements, epigenetics, and non-genetic inheritance" at the biennial ESEB (European Society for Evolutionary Biology) congress in 2017.

9.1.2. Scientific Events Selection

9.1.2.1. Chair of Conference Program Committees

- Leen Stougie was Chair of the Program Committee of *MAPSP* in 2017.

9.1.2.2. Member of the Conference Program Committees

- Pierluigi Crescenzi was a member of the Program Committee of *ICTCS* in 2017.
- Roberto Grossi was a member of the Program Committee of *ICALP* in 2017.
- Vincent Lacroix was a member of the Program Committee of *JOBIM* in 2017.
- Giuseppe Italiano was a member of the Program Committee of *SPIRE*, *FAW*, and *STACS* in 2017.
- Nadia Pisanti was a member of the program committee of *WALCOM*, *BIOINFORMATICS*, *CPM*, *ISBRA*, *CIBB*, *Hi BI BI*, emphWABI, and *ICCABS* in 2017.
- Marie-France Sagot was a member of the Program Committee of *ISMB/ECCB*, *PSC*, *RECOMB-CG*, and *WABI* in 2017.
- Alexander Schönhuth was a member of the Program Committee of *RECOMB*, *ISMB*, *ACM-BCB*, *GCB* (German bioinformatics conference), and *BioSB* (Dutch bioinformatics conference) in 2017.
- Leen Stougie was member of the Program Committee of *RECOMB* in 2017.

9.1.2.3. Reviewer

Besides the above, various other members of ERABLE have been reviewer for other international conferences, such as *SODA* etc.

9.1.3. Journal

9.1.3.1. Member of the Editorial Boards

- Pierluigi Crescenzi is member of the Editorial Board of *Journal of Computer and Systems Science and Electronic Notes on Theoretical Computer Science*.
- Roberto Grossi is member of the Editorial Board of *Theory of Computing Systems (TOCS)* and pf *RAIRO – Theoretical Informatics and Applications*.
- Giuseppe Italiano is member of Editorial Board of *Algorithmica* and *Theoretical Computer Science*.
- Alberto Marchetti-Spaccamela is member of the Editorial Board of *Theoretical Computer Science*.
- Arnaud Mary is Editor-in-Chief of a special issue of *Discrete Applied Mathematics* dedicated to *WEPA* 2016.
- Nadia Pisanti is since 2012 member of Editorial Board of *International Journal of Computer Science and Application (IJCSA)* and 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*, *Journal of Discrete Algorithms*, and *Lecture Notes in BioInformatics*.
- Leen Stougie is member of the Editorial Board of *Surveys in Operations Research and Management Science* since 2011, and of *Journal of Industrial and Management Optimization* since 2013.
- Cristina Vieira is Executive Editor of *Gene*, and since 2014 member of the Editorial Board of *Mobile DNA*.

9.1.3.2. Reviewer - Reviewing Activities

Members of ERABLE have reviewed papers for the following journals: *Theoretical Computer Science*, *Algorithmica*, *IEEE/ACM Transactions in Computational Biology and Bioinformatics (TCBB)*, *Algorithms for Molecular Biology*, *Scientific Reports*, *Journal of Computational Biology*, *BMC Bioinformatics*, *Computing and Informatics*, *BMC Evolutionary Biology*, *Genetica*, *Gene*, *Genome Biology and Evolution*, *Genetical Research*, *Genome Research*, *Molecular Biology and Evolution*, *Insect Biochemistry and Molecular Biology*, *PLoS Genetics*, *Mutation research*, *mBio*, *Frontiers in Microbiology*, *Infection, genetics and evolution*, *PLoS Biology*. *Nature Communications*.

9.1.4. Invited Talks

Giuseppe Italiano gave talks at the 4th Workshop on Graph Theory, Algorithms and Applications, Erice, Italy, 5-13 May, 2017 on "2-Connectivity in Directed Graphs", and at the 10th International Conference on Algorithms and Complexity (CIAC 2017), Athens, Greece, May 24-26, 2017 (http://www.corelab.ntua.gr/ciac2017/) on "2-Edge- and 2-Vertex- Connectivity in Directed Graphs".

Vincent Lacroix gave talks at the MIAT INRA Seminar, Toulouse, Feb 2017 on "De novo identification of SNPs from RNA-seq data in non-model species", and at the Aviesan ITMO Santé 9th annual meeting (https://its.aviesan.fr/index.php?pagendx=1046), Lyon, Nov 2017 on "Complementarity of assembly-first and mapping-first approaches for alternative splicing annotation and differential analysis from RNA-seq data".

Alexander Schönhuth gave talks at the FCHealth Meeting, Helsinki (organiser: Veli Mäkinen) on "Disentangling the genetic diversity of pathogens at the strain level using overlap graphs", the Computational Genomics Summer Institute (http://computationalgenomics.bioinformatics.ucla.edu), University of California at Los Angeles on "Polyploid Genome Assembly", and the Bertinoro Computational Biology (BCB) 2017 (https://bertinoro2017.wordpress.com, organisers: Sohrab Shah, UBC Vancouver; Jens Lagergren, Stockholm), on "Discovering somatic variants in single cells".

Leen Stougie gave a talk at the Workshop on Stochastic Programming Honoring Maarten van der Vlerk, Groningen, The Netherlands, 10-11 August 2017, on "Stochastic and Robust Scheduling".

9.1.5. Leadership within the Scientific Community

Giuseppe Italiano is Member of the Council of EATCS, the European Association for Theoretical Computer Science.

Leen Stougie is Member of the general board ofthe Dutch Network on the Mathematics of Operations Research (Landelijk Netwerk Mathematische Besliskunde (LNMB)). He is also Chairman Program Committee Econometrics and OR, VU Amsterdam and Member of the Board of the research school ABRI-VU, Amsterdam.

Marie-France Sagot and Fabrice Vavre are members of the Steering Committee of the LabEx Ecofect (http://ecofect.universite-lyon.fr/).

9.1.6. Scientific Expertise

Marie-France Sagot is member of the Advisory Board of the CWI, Amsterdam, The Netherlands, and chair of the "Commissions Scientifiques Spécialisées" (CSS) of the INRA for the Department of Applied Mathematics and Computer Science.

Fabrice Vavre is president of the Section 29 of the Comité National de la Recherche Scientifique (CoNRS).

9.1.7. Research Administration

Hubert Charles is director of the Biosciences Department of the Insa-Lyon and co-director of studies of the "Bioinformatique et Modélisation (BIM)" track.

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

Alberto Marchetti-Spaccamela is Director of the Department of Computer Engineering and Management Antonio Ruberti at Sapienza University since 2013.

Nadia Pisanti was since 2013 and until October 31st 2017 member of the Board of the Regional PhD School of Computer Science at the University of Pisa, Italy. Since November 1st 2017, she is 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).

Alexander Schönhuth is member of the Scientific Board of BioSB (the Dutch organisation for bioinformatics) since May 2017.

Leen Stougie is since April 2017 Leader of the Life Science Group at CWI.

Alain Viari is since 2012 Deputy Scientific Director at Inria responsible for the ICST for Life and Environmental Sciences. He thus represents Inria at several national instances related to Life Sciences and Health (Allenvi, Aviesan, Ibisa, etc.). He 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.

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

9.2.1.1. 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 (https://www.bee-lyon-univ.fr/). She teaches genetics 192 hours per year at the University and at the ENS-Lyon. Hubert Charles is responsible for the Master of Modelling and Bioinformatics (BIM) at the Insa of Lyon (http://biosciences.insa-lyon.fr/). He teaches 192 hours per year in statistics and biology. Vincent Lacroix is responsible for several courses of the newly created Master in Bioinformatics (https://www.bioinfo-lyon.fr/) (L3: Advanced Bioinformatics, M1: Methods for Data Analysis in Genomics,

M1: Methods for Data Analysis in Transcriptomics, M1: Bioinformatics Project, M2: Ethics). He teaches 192 hours per year in bioinformatics. Arnaud Mary is responsible for two courses of the Bioinformatics Curriculum at the University (L2: Introduction to Bioinformatics and Biostatistics, M1: Object Oriented Programming) and one at Insa (Discrete Mathematics). He has been teaching 150 hours per year for his three first years as Assistant and then as Associated Professor, and will now, as of September 2017, teach 192 hours per year. Blerina Sinaimeri taught 30 hours per year in graph algorithms for the M1 students of the Master in Bioinformatics, and approximately 12h per year (Discrete Mathematics) at Insa. In 2017, she also taught 24h per year at the Master in Computer Science of the ENS-Lyon.

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

Vincent Lacroix was an instructor in NGS data analysis training for the CNRS Formation in the last 3 years since 2015, a course coordinated by Eric Rivals from the LIRMM, Montpellier.

Alain Viari was a lecturer at the 3rd International School on Breast Cancer organised at the Curie Institute in June 2017. Together with Vincent Lacroix, he participated to a transdisciplinary module for PhDs on Ethics organised by the University of Lyon.

All French members of the ERABLE team are affiliated to the doctoral school E2M2 (Ecology-Evolution-Microbiology-Modelling, http://e2m2.universite-lyon.fr/).

9.2.1.2. Italy & The Netherlands

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

Dutch researchers teach between 40 and 270 hours per year, again at the undergraduate and Master levels, in pure computer science (*e.g.* Algorithm Engineering, Randomised Algorithms), applied mathematics (*e.g.* Operational Research, Advanced Linear Programming) and computational biology (*e.g.* Biological Network Analysis).

9.2.2. Supervision

The following PhDs were defended in ERABLE in 2017:

- Hélène Lopez-Maestre, University of Lyon 1, Feb 2017, supervisors: V. Lacroix, C. Vieira.
- Laura Urbini, University of Lyon 1, October 2017, supervisors: C. Matias, M.-F. Sagot, B. Sinaimeri.
- Martin van Ee, December 2017, supervisor: L. Stougie.

The following are the PhDs in progress:

- Audric Cologne, University of Lyon 1 (funded by Inserm and Inria, co-supervisors: Patrick Edery – Federation of Health Research of Lyon-Est, Vincent Lacroix)
- Mattia Gastaldello, Sapienza University of Rome and University of Lyon 1 (funded by "Vinci Program-Université Franco-Italienne", co-supervisors: Tiziana Calamoneri, Sapienza University of Rome; Marie-France Sagot)
- Leandro Ishi Soares de Lima, University of Lyon 1 (funded by the Brazilian "Science without Borders" program, co-supervisors: Giuseppe Italiano, Vincent Lacroix, Marie-France Sagot)
- Carol Moraga Quinteros, University of Lyon 1 (funded by Conicyt Chile, co-supervisors: Rodrigo Gutierrez – Catholic University of Chile, Marie-France Sagot)
- Henri Taneli Pusa, University of Lyon 1 (funded by H2020-MSCA-ETN-2014 project MicroWine, co-supervisors: Alberto Marchetti-Spaccamela, Arnaud Mary, Marie-France Sagot)
- Camille Sessegolo, University of Lyon 1 (funded by ANR Aster; co-supervisors: Vincent Lacroix, Arnaud Mary)
- Irene Ziska, University Lyon 1 (funded by Inria Cordi-S, co-supervisors: Susana Vinga – Instituto Superior Técnico at Lisbon; Marie-France Sagot)

Besides the PhD students indicated above, who are physically located within one of the premises of ERABLE, the project-team has PhD students in co-supervision who spend the majority or the whole of their time in the premises of other teams. These include: Scheila G. Mucha (funded by Brazil, co-supervisors: Arnaldo Zaha – Federal University of Rio Grande do Sul in Brazil, Marie-France Sagot), Rita Ramos (funded by Portuguese FCT, co-supervisors: Claudia Nunes dos Santos – ITQB Lisbon, Marie-France Sagot), André Veríssimo (funded by Portuguese FCT, co-supervisors: Susana Vinga – Instituto Superior Técnico in Lisbon, Marie-France Sagot). Furthermore, ERABLE had for 18 months a "sandwich" (meaning part-time) PhD student in the team, Alex di Genova, whose supervisor is Gonzalo Ruz from the University Adolfo Ibañez in Santiago, Chile (A. di Genova defended his PhD on November 14 2017 and has since joined ERABLE as a postdoc).

9.2.3. Juries

The following are the PhD or HDR juries to which members of ERABLE participated in 2017.

- Giuseppe Italiano: External Examiner of the PhD of Mathias baek Tejs Knudsen, DIKU, Copenhagen University, August 4, 2017.
- Nadia Pisanti: External Reviewer of the PhD of Mika Amit, University of Haifa, Israel, July 3rd 2017.

9.3. Popularisation

- Ricardo Andrade participated to the "Fête de la Science" at Inria.
- Vincent Lacroix co-organised with Nicolas Lechopier a conference on synthetic biology followed by a live radio session involving the speaker Catherine Bourgain, and the students from the Master in Bioinformatics http://sciencespourtous.univ-lyon1.fr/sciences-lemission-science-democratie-parole-aux-etudiants/.
- Blerina Sinaimeri participated to the Conference "Filles et mathématiques : Une équation lumineuse !" (see https://filles-et-math.sciencesconf.org/program) and to the "Fête de la Science" at Inria.
- Alain Viari as Deputy Scientific Director for Life Sciences at Inria regularly gives talks to external partners and general audiences to present the Life Sciences topics at Inria.

10. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] H. LOPEZ-MAESTRE. *Analyses and methods for transcriptomic data from non-model species: Variation in the expression of transposable elements (and genes) and nucleotide variants*, Université Claude Bernard Lyon 1, February 2017, https://hal.inria.fr/tel-01575640

[2] L. URBINI. *Models and algorithms to study the common evolutionary history of hosts and symbionts*, Université Claude Bernard - Lyon I, October 2017, https://hal.inria.fr/tel-01673445

Articles in International Peer-Reviewed Journals

[3] M. BAILLY-BECHET, P. MARTINS-SIMÕES, G. SZÖLLŐSI, G. MIALDEA, M.-F. SAGOT, S. CHARLAT. *How Long Does Wolbachia Remain on Board?*, in "Molecular Biology and Evolution", 2017, vol. 34, pp. 1183 - 1193 [*DOI :* 10.1093/MOLBEV/MSX073], https://hal.inria.fr/hal-01524867

[4] L. BEAUDOU, A. MARY, L. NOURINE. *Algorithms for k-meet-semidistributive lattices*, in "Theoretical Computer Science", 2017, vol. 658, pp. 391 - 398 [*DOI :* 10.1016/J.TCS.2015.10.029], https://hal.inria.fr/hal-01571243

[5] A. CONTE, R. GROSSI, A. MARINO, R. RIZZI. *Efficient enumeration of graph orientations with sources*, in "Discrete Applied Mathematics", August 2017 [*DOI :* 10.1016/J.DAM.2017.08.002], https://hal.inria.fr/hal-01609009

[6] H. DAVIES, D. GLODZIK, S. MORGANELLA, L. R. YATES, J. STAAF, X. ZOU, M. RAMAKRISHNA, S. MARTIN, S. BOYAULT, A. M. SIEUWERTS, P. T. SIMPSON, T. A. KING, K. RAINE, J. E. EYFJORD, G. KONG, Å. BORG, E. BIRNEY, H. G. STUNNENBERG, M. J. VAN DE VIJVER, A. BØRRESEN-DALE, J. W. M. MARTENS, P. N. SPAN, S. R. LAKHANI, A. VINCENT-SALOMON, C. SOTIRIOU, A. TUTT, A. M. THOMPSON, S. VAN LAERE, A. L. RICHARDSON, A. VIARI, P. J. CAMPBELL, M. R. STRATTON, S. NIK-ZAINAL. *HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures*, in "Nature Medicine", 2017, vol. 23, n^{o} 4, pp. 517 - 525 [*DOI :* 10.1038/NM.4292], https://hal.inria.fr/hal-01525050

[7] H. DAVIES, S. MORGANELLA, C. A. PURDIE, S. J. JANG, E. BORGEN, H. RUSSNES, D. GLODZIK, X. ZOU, A. VIARI, A. L. RICHARDSON, A.-L. BØRRESEN-DALE, A. THOMPSON, J. E. EYFJORD, G. KONG, M. R. STRATTON, S. NIK-ZAINAL. *Whole-Genome Sequencing Reveals Breast Cancers with Mismatch Repair Deficiency*, in "Cancer Research", 2017, vol. 77, n^{o} 18, pp. 4755-4762 [*DOI :* 10.1158/0008-5472.CAN-17-1083], https://hal.inria.fr/hal-01599736

[8] A. DUDNIK, F. ALMEIDA, R. ANDRADE, B. AVILA, P. BAÑADOS, B. DIANE, J.-E. BASSARD, M. BENKOULOUCHE, M. BOTT, A. BRAGA, D. BREITEL, R. BRENNAN, L. BULTEAU, C. CHANFORAN, I. COSTA, R. S. C. COSTA, M. DOOSTMOHAMMADI, N. FARIA, C. FENG, R. FERRO, A. FERNANDES, P. FERREIRA, A. FOITO, S. FREITAG, C. JARDIM, G. GARCIA, P. GASPAR, J. GODINHO-PEREIRA, B. HAMBERGER, A. HARTMANN, H. HEIDER, S. LI, A. JULIEN-LAFERRIÈRE, N. KALLSCHEUER, W. KERBE, O. P. KUIPERS, N. LOVE, A. MARCHETTI-SPACCAMELA, J. MARIENHAGEN, C. MARTIN, A. MARY, V. MAZUREK, C. MEINHART, D. M. SEVILLANO, R. MENEZES, M. NAESBY, M. H. H. NØRHOLM, F. T. OKKELS, J. OLIVEIRA, M. OTTENS, D. PARROT, L. PEI, I. ROCHA, R. ROSADO-RAMOS, C. ROUSSEAU, M.-F. SAGOT, C. NUNES DOS SANTOS, M. SCHMIDT, T. SHELENGA, L. SHEPHERD, A. RITA SILVA, M. HENRIQUES DA SILVA, O. SIMON, S. G. STAHLHUT, A. SOLOPOVA, A. SOROKIN, D. STEWART, L. STOUGIE, M. TRICK, S. SU, V. THOLE, O. TIKHONOVA, P. VAIN, A. VERÍSSIMO, A. VILA-SANTA, S. VINGA, M. VOGT, L. WANG, L. WANG, W. WEI, S. YOUSSEF, A. RUTE, J. FÖRSTER. *BacHBerry: BACterial Hosts for production of Bioactive phenolics from bERRY fruits*, in "Phytochemistry Reviews", 2017, pp. 1-36 [*DOI :* 10.1007/S11101-017-9532-2], https://hal.inria.fr/hal-01673596

[9] O. DURON, F. BINETRUY, V. NOËL, J. CREMASCHI, K. MCCOY, C. ARNATHAU, O. PLANTARD, J. GOOLSBY, A. PÉREZ DE LEÓN, D. HEYLEN, A. R. VAN OOSTEN, Y. GOTTLIEB, G. BANETH, A. GUGLIELMONE, A. ESTRADA-PEÑA, M. OPARA, L. ZENNER, F. VAVRE, C. CHEVILLON. *Evolutionary changes in symbiont community structure in ticks*, in "Molecular Ecology", 2017, vol. 26, n^{o} 11, pp. 2905–2921 [*DOI :* 10.1111/MEC.14094], https://hal.inria.fr/hal-01523998

[10] J.-P. FOY, L. BAZIRE, S. ORTIZ-CUARAN, S. DENEUVE, J. KIELBASSA, E. THOMAS, A. VIARI, A. PUISIEUX, P. GOUDOT, C. BERTOLUS, N. FORAY, Y. KIROVA, P. VERRELLE, P. SAINTIGNY. *A 13-gene expression-based radioresistance score highlights the heterogeneity in the response to radiation therapy across HPV-negative HNSCC molecular subtypes*, in "BMC Medicine", December 2017, vol. 15, n^{o} 1, pp. 703 - 719 [*DOI :* 10.1186/S12916-017-0929-Y], https://hal.inria.fr/hal-01599731

[11] P. FOY, C. BERTOLUS, M.-C. MICHALLET, S. DENEUVE, R. INCITTI, N. BENDRISS-VERMARE, M.-A. ALBARET, S. ORTIZ-CUARAN, E. THOMAS, A. COLOMBE, C. PY, N. GADOT, J.-P. MICHOT, J. FAYETTE, A. VIARI, B. VAN DEN EYNDE, P. GOUDOT, M. DEVOUASSOUX-SHISHEBORAN, A. PUISIEUX, C. CAUX, P. ZROUNBA, S. LANTUEJOUL, P. SAINTIGNY. *The immune microenvironment of HPV-negative oral*

squamous cell carcinoma from never-smokers and never-drinkers patients suggests higher clinical benefit of IDO1 and PD1/PD-L1 blockade, in "Annals of Oncology", April 2017, vol. 28, n° 8, pp. 1934–1941 [*DOI :* 10.1093/ANNONC/MDX210], https://hal.inria.fr/hal-01560943

[12] D. GLODZIK, S. MORGANELLA, H. DAVIES, P. T. SIMPSON, Y. LI, X. ZOU, J. DIEZ-PEREZ, J. STAAF, L. B. ALEXANDROV, M. SMID, A. B. BRINKMAN, I. H. RYE, H. RUSSNES, K. RAINE, C. A. PURDIE, S. R. LAKHANI, A. M. THOMPSON, E. BIRNEY, H. G. STUNNENBERG, M. J. VAN DE VIJVER, J. W. M. MARTENS, A.-L. BØRRESEN-DALE, A. L. RICHARDSON, G. KONG, A. VIARI, D. EASTON, G. EVAN, P. J. CAMPBELL, M. R. STRATTON, S. NIK-ZAINAL. *A somatic-mutational process recurrently duplicates germline susceptibility loci and tissue-specific super-enhancers in breast cancers*, in "Nature Genetics", 2017, vol. 49, n° 3, pp. 341 - 348 [*DOI :* 10.1038/NG.3771], https://hal.inria.fr/hal-01525728

[13] C. GOUBERT, H. HENRI, G. MINARD, C. VALIENTE MORO, P. MAVINGUI, C. VIEIRA, M. BOULESTEIX. *High-Throughput Sequencing of Transposable Element Insertions Suggests Adaptive Evolution of the Invasive Asian Tiger Mosquito Towards Temperate Environments*, in "Molecular Ecology", 2017, pp. 1-14 [*DOI :* 10.1111/MEC.14184], https://hal.inria.fr/hal-01546749

[14] R. GROSSI, G. MENCONI, N. PISANTI, R. TRANI, S. VIND. *Fast Output-Sensitive Pattern Discovery in Massive Sequences using the Motif Trie*, in "Theoretical Computer Science", 2017, 25 p. [*DOI :* 10.1016/J.TCS.2017.04.012], https://hal.inria.fr/hal-01525745

[15] A. HARTMANN, A. VILA-SANTA, N. KALLSCHEUER, M. VOGT, A. JULIEN-LAFERRIÈRE, M.-F. SAGOT, J. MARIENHAGEN, S. VINGA. *OptPipe - a pipeline for optimizing metabolic engineering targets*, in "BMC Systems Biology", December 2017, vol. 11, pp. 1-9 [*DOI :* 10.1186/S12918-017-0515-0], https://hal.inria.fr/hal-01672905

[16] S. HUET, E. SZAFER-GLUSMAN, B. TESSON, L. XERRI, W. J. FAIRBROTHER, K. MUKHYALA, C. BOLEN, E. PUNNOOSE, L. TONON, C. CHASSAGNE-CLÉMENT, P. FEUGIER, A. VIARI, F. JARDIN, G. SALLES, P. SUJOBERT. *BCL2 mutations do not confer adverse prognosis in follicular lymphoma patients treated with rituximab*, in "American Journal of Hematology", 2017, vol. 92, n° 6, pp. 515 - 519 [*DOI :* 10.1002/AJH.24701], https://hal.inria.fr/hal-01524901

[17] S. HUET, L. XERRI, B. TESSON, S. MARESCHAL, S. TAIX, L. MESCAM-MANCINI, E. SOHIER, C. CARRÈRE, J. LAZAROVICI, O. CASASNOVAS, L. TONON, S. BOYAULT, S. HAYETTE, C. HAIOUN, B. FABIANI, A. VIARI, F. JARDIN, G. SALLES. *EZH2 alterations in follicular lymphoma: biological and clinical correlations*, in "Blood Cancer Journal", 2017, vol. 7, n° 4 [*DOI :* 10.1038/BCJ.2017.32], https://hal.inria.fr/hal-01566524

[18] Y. S. JU, I. MARTINCORENA, M. GERSTUNG, M. PETLJAK, L. B. ALEXANDROV, R. RAHBARI, D. C. WEDGE, H. R. DAVIES, M. RAMAKRISHNA, A. FULLAM, S. MARTIN, C. ALDER, N. PATEL, S. GAMBLE, S. O'MEARA, D. D. GIRI, T. SAUER, S. E. PINDER, C. A. PURDIE, Å. BORG, H. STUNNENBERG, M. VAN DE VIJVER, B. K. T. TAN, C. CALDAS, A. TUTT, N. T. UENO, L. J. VAN 'T VEER, J. W. M. MARTENS, C. SOTIRIOU, S. KNAPPSKOG, P. N. SPAN, S. R. LAKHANI, J. E. EYFJÖRD, A.-L. BØRRESEN-DALE, A. RICHARDSON, A. M. THOMPSON, A. VIARI, M. E. HURLES, S. NIK-ZAINAL, P. J. CAMPBELL, M. R. STRATTON. *Somatic mutations reveal asymmetric cellular dynamics in the early human embryo*, in "Nature", 2017, vol. 543, n° 7647, pp. 714 - 718 [*DOI :* 10.1038/NATURE21703], https://hal.inria.fr/hal-01525712

[19] E. LERAT, M. FABLET, L. MODOLO, H. LOPEZ-MAESTRE, C. VIEIRA. *TEtools facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of*

piRNA genes, in "Nucleic Acids Research", 2017, vol. 45, 13 p. [*DOI :* 10.1093/NAR/GKW953], https://hal.inria.fr/hal-01524877

[20] T. LEFÉBURE, C. MORVAN, F. MALARD, C. FRANÇOIS, L. KONECNY-DUPRÉ, L. GUÉGUEN, M. WEISS-GAYET, A. SEGUIN-ORLANDO, L. ERMINI, C. D. SARKISSIAN, N. P. CHARRIER, D. EME, F. MERMILLOD-BLONDIN, L. DURET, C. VIEIRA, L. ORLANDO, C. J. DOUADY. *Less effective selection leads to larger genomes*, in "Genome Research", 2017, vol. 27, pp. 1016-1028 [*DOI :* 10.1101/GR.212589.116], https://hal.inria.fr/hal-01544879

[21] L. LIMA, B. SINAIMERI, G. SACOMOTO, H. LOPEZ-MAESTRE, C. MARCHET, V. MIELE, M.-F. SAGOT, V. LACROIX. *Playing hide and seek with repeats in local and global de novo transcriptome assembly of short RNA-seq reads*, in "Algorithms for Molecular Biology", December 2017, vol. 12, n° 1, 2 p. [*DOI :* 10.1186/S13015-017-0091-2], https://hal.inria.fr/hal-01474524

[22] H. LOPEZ-MAESTRE, E. A. G. CARNELOSSI, V. LACROIX, N. BURLET, B. MUGAT, S. CHAMBEYRON, C. M. A. CARARETO, C. VIEIRA. *Identification of misexpressed genetic elements in hybrids between Drosophila-related species*, in "Scientific Reports", 2017, vol. 7, 40618 p. , https://hal.archives-ouvertes.fr/hal-01524879

[23] A. MALIZIA, K. A. OLSEN, T. TURCHI, P. CRESCENZI. *An ant-colony based approach for real-time implicit collaborative information seeking*, in "Information Processing and Management", 2017, vol. 53, n° 3, pp. 608 - 623 [*DOI :* 10.1016/J.IPM.2016.12.005], https://hal.inria.fr/hal-01525753

[24] A. MELANI, M. BERTOGNA, V. BONIFACI, A. MARCHETTI-SPACCAMELA, G. BUTTAZZO. *Schedulability Analysis of Conditional Parallel Task Graphs in Multicore Systems*, in "IEEE Transactions on Computers", 2017, vol. 66, n° 2, pp. 339-353 [*DOI :* 10.1109/TC.2016.2584064], https://hal.inria.fr/hal-01556802

[25] A. MELANI, M. BERTOGNA, R. DAVIS, V. BONIFACI, A. MARCHETTI-SPACCAMELA, G. BUTTAZZO. *Exact Response Time Analysis for Fixed Priority Memory-Processor Co-Scheduling*, in "IEEE Transactions on Computers", 2017, vol. 66, pp. 631 - 646 [*DOI :* 10.1109/TC.2016.2614819], https://hal.inria.fr/hal-01556792

[26] D. MONNIN, N. KREMER, E. DESOUHANT, F. VAVRE. *Impact of Wolbachia on oxidative stress sensitivity in the parasitic wasp Asobara japonica*, in "PLoS ONE", 2017, vol. 5, 10 p. [*DOI :* 10.1371/JOURNAL.PONE.0175974.S002], https://hal.inria.fr/hal-01523927

[27] G. RODRIGUES GALVAO, C. BAUDET, Z. DIAS. *Sorting Circular Permutations by Super Short Reversals*, in "IEEE/ACM Transactions on Computational Biology and Bioinformatics", 2017, vol. 14, n° 3, pp. 620-633 [*DOI :* 10.1109/TCBB.2016.2515594], https://hal.inria.fr/hal-01317003

[28] G. TOLOSA, L. BECCHETTI, E. FEUERSTEIN, A. MARCHETTI-SPACCAMELA. *Performance Improvements for Search Systems using an Integrated Cache of Lists+Intersections*, in "Information Retrieval Journal", 2017, vol. 20, n° 3, pp. 172-198 [*DOI :* 10.1007/978-3-319-11918-2_22], https://hal.inria.fr/hal-01528536

[29] L. VAN IERSEL, S. KELK, G. STAMOULIS, L. STOUGIE, O. BOES. *On Unrooted and Root-Uncertain Variants of Several Well-Known Phylogenetic Network Problems*, in "Algorithmica", 2017, vol. 64, n° 4, pp. 621 - 637 [*DOI :* 10.1093/SYSBIO/SYV020], https://hal.inria.fr/hal-01599716

International Conferences with Proceedings

[30] V. ACUNA, R. GROSSI, G. F. ITALIANO, L. LIMA, R. RIZZI, G. SACOMOTO, M.-F. SAGOT, B. SINAIMERI. *On Bubble Generators in Directed Graphs*, in "WG 2017 - 43rd International Workshop on Graph-Theoretic Concepts in Computer Science", Eindhoven, Netherlands, Lecture Notes in Computer Science, Springer, June 2017, vol. 10520, pp. 18-31 [*DOI :* 10.1007/978-3-319-68705-6_2], https://hal.inria.fr/hal-01647516

[31] G. BERNARDINI, N. PISANTI, S. PISSIS, G. ROSONE. *Pattern Matching on Elastic-Degenerate Text with Errors*, in "SPIRE 2017 - 24th International Symposium on String Processing and Information Retrieval", Palermo, Italy, SPIRE 2017: String Processing and Information Retrieval, Springer, September 2017, vol. 10508, pp. 74-90 [*DOI :* 10.1007/978-3-319-67428-5_7], https://hal.inria.fr/hal-01673585

[32] A. BJELDE, Y. DISSER, J. HACKFELD, C. HANSKNECHT, M. LIPMANN, J. MEISSNER, K. SCHEWIOR, M. SCHLÖTER, L. STOUGIE. *Tight Bounds for Online TSP on the Line*, in "ACM-SIAM Symposium on Discrete Algorithms (SODA)", Barcelona, Spain, January 2017, pp. 994 - 1005 [*DOI :* 10.1137/1.9781611974782.63], https://hal.inria.fr/hal-01549685

[33] M. BORASSI, P. CRESCENZI, L. TREVISAN. *An Axiomatic and an Average-Case Analysis of Algorithms and Heuristics for Metric Properties of Graphs*, in "SODA 2017 - Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms", Barcelona, Spain, Proceedings of the Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, January 2017, pp. 920 - 939 [*DOI :* 10.1137/1.9781611974782.58], https://hal.inria.fr/hal-01525752

[34] N. BOUSQUET, A. MARY, A. PARREAU. *Token Jumping in minor-closed classes*, in "Fundamentals of Computation Theory (FCT 2017)", Bordeaux, France, R. KLASING, M. ZEITOUN (editors), Lecture Notes in Computer Science, Springer, September 2017, vol. 10472, pp. 136-149, https://arxiv.org/abs/1706.09608 [*DOI :* 10.1007/978-3-662-55751-8_12], https://hal.archives-ouvertes.fr/hal-01634505

[35] A. CONTE, R. GROSSI, A. MARINO, L. TATTINI, L. VERSARI. *A Fast Algorithm for Large Common Connected Induced Subgraphs*, in "AlCoB 2017: 4th International Conference on Algorithms for Computational Biology", Aveiro, Portugal, June 2017, vol. 37, pp. 828 - 74 [*DOI :* 10.1021/CI9601675], https://hal.inria.fr/hal-01555996

[36] A. CONTE, R. GROSSI, A. MARINO, T. UNO, L. P. VERSARI. *Listing Maximal Independent Sets with Minimal Space and Bounded Delay*, in "International Symposium on String Processing and Information Retrieval (SPIRE)", Palermo, Italy, G. FICI, M. SCIORTINO, R. VENTURINI (editors), Lecture Notes in Computer Science, Springer, September 2017, vol. 10508, n^{o} 1-2, pp. 144-160 [*DOI :* 10.1007/978-3-319-67428-5_13], https://hal.inria.fr/hal-01609012

[37] R. P. GROSSI, C. S. ILIOPOULOS, C. LIU, N. P. PISANTI, S. P. PISSIS, A. RETHA, G. ROSONE, F. VAYANI, L. P. VERSARI. *On-line pattern matching on similar texts*, in "28th Annual Symposium on Combinatorial Pattern Matching (CPM'17)", Varsovie, Poland, 2017, pp. 7 - 8 [*DOI :* 10.4230/LIPICS.CPM.2017.07], https://hal.inria.fr/hal-01526650

Scientific Popularization

[38] P. SAINTIGNY, P. FOY, A. FERRARI, P. CASSIER, A. VIARI, A. PUISIEUX. *Apport et défis des Big Data en cancérologie*, in "Bulletin du Cancer", 2017, vol. 104, n^{o} 3, pp. 281 - 287 [*DOI :* 10.1016/J.BULCAN.2016.10.020], https://hal.inria.fr/hal-01525736

Other Publications

[39] C. BENOIT-PILVEN, C. MARCHET, E. CHAUTARD, L. LIMA, M.-P. LAMBERT, G. SACOMOTO, A. REY, C. BOURGEOIS, D. AUBOEUF, V. LACROIX. *Annotation and differential analysis of alternative splicing using de novo assembly of RNAseq data*, November 2017, working paper or preprint [*DOI :* 10.1101/074807], https://hal.archives-ouvertes.fr/hal-01643169