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 2016

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:

3. - Data and knowledge
3.1. - Data
3.1.1. - Modeling, representation
3.1.4. - Uncertain data
3.3. - Data and knowledge analysis
3.3.2. - Data mining
3.3.3. - Big data analysis
7. - Fundamental Algorithmics
7.2. - Discrete mathematics, combinatorics
7.3. - Optimization
7.9. - Graph theory
7.10. - Network science
7.11. - Performance evaluation

Other Research Topics and Application Domains:

1. - Life sciences
1.1. - Biology
1.1.1. - Structural biology
1.1.2. - Molecular biology
1.1.5. - Genetics
1.1.6. - Genomics
1.1.8. - Evolutionnary biology
1.1.9. - Bioinformatics
1.1.11. - Systems biology
1.1.12. - Synthetic biology
1.2. - Ecology
1.2.1. - Biodiversity
1.4. - Pathologies
2. - Health
2.2. - Physiology and diseases
2.2.3. - Cancer
2.2.4. - Infectious diseases, Virology
2.3. - Epidemiology

1. Members

Research Scientists

Gunnar Klau [CWI, The Netherlands, Senior Researcher]
Marie-France Sagot [Team leader, Inria, Senior Researcher, HDR]
Blerina Sinaimeri [Inria, Researcher]
Fabrice Vavre [CNRS, Senior Researcher, HDR]
Alain Viari [Inria, Senior Researcher & Deputy Scientific Director for ICST for Life and Environmental Sciences at Inria]

Faculty Members

Pierluigi Crescenzi [University of Florence, Italy, Full Professor]
Hubert Charles [INSA Lyon, Full Professor, HDR]
Roberto Grossi [University of Pisa, Italy, Full Professor]
Vincent Lacroix [University Lyon I, Associate Professor]
Alberto Marchetti-Spaccamela [Sapienza University of Rome, Italy, Full Professor]
Arnaud Mary [University Lyon I, Associate Professor]
Nadia Pisanti [University of Pisa, Italy, Assistant Professor until October 2016, Associate Professor since November 2016]
Leen Stougie [Free University Amsterdam & CWI, The Netherlands, Full Professor]
Cristina Vieira [University Lyon I, Full Professor, HDR]

Engineers

Clara Inae Benoit-Pilvin [INSERM, since Oct 2016]
Martin Wannagat [Inria, ADT Engineer, since Dec 2016]

PhD Students

Audric Cologne [Inserm & Inria, co-supervised by Patrick Edery and Vincent Lacroix, from Oct 2016]
Alex Di Genova [University of Santiago, Chile, co-supervised by Alejandro Maass (CMM) and Eric Goles (University Adolfo Ibañez), will spend 18-24 months in ERABLE, Lyon, starting from Sept 2015]
Mattia Gastaldello [Sapienza University of Rome and University of Lyon 1, grant by the Vinci Program – Université Franco-Italienne, co-supervised by Tiziana Calamoneri (Sapienza) and Marie-France Sagot, will spend half of his PhD in Lyon]
Leandro Ishi Soares de Lima [Science Without Frontiers, Ministry of Research Brazil, co-supervised by Giuseppe Italiano, Tor Vergata University of Rome, Vincent Lacroix, and Marie-France Sagot]
Alice Julien-Laferrière [Inria, grant by FP7 KBBE project, co-supervised by Vincent Lacroix, Marie-France Sagot, and Susana Vinga, IDMEC-IST, Lisbon, Portugal, until Nov 2016, defence Dec 8, 2016]
Hélène Lopez-Maestre [University Lyon 1, co-supervised by Vincent Lacroix and Cristina Vieira, until Sept 2016, defence Jan 2017]
Carol Moraga Quinteros [University of Lyon 1, PhD Student, from Oct 2016, co-supervised by Rodrigo Gutierrez and Marie-France Sagot, funded by Conicyt, Chile]
Scheila Mucha [Capes-Cofecub, Brazilian Sandwich PhD for one year until June 2015, renewed 6 months more from October 2015 funded by ERABLE, supervised by Arnaldo Zaha, Federal University of Rio Grande do Sul, Brazil]
Henri Taneli Pusa [Inria, grant by H2020-MSCA-ETN-2014 project, co-supervised by Alberto Marchetti-Spaccamela, Arnaud Mary, and Marie-France Sagot]
Laura Urbini [Inria, from October 2014, co-supervised by Catherine Matias, University Pierre et Marie Curie, Paris, and Marie-France Sagot]
André Veríssimo [IDMEC-IST, Lisbon, funded by FCT Portugal, co-supervised by Susana Vinga, IDMEC-IST, and Marie-France Sagot, spends a few months per year in Lyon]
Martin Wannagat [Inria, grant by European Research Council, co-supervised by Alberto Marchetti-Spaccamela, Marie-France Sagot, and Leen Stougie, defended in June 2016, until June 2016]

Post-Doctoral Fellows

Ricardo de Andrade Abrantes [Institute of Mathematics and Statistics, University of São Paulo, Brazil and Erable, Postdoc funded by Fapesp, working also with Erable]
Mariana Galvão Ferrarini [Inria, grant by European Research Council and Inria FRM]

Andrea Marino [University of Pisa]
Delphine Parrot [Inria, grant by FP7 KBBE project, until Aug 2016]
Martin Wannagat [Inria, grant by European Research Council, from Sept to Nov 2016]

Administrative Assistant
Marina Da Graça [Inria]

Others
Audric Cologne [Inria, Master 2, from Feb 2016 until Jun 2016]
Louis Duchemin [Inria, Master 1, from Feb 2016 until Jun 2016]
Irene Ziska [Free University of Berlin, Master, from Nov 2016 until Dec 2016]
Laurent Jacob [CNRS & LBBE, Researcher, external collaborator]
Vincent Miele [CNRS & LBBE, Research engineer, external collaborator]
Anne Morgat [SIB Geneva, Researcher, external collaborator]
Susana Vinga [IDMEC-IST Lisbon, Researcher, external collaborator]
Ana Tereza Vasconcelos [LNCC Brazil, Researcher, external collaborator, co-responsible for LIA LIRIO]

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 substracting 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 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. The first four axes thus fall in the first category, and the fifth one in the second. As concerns the first four axes, the model organisms or systems chosen will be those studied by the biologists among our permanent members or among our close collaborators. Currently these 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: 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 5: 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

The main area of application of ERABLE is biology understood in its more general sense, with a special focus on symbiosis and on intracellular interactions.

5. New Software and Platforms

5.1. AcypiCyc

FUNCTIONAL DESCRIPTION

Database of the metabolic network of *Acyrthosiphon pisum*.

- Participants: Patrice Baa Puyoule, Hubert Charles, Stefano Colella, Ludovic Cottret, Marie-France Sagot, Augusto Vellozo and Amélie Veron
- Contact: Hubert Charles
- URL: http://acypicyc.cycadsys.org/

5.2. AlViE

FUNCTIONAL DESCRIPTION

ALVIE is a post-mortem algorithm visualisation Java environment, which is based on the interesting event paradigm. The current distribution of ALVIE includes more than forty visualisations. Almost all visualisations include the representation of the corresponding algorithm C-like pseudo-code. The ALVIE distribution allows a programmer to develop new algorithms with their corresponding visualisation: the included Java class library, indeed, makes the creation of a visualisation quite an easy task (once the interesting events have been identified).

- Participants: Pierluigi Crescenzi, Giorgio Gambosi, Roberto Grossi, Carlo Nocentini, Tommaso Papini, Walter Verdese
- Contact: Pierluigi Crescenzi
- URL: http://javamm.sourceforge.net/piluc/software/alvie.html

5.3. Cassis

FUNCTIONAL DESCRIPTION

Algorithm for precisely detecting genomic rearrangement breakpoints.

- Participants: Christian Baudet, Christian Gautier, Claire Lemaitre, Marie-France Sagot, Eric Tannier
- Contact: Christian Baudet (not Inria), Claire Lemaitre (Inria GenScale), Marie-France Sagot (Inria ERABLE)
- URL: http://pbil.univ-lyon1.fr/software/Cassis/

5.4. Cidane

FUNCTIONAL DESCRIPTION

CIDANE is a novel framework for genome-based transcript reconstruction and quantification from RNA-seq reads.

- Participants: Stefan Canzar, Sandra Andreotti, David Weese, Kurt Reinert, Gunnar Klau
- Contact: Stefan Canzar (not Inria)
- URL: http://ccb.jhu.edu/software/cidane/

5.5. Coala

FUNCTIONAL DESCRIPTION

COALA stands for "CO-evolution Assessment by a Likelihood-free Approach". It is thus a likelihood-free method for the co-phylogeny reconstruction problem which is based on an Approximative Bayesian Computation (ABC).

- Participants: Christian Baudet, Pierluigi Crescenzi, Beatrice Donati, Christian Gautier, Catherine Matias, Marie-France Sagot, Blerina Sinaimeri
- Contact: Christian Baudet (not Inria), Marie-France Sagot and Blerina Sinaimeri
- URL: http://coala.gforge.inria.fr/

5.6. CophyTrees

FUNCTIONAL DESCRIPTION

COPHYTREES is a visualiser for host-parasite and gene-species trees evolution.

- Participants: Laurent Bulteau
- Contact: Laurent Bulteau (not Inria), Blerina Sinaimeri (for Inria ERABLE)
- URL: http://eucalypt.gforge.inria.fr/viewer.html

5.7. C3Part & Isofun

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. It is based on the notion of "common connected components" between graphs.

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

5.8. CycADS

FUNCTIONAL DESCRIPTION

Cyc annotation database system.

- Participants: Patrice Baa Puyoule, Hubert Charles, Stefano Colella, Ludovic Cottret, Marie-France Sagot and Augusto Vellozo
- Contact: Hubert Charles
- URL: http://www.cycadsys.org/

5.9. Dinghy

FUNCTIONAL DESCRIPTION

DINGHY is a visualisation program for network pathways of up to 150 reactions.

- Participants: Laurent Bulteau, Alice Julien-Laferrière, Delphine Parrot
- Contact: Laurent Bulteau (not Inria), Alice Julien-Laferrière, Delphine Parrot (not Inria), Marie-France Sagot (for Inria ERABLE)
- URL: http://dinghy.gforge.inria.fr/

5.10. Eucalypt

FUNCTIONAL DESCRIPTION

EUCALYPT stands for "EnUmerator of Co-evolutionary Associations in PoLYnomial-Time delay". It is an algorithm for enumerating all optimal (possibly time-unfeasible) mappings of a parasite tree unto a host tree.

- Participants: Christian Baudet, Pierluigi Crescenzi, Beatrice Donati, Marie-France Sagot, Blerina Sinaimeri
- Contact: Christian Baudet (not Inria), Beatrice Donati (not Inria), and Marie-France Sagot (Inria ERABLE)
- URL: http://eucalypt.gforge.inria.fr/index.html

5.11. Gobbolino & Touché

FUNCTIONAL DESCRIPTION

GOBBOLINO and TOUCHÉ were 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. Biologically, stories correspond to alternative metabolic pathways that may explain some stress that affected the metabolites corresponding to the black nodes by changing their concentration (measured by metabolomics experiments).

- Participants: Vicente Acuña, Etienne Birmelé, Ludovic Cottret, Pierluigi Crescenzi, Fabien Jourdan, Vincent Lacroix, Alberto Marchetti-Spaccamela, Andrea Marino, Paulo Vieira Milreu, Marie-France Sagot, Leen Stougie
- Contact: Paulo Vieira Milreu (not Inria), Marie-France Sagot (Inria ERABLE)
- URL: http://gforge.inria.fr/projects/gobbolino

5.12. HapCol

FUNCTIONAL DESCRIPTION

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

- Participants: Paola Bonizzoni, Riccardo Dondi, Gunnar Klau, Yuri Pirola, Nadia Pisanti, Simone Zaccaria
- Contact: Gunnar Klau, Nadia Pisanti, Paola Bonizzoni (not Inria)
- URL: http://hapcol.algolab.eu/

5.13. KisSNP & DiscoSNP

FUNCTIONAL DESCRIPTION

Algorithm for identifying SNPs without a reference genome by comparing raw reads. KISSNP has now given birth to DISCOSNP in a work involving V. Lacroix from ERABLE and the GenScale Inria Team at Rennes (contact: pierre.peterlongo@inria.fr).

- Participants: Vincent Lacroix, Pierre Peterlongo
- Contact: Pierre Peterlongo (EPI GenScale)
- URL: http://colibread.inria.fr/software/discosnp/

5.14. KisSplice

FUNCTIONAL DESCRIPTION

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

- Participants: Lilia Brinza, Alice Julien-Laferrière, Janice Kielbassa, Vincent Lacroix, Leandro Ishi Soares de Lima, Camille Marchet, Vincent Miele, Gustavo Sacomoto
- Contact: Vincent Lacroix
- URL: http://kissplice.prabi.fr/

5.15. kissDE

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 NGS data pre-processing and gives as output a list of condition specific variants.

- Participants: Clara Benoit-Pilven, Lilia Brinza, Janice Kielbassa, Vincent Lacroix, Camille Marchet and Vincent Miele
- Contact: Vincent Lacroix
- URL: http://kissplice.prabi.fr/tools/kissDE/

5.16. KisSplice2RefTranscriptome

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 RNAseq data only, we obtain a list of condition-specific SNPs stratified by functional impact.

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

5.17. KisSplice2RefGenome

FUNCTIONAL DESCRIPTION

KISSPLICE (see above) identifies variations in RNAseq data, without a reference genome. In many applications however, a reference genome is available. KISSPLICE2REFGENOME enables to facilitate the interpretation of KISSPLICE's results after mapping them to a reference genome.

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

5.18. Lasagne

FUNCTIONAL DESCRIPTION

LASAGNE is a Java application which allows the user to compute distance measures on graphs by making a clever use either of the breadth-first search or of the Dijkstra algorithm. In particular, the current version of LASAGNE can compute the exact value of the diameter of a graph: the graph can be directed or undirected and it can be weighted or unweighted. Moreover, LASAGNE can compute an approximation of the distance distribution of an undirected unweighted graph. These two features are integrated within a graphical user interface along with other features, such as computing the maximum (strongly) connected component of a graph.

- Participants: Pierluigi Crescenzi, Roberto Grossi, Michel Habib, Claudio Imbrenda, Leonardo Lanzi, Andrea Marino
- Contact: Pierluigi Crescenzi
- URL: http://lasagne-unifi.sourceforge.net/

5.19. MeDuSa

FUNCTIONAL DESCRIPTION

MEDUSA (Multi-Draft based Scaffolder) is an algorithm for genome scaffolding. It exploits information obtained from a set of (draft or closed) genomes from related organisms to determine the correct order and orientation of the contigs.

- Participants: Emmanuelle Bosi, Sara Brunetti, Pierluigi Crescenzi, Beatrice Donati, Renato Fani, Marco Fondi, Marco Galardini, Pietro Lió, Marie-France Sagot,
- Contact: Pierluigi Crescenzi, Marco Fondi (not Inria)
- URL: http://combo.dbe.unifi.it/medusa

5.20. MetExplore

FUNCTIONAL DESCRIPTION

Web server to link metabolomic experiments and genome-scale metabolic networks.

- Participants: Michael Barrett, Hubert Charles, Ludovic Cottret, Fabien Jourdan, Marie-France Sagot, Florence Vinson, David Wildridge
- Contact: Fabien Jourdan (not Inria), Marie-France Sagot
- URL: http://metexplore.toulouse.inra.fr/metexplore/

5.21. Migal

FUNCTIONAL DESCRIPTION

Algorithm for comparing RNA structures.

- Participants: Julien Allali and Marie-France Sagot
- Contact: Marie-France Sagot
- URL: http://www-igm.univ-mlv.fr/~allali/logiciels/index.en.php

5.22. Mirinho

FUNCTIONAL DESCRIPTION

Predicts, at a genome-wide scale, microRNA candidates.

- Participants: Christian Gautier, Cyril Fournier, Christine Gaspin, Susan Higashi, Marie-France Sagot
- Contact: Susan Higashi (not Inria), Marie-France Sagot
- URL: http://mirinho.gforge.inria.fr/

5.23. Motus & MotusWEB

FUNCTIONAL DESCRIPTION

Algorithm for searching and inferring coloured motifs in metabolic networks (web-based version - offers different functionalities from the downloadable version).

- Participants: Ludovic Cottret, Fabien Jourdan, Vincent Lacroix, Odile Rogier and Marie-France Sagot
- Contact: Vincent Lacroix
- URL: http://doua.prabi.fr/software/motus and http://pbil.univ-lyon1.fr/software/motus_web/

5.24. MultiPus

FUNCTIONAL DESCRIPTION

MULTIPUS (for MULTIple species for the synthetic Production of Useful biochemical Substances) is an algorithm that, given a microbial consortium given 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: Laurent Bulteau, Alice Julien-Laferrière, Arnaud Mary, Alberto Marchetti-Spaccamela, Delphine Parrot, Marie-France Sagot, Leen Stougie and Susana Vinga
- Contact: Alice Julien-Laferrière, Arnaud Mary, Marie-France Sagot
- URL: http://multipus.gforge.inria.fr/

5.25. PepLine

FUNCTIONAL DESCRIPTION

Pipeline for the high-throughput analysis of proteomic data.

- Participants: Jérôme Garin, Alain Viari
- Contact: Alain Viari
- URL: Available upon request to the contact person

5.26. Pitufo and family

FUNCTIONAL DESCRIPTION

Algorithms to enumerate all minimal sets of precursors of target compounds in a metabolic network.

- Participants: Vicente Acuña Aguayo, Ludovic Cottret, Alberto Marchetti-Spaccamela, Fabio Henrique Viduani Martinez, Paulo Vieira Milreu, Marie-France Sagot, Leen Stougie
- Contact: Paulo Vieira Milreu (not Inria), Marie-France Sagot
- URL: https://sites.google.com/site/pitufosoftware/home

5.27. RepSeek

FUNCTIONAL DESCRIPTION

Finding approximate repeats in large DNA sequences.

- Participants: Guillaume Achaz, Eric Coissac, Alain Viari
- Contact: Guillaume Achaz (not Inria), Alain Viari
- URL: http://wwwabi.snv.jussieu.fr/public/RepSeek/

5.28. Rime

FUNCTIONAL DESCRIPTION

RIME detects long similar fragments occurring at least twice in a set of biological sequences.

- Participants: Maria Federico, Pierre Peterlongo, Nadia Pisanti, Marie-France Sagot
- Contact: Maria Federico (not Inria), Nadia Pisanti, Marie-France Sagot
- URL: https://code.google.com/p/repeat-identification-rime/

5.29. Sasita

FUNCTIONAL DESCRIPTION

SASITA is a software for the exhaustive enumeration of minimal stoichiometrically valid precursor sets in metabolic networks.

- Participants: Vicente Acuña, Ricardo Andrade, Alberto Marchetti-Spaccamela, Marie-France Sagot, Leen Stougie, Martin Wannagat
- Contact: Marie-France Sagot, Ricardo Andrade, Martin Wannagat
- URL: http://sasita.gforge.inria.fr/

5.30. Smile

FUNCTIONAL DESCRIPTION

Motif inference algorithm taking as input a set of biological sequences. A visualiser is currently being developed.

- Participants: Ricardo Andrade (visualiser), Mariana Ferrarini (visualiser), Laurent Marsan, Marie-France Sagot
- Contact: Ricardo Andrade, Marie-France Sagot
- URL: Soon available

5.31. Totoro & Kotoura

FUNCTIONAL DESCRIPTION

We proposed two methods to decipher the reaction changes during a metabolic transient state using measurements of metabolic concentrations. We called these *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.

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.

- Participants: Ricardo Andrade, Laurent Bulteau, Louis Duchemin, Alice Julien-Laferrière, Alberto Marchetti-Spaccamela, Arnaud Mary, Vincent Lacroix, Marie-France Sagot, Leen Stougie, Philippe Veber, Susana Vinga
- Contact: Alice Julien-Laferrière, Arnaud Mary, Ricardo Andrade, Marie-France Sagot
- URL: http://hyperstories.gforge.inria.fr/

5.32. WhatsHap and pWhatsHap

FUNCTIONAL DESCRIPTION

WHATSHAP is a DP approach for haplotype assembly from long reads that works until 20x coverage, solves the minimum error correction problem exactly. PWHATSHAP is a parallelisation of the core dynamic programming algorithm of WHATSHAP done by M. Aldinucci, A. Bracciali, T. Marschall, M. Patterson, N. Pisanti, and M. Torquati.

- Participants: Gunnar Klau, Tobias Marschall, Murray Patterson, Nadia Pisanti, Alexander Schönhuth, Leen Stougie, Leo van Iersel
- Contact: Alexander Schönhuth (not Inria), Gunnar Klau, Nadia Pisanti
- URL: https://bitbucket.org/whatshap/whatshap and https://bitbucket.org/whatshap/whatshap/branch/parallel

6. New Results

6.1. General comments

We present in this section the main results obtained in 2016. Some were already in preparation or submitted at the end of 2015. This will be indicated whenever it is the case.

We tried to organise the results following the five main axes of research of the team. Clearly, in some cases, a result obtained overlaps more than one axis. 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 rumour spreading (by Pierluigi Crescenzi) [9] or on general network analysis (by Pierluigi Crescenzi or Roberto Grossi) [31], [36], [40], [39], [37], [38], [10], [42] could in the future become relevant for life sciences (biology or ecology). In the other direction, algorithmic ideas that were developed in the context of a problem in life sciences could prove useful for solving more general problems (possibly with other applications). This was the case of some of the ideas explored in previous years to deal with de Bruijn graphs in the context of NGS analysis that led to the team fruitfully collaborating with a group of researchers at the ETH in Switzerland on a problem related to transport systems [34].

Below however, we preferred to only indicate the theoretical results related to problems closely resembling questions that have already been addressed by us in computational biology. Notice that such CS results 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 are not mentioned either 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

RNA-seq NGS algorithms and data analysis
SNPs (Single Nucleotide Polymorphisms) are genetic markers whose precise identification is a prerequisite for association studies. Methods to identify them are currently well developed for model species, but rely on the availability of a (good) reference genome, and therefore cannot be applied to non-model species. They are also mostly tailored for whole genome (re-)sequencing experiments, whereas in many cases, transcriptome sequencing can be used as a cheaper alternative which already enables to identify SNPs located in transcribed regions. In a paper accepted this year [18], we proposed the use of a previously developed method, KISSPLICE, that identifies, quantifies and annotates SNPs without any reference genome, using RNA-seq data only. 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 recall of our method and discussed them with respect to other methods 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. KISSPLICE can be used for any species to annotate SNPs and predict their impact on the protein sequence. We further enable to test for the association of the identified SNPs with a phenotype of interest.

We participated also in two other works, one computational and the other biological, on alternative splicing in Human.

The first is associated to the ANR Colib'read project in which we were one of the partners. A Colib'read Galaxy tools suite was developed that should enable a broad range of life science researchers to analyse raw NGS data, allows the maximum biological information to be retained in the data, and uses a very low memory footprint [17]. The algorithms implemented in the tools are based on the use of a de Bruijn graph and of a bloom filter. The analyses can be performed in a few hours, using small amounts of memory. Applications using real data further demonstrate the good accuracy of these tools compared to classical approaches.

KISSPLICE was also used in the context of myotonic dystrophy (DM), which is caused by the expression of mutant RNAs containing expanded CUG repeats that sequester muscleblind-like (MBNL) proteins, leading to alternative splicing changes. Cardiac alterations, characterised by conduction delays and arrhythmia, are the second most common cause of death in DM. Using RNA sequencing, the authors of [14] identified novel splicing alterations in DM heart samples, including a switch from adult exon 6B towards fetal exon 6A in the cardiac sodium channel, SCN5A. They found that MBNL1 regulates alternative splicing of SCN5A mRNA and that the splicing variant of SCN5A produced in DM presents a reduced excitability compared to the control adult isoform. Importantly, reproducing splicing alteration of Scn5a in mice is sufficient to promote heart arrhythmia and cardiac-conduction delay, two predominant features of myotonic dystrophy. Misregulation of the alternative splicing of SCN5A may therefore contribute to a subset of the cardiac dysfunctions observed in myotonic dystrophy.

We introduced CIDANE, a novel framework for genome-based transcript reconstruction and quantification from RNA-seq reads [8]. CIDANE assembles transcripts efficiently with significantly higher sensitivity and precision than existing tools. Its algorithmic core not only reconstructs transcripts *ab initio*, but also allows the use of the growing annotation of known splice sites, transcription start and end sites, or full-length transcripts, which are available for most model organisms. CIDANE supports the integrated analysis of RNA-seq and additional gene-boundary data and recovers splice junctions that are invisible to other methods.

Landscape of somatic mutations in breast cancer whole-genome sequences
In the context of the International Cancer Genome Consortium (ICGC), we conducted a whole-genome, exome, RNASeq and methylome characterisation of 560 breast cancers. The results were published this year in three main papers.

The first one describes the general landscape of somatic mutations and rearrangements in all subtypes of breast cancers [21]. This allowed to extend our current repertoire of probable breast cancer drivers to 93 genes. The mutational signature analysis was extended to genome rearrangements as well and revealed six typical rearrangement signatures. Three of them, characterised by tandem duplications or deletions, appear associated with defective homologous- recombination-based DNA repair (BRCA1/2). This analysis highlighted the repertoire of cancer genes and mutational processes operating in human, and represented a progress towards obtaining a comprehensive account of the somatic genetic basis of breast cancer.

This first analysis was then used to link known and novel drivers and mutational signatures to gene expression (transcriptome) of 266 cases [28]. One important and still debated question is to know to what extend somatic aberrations could trigger an immune-response. Our data suggested that substitutions of a particular type could be more effective in doing so than others.

Finally, in the context of ICGC, France was in charge of the analysis of a clinically specific subgroup of breast cancers, called HER2-positive, characterised by the HER2/ERBB2 amplification and over-expression. This is a subgroup for which several efficient targeted therapies (trastuzumab) are now available. However, resistance to treatment has been observed, revealing the underlying diversity of these cancers. An in-depth genomic and transcriptomic characterisation of 64 HER2-positive breast tumour was carried out. We delineated four subgroups, based on the expression data, each of them with distinctive genomic features in terms of somatic mutations, copy-number changes or structural variations [12]. The results suggested that, despite being clinically delineated by a specific gene amplification, HER2-positive tumours actually melt into the luminal-basal breast cancer spectrum rather, probably following their "cell-of-origin" fate and suggesting that the ERBB2 amplification is an embedded event in the natural history of these tumours. Finally, WGS data allowed us to gain more information about the amplification process itself and brought some indications about how (and maybe when) it arose. Whole genome paired-end sequencing provides two important experimental clues to this purpose: a) high dynamics and resolution analysis of copy numbers, and b) ability to pinpoint large scale structural rearrangements by using clipping and abnormal mapping of read pairs. We could show that, in several cases, the observed sequence of copy numbers as well as the orientation of clipped reads was consistent with a breakage-fusion-bridge folding mechanism (BFB). However, the observation of long distance and inter-chromosomal rearrangements further showed that the amplification is a complex event (or sequence of events), likely involving several amplicons on the same or different chromosomes and several intertwined mechanisms. Indeed one of the features of HER2+ tumours is the ubiquitous presence of firestorms, corresponding to multiple closely spaced amplicons on highly rearranged chromosomal arms. It is therefore tempting to combine two mechanisms to explain the complex amplification patterns observed: chromothripsis, which will generate a mosaic of fragments (but no amplification per se), followed by a BFB amplification of chromosomal arm(s). This work was done at the “Plateforme Bioinformatique Gilles Thomas” located at Centre Léon Bérard (Lyon).

Sequence comparison

Sequence comparison is a fundamental step in many important computational biology tasks, in particular the reconstruction of genomes, a first key step before being able to identify the molecular elements present in them.

Traditional algorithms for measuring approximation in sequence comparison are based on the notions of distance or similarity, and are generally computed through sequence alignment techniques. As circular molecular structures are a common phenomenon in nature, a caveat of the adaptation of alignment techniques for circular sequence comparison is that they are computationally expensive, requiring from super-quadratic to cubic time in the length of the sequences. We introduced a new distance measure based on q-grams, and showed how it can be applied effectively and computed efficiently for circular sequence comparison [15]. Experimental results, using real DNA, RNA, and protein sequences as well as synthetic data, demonstrated orders-of-magnitude superiority of our approach in terms of efficiency, while maintaining an accuracy very competitive in relation to the state of the art.

Data structures for text indexing and string (sequence) comparison

Suffix trees are important data structures for text indexing and string algorithms. For any given string w of length $n = |w|$, a suffix tree for w takes $O(n)$ vertices and links. It is often presented as a compacted version of a suffix trie for w, where the latter is the trie (or digital search tree) built on the suffixes of w. The compaction process replaces each maximal chain of unary vertices with a single arc. For this, the suffix tree requires that the labels of its arcs are substrings encoded as pointers to w (or equivalent information). On the contrary, the arcs of the suffix trie are labeled by single symbols but there can be $\Theta(n^2)$ vertices and links for suffix tries in the worst case because of their unary vertices. It was an interesting question if the suffix trie can be stored using $O(n)$ vertices. We addressed it and thus presented the linear-size suffix trie, which guarantees $O(n)$ vertices [11]. We used a new technique for reducing the number of unary vertices to $O(n)$, that stems from some results on anti-dictionaries. For instance, by using the linear-size suffix trie, we are able to check whether a pattern p of length $m = |p|$ occurs in w in $O(m \log |\Sigma|)$ time and we can find the longest common substring of two strings w_1 and w_2 in $O((|w_1| + |w_2|) \log |\Sigma||)$ time for an alphabet $\Sigma|$.

Haplotype assembly
Haplotype assembly is the computational problem of reconstructing haplotypes in diploid organisms and is of fundamental importance for characterising the effects of single-nucleotide polymorphisms on the expression of phenotypic traits. Haplotype assembly highly benefits from the advent of "future-generation" sequencing technologies and their capability to produce long reads at increasing coverage. Existing methods are not able to deal with such data in a fully satisfactory way, either because accuracy or performances degrade as read length and sequencing coverage increase or because they are based on restrictive assumptions.

By exploiting a feature of future-generation technologies – the uniform distribution of sequencing errors – we designed an exact algorithm, called HAPCOL, that is exponential in the maximum number of corrections for each single-nucleotide polymorphism position and that minimises the overall error-correction score [22]. We performed an experimental analysis, comparing HAPCOL to the current state-of-the-art combinatorial methods both on real and simulated data. On a standard benchmark of real data, we showed that HAPCOL is competitive with state-of-the-art methods, improving the accuracy and the number of phased positions. Furthermore, experiments on realistically simulated datasets revealed that HAPCOL requires significantly less computing resources, especially memory. Thanks to its computational efficiency, HAPCOL can overcome the limits of previous approaches, allowing to phase datasets with higher coverage and without the traditional all-heterozygous assumption.

HAPCOL is based on MEC (Minimum error correction) which is computationally hard to solve. However, some approximation-based or fixed-parameter approaches have been proved capable of obtaining accurate results on real data. In another work [5], we then attempted to expand the current characterisation of the computational complexity of MEC from such approximation and fixed-parameter tractability points of view. We showed that MEC is not approximable within a constant factor, whereas it is approximable within a logarithmic factor in the size of the input. Furthermore, we answered open questions on the fixed-parameter tractability for parameters of classical or practical interest: the total number of corrections and the fragment length. In addition, we presented a direct 2-approximation algorithm for a variant of the problem that has also been applied in the framework of clustering data. Finally, since polyploid genomes, such as those of plants and fishes, are composed of more than two copies of the chromosomes, we introduced a novel formulation of MEC, namely the k-ploid MEC problem, that extends the traditional problem to deal with polyploid genomes. We showed that the novel formulation remains both computationally hard and hard to approximate. Nonetheless, from the parameterised point of view, we proved that the problem is tractable for parameters of practical interest such as the number of haplotypes and the coverage, or the number of haplotypes and the fragment length.

6.3. Inferring and analysing the networks of molecular elements

Metamodules in transcriptomic analysis

The human microbiome plays a key role in health and disease. Thanks to comparative metatranscriptomics, the cellular functions that are deregulated by the microbiome in disease can now be computationally explored. Unlike gene-centric approaches, pathway-based methods provide a systemic view of such functions; however, they typically consider each pathway in isolation and in its entirety. They can therefore overlook the key differences that (i) span multiple pathways, (ii) contain bidirectionally deregulated components, (iii) are confined to a pathway region. To capture these properties, computational methods that reach beyond the scope of predefined pathways are needed.

By integrating an existing module discovery algorithm into comparative metatranscriptomic analysis, we developed METAMODULES, a novel computational framework for automated identification of the key functional differences between health- and disease-associated communities [20]. Using this framework, we recovered significantly deregulated subnetworks that were indeed recognised to be involved in two well-studied, microbiome-mediated oral diseases, such as butanoate production in periodontal disease and metabolism of sugar alcohols in dental caries. More importantly, our results indicated that our method can be used for hypothesis generation based on automated discovery of novel, disease-related functional subnetworks, which would otherwise require extensive and laborious manual assessment.

Metabolic environmental dialog
What an organism needs at least from its environment to produce a set of metabolites, *e.g.* target(s) of interest and/or biomass, has been called a minimal precursor set. Early approaches to enumerate all minimal precursor sets took into account only the topology of the metabolic network (topological precursor sets). Due to cycles and the stoichiometric values of the reactions, it is often not possible to produce the target(s) from a topological precursor set in the sense that there is no feasible flux. Although considering the stoichiometry makes the problem harder, it enables to obtain biologically reasonable precursor sets that we call stoichiometric. Recently a method to enumerate all minimal stoichiometric precursor sets was proposed in the literature. The relationship between topological and stoichiometric precursor sets had however not yet been studied.

Such relationship was explored in a recently accepted paper [3]. In there, we also presented two algorithms that enumerate all minimal stoichiometric precursor sets. The first one is of theoretical interest only and is based on the above mentioned relationship. The second approach solves a series of mixed integer linear programming (MILP) problems. We compared the computed minimal precursor sets to experimentally obtained growth media of several *Escherichia coli* strains using genome-scale metabolic networks.

The results showed that the second approach, called SASITA, efficiently enumerates minimal precursor sets taking stoichiometry into account, and allows for broad *in silico* studies of strains or species interactions that may help to understand *e.g.* pathotype and niche-specific metabolic capabilities.

This work was also part of the PhD of Martin Wannagat, defended in June 2016 [2].

Metabolic hyperstories
In the context of a PhD in the team (whose defence took place in Dec 8, 2016) [1] and using metabolomics data, we focused on inferring the metabolic behaviour of an organism when it is subjected to a change in conditions. In this case, one can infer the reactions impacted when the changes are controlled and known (*e.g.* exposition to toxic compounds, changes in culture conditions). However, understanding how the metabolism of an organism changes of equilibrium is also of interest to infer the processes related for example to a transition between a commensal or beneficial bacterium to a pathogenic one. This question led to two different methods. The first, that we called TOTORO (for TOpological analysis of Transient metabOlic RespOnse), is based on the topology of metabolic networks to infer the reactions involved in a transient state, when an organism goes from one state of growth to another. We proposed a novel definition using the directed hypergraph representation and discussed its application on a dataset of Yeast exposed to cadmium. We showed that this method suggests more complete solutions of the reactions impacted during the metabolic shift. The second method, called KOTOURA (for Kantitative analysis Of Transient metabOlic and regUlatory Response And control), offers a constraint-based perspective in a more quantitative approach. We applied it to a simulated dataset and we are currently trying to infer the possible quantitative responses to mutations with a more complete kinetic model. An image previously used is that condition-specific models provide a snapshot of the metabolism of

an organism, whether it is at the evolutionary-time scale or at the scale of a specific environmental condition describing a physiological process. Our idea here is thus to infer the transitions between those snapshots.

Besides the PhD manuscript, two papers are in preparation to present this work. They should be submitted in early 2017. A prototype for the two methods is available at: http://hyperstories.gforge.inria.fr/.

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

Robustness of the parsimonious reconciliation method in cophylogeny

The currently most used method in cophylogenetic studies is the so-called *phylogenetic tree reconciliation*. In this model, we are given the phylogenetic tree of the hosts H, the one of the symbionts S, and a mapping ϕ from the leaves of S to the leaves of H indicating the known symbiotic relationships among present-day organisms. The common evolutionary history of the hosts and of their symbionts is then explained through a number of macroevolutionary events (four in general). A reconciliation is then a function λ which is an extension of the mapping ϕ between leaves to a mapping that includes all internal nodes and that can be constructed using the different types of events considered. An optimal reconciliation is usually defined in a parsimonious way: a cost is associated to each event and a solution of minimum total cost is searched for.

An important issue in this model is that it makes strong assumptions on the input data which may not be verified in practice. We examine two cases where this situation happens. The first is related to a limitation in the currently available methods for tree reconciliation where the association ϕ of the leaves is for now required to be a function. This is not realistic as a single symbiont species can infect more than one host. For each present-day symbiont involved in a multiple association, one is currently forced to choose a single one. The second case addresses a different type of problem related to the phylogenetic trees of hosts and symbionts. These indeed are assumed to be correct, which may not be the case. In this work, we addressed the problem of correctly rooting a phylogenetic tree.

We thus explored the robustness of the parsimonious tree reconciliation method under "editing" (multiple associations) or "small perturbations" of the input (rooting problem) [29].

An extended version of this paper has been submitted to *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Insights on the virulence of swine respiratory tract mycoplasmas through genome-scale metabolic modelling

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* was shown to be virtually asymptomatic, *M. hyopneumoniae* is known to be the causative agent of enzootic pneumonia and *M. hyorhinis* to be present in cases of pneumonia, polyserositis and arthritis. Nonetheless, the elevated genomic resemblance among these three mycoplasmas combined with their different levels of pathogenicity is an indication that they have unknown mechanisms of virulence and differential expression. In 2015, we performed whole-genome metabolic network reconstructions for these three mycoplasmas. The results obtained were then submitted for publication to *BMC Genomics*. The paper has since been published [13].

Maximal chain subgraphs and covers of bipartite graphs motivated by analysis of cytoplasmic incompatibility

In a previous work of the team (Nor *et al. American Naturalist*, 182(1):15-24, 2013; Nor*et al. Information and Computation*, 213:23-32, 2012), we showed that a minimum chain subgraph cover of a given bipartite graph provides a good model for identifying the minimum genetic architecture enabling to explain one type of manipulation, called *cytoplasmic incompatibility*, by some parasite bacteria on their hosts. This phenomenon results in the death of embryos produced in crosses between males carrying the infection and uninfected females. The observed cytoplasmic compatibility relationships, can then be represented by a bipartite graph with males and females in different classes. Moreover, as different minimum (resp. minimal) covers may correspond to solutions that differ in terms of their biological interpretation, the capacity to enumerate all such minimal chain covers becomes crucial.

We recently addressed three related problems that bear some interest for the above problem besides raising interesting theoretical questions [35]. One is the enumeration of all the maximal *edge induced* chain subgraphs of a bipartite graph, for which we provided a polynomial delay algorithm. We gave bounds on the number of maximal chain subgraphs for a bipartite graph and used them to establish the input-sensitive complexity of the enumeration problem. The second problem we treated was the one of finding the minimum number of chain subgraphs needed to cover all the edges a bipartite graph. For this, we provided an exact exponential algorithm with a non trivial complexity. Finally, we approached the problem of enumerating all minimal chain subgraph covers of a bipartite graph and showed that it can be solved in quasi-polynomial time.

An extended version of the conference paper has been submitted to a journal in December 2016.

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

On the Complexity of Quadratic-Time Solvable Problems

Quadratic-time solvable problems may be classified into two classes: problems that are solvable in *truly subquadratic* time (that is, in time $(n^{2-\epsilon})$ for some $\epsilon > 0$) and problems that are not, unless the well known Strong Exponential Time Hypothesis (in short, SETH) is false. We proved that some quadratic-time solvable problems are indeed easier than expected [6]. We provided an algorithm that computes the transitive closure of a directed graph in time $(mn^{\frac{\omega+1}{4}})$, where m denotes the number of edges in the transitive closure and ω is the exponent for matrix multiplication. As a side effect of our analysis, we were able to prove that our algorithm runs in time $(n^{\frac{5}{3}})$ if the transitive closure of the graph is sparse. The same time bounds hold if we want to check whether a graph is transitive, by replacing m with the number of edges in the graph itself. As far as we know, this gives us the fastest algorithm for checking whether a sparse graph is transitive. Finally, we applied our algorithm to the comparability graph recognition problem (which dates back to 1941): also in this case, we obtained the first truly subquadratic algorithm. We then dealt with some hardness results. In particular, we started from an artificial quadratic-time solvable variation of the k-SAT problem and constructed a graph of Karp reductions, proving that a truly subquadratic-time algorithm for any of the problems in the graph falsifies SETH. More specifically, the analysed problems were the following: computing the subset graph, finding dominating sets, computing the betweenness centrality of a vertex, computing the minimum closeness centrality, and computing the hyperbolicity of a pair of vertices. We were also able to include in our framework three proofs that had already appeared in the literature, concerning the problems of distinguishing between split graphs of diameter 2 and diameter 3, of solving the local alignment of strings, and of finding two orthogonal binary vectors inside a collection.

Enumeration of solutions produced by closure operations

In enumeration problems, we are interested in listing a set of elements, which can be of exponential cardinality in the size of the input. The complexity of such problems is thus measured in terms of their input and output sizes. An enumeration algorithm with a complexity polynomial in both sizes is called output polynomial or total polynomial time. Another more precise notion of complexity is related to the *delay*, that is to the time between the production of two consecutive solutions. We are especially interested in problems solvable with a delay polynomial in the input size. These are considered as the tractable problems in enumeration complexity.

We addressed the problem of generating all elements obtained by the saturation of an initial set by some operations [41]. More precisely, we proved that we can generate the closure by polymorphisms of a boolean relation with a polynomial delay. This implies for instance that we can compute with polynomial delay the closure of a family of sets by any set of "set operations" (*e.g.* union, intersection, difference, symmetric difference, etc.). To do so, we proved that for any set of operations $\mathcal{F}$, one can decide in polynomial time whether an element belongs to the closure by $\mathcal{F}$ of a family of sets. When the relation is over a domain larger than two elements, our generic enumeration method fails for some cases since the associated decision problem is NP-hard, and we then provide an alternative algorithm.

6.6. Going towards control

Combinatorial approach for microbial consortia synthetic design
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. This is the problem we addressed in a paper accepted this year [16].

We thus 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 the interaction among the species in the consortium could improve the production line. In mathematical terms, given a weighted directed hypergraph $\mathcal{H}$, the problem is to enumerate all directed sub-hypergraphs whose sets of vertices and of hyperarcs are included in those of $\mathcal{H}$, enable to produce the set of targets of interest from a subset of the sources of $\mathcal{H}$, and are of minimum weight. We called this the Directed Steiner Hypertree (DSH) problem.

We showed that the main issue in terms of the complexity of the problem comes from the hyperarcs with multiple source vertices (we called those the *tentacular hyperarcs*), not from the possible multiplicity of the target vertices. This is not the only issue though, and we thus further demonstrated that even when there is only one target that needs to be reached, the problem remains NP-hard. When both parameters, number of tentacular hyperarcs and of targets, are fixed, the problem becomes tractable. We then explored two methods for addressing it. One is a dynamic programming approach, and the other logic programming using ASP (Answer Set Programming). The second was more efficient for now, and the software MULTIPUS is thus based on it.

As initial validations of the model and of the method, we applied it to two case-studies taken from the literature.

This work was also part of the PhD of Alice Julien-Laferrière defended in December 2016 [1].

7. Partnerships and Cooperations

7.1. Regional Initiatives

7.1.1. ICEbErg

- Title: Integrating Co-phylogeny in the analysis of Ecological nEtworks
- Coordinator: B. Sinaimeri and S. Dray
- ERABLE participant(s): B. Sinaimeri
- Type: Inter-departmental project funded by the LBBE (Sept 2016 - Sept 2017)
- Web page: Not available

7.2. National Initiatives

7.2.1. ANR

7.2.1.1. ABS4NGS

- Title: Solutions Algorithmiques, Bioinformatiques et Logicielles pour le Séquençage Haut Débit
- Coordinator: E. Barillot

- ERABLE participant(s): V. Lacroix
- Type: ANR (2012-2016)
- Web page: https://sites.google.com/site/abs4ngs/

7.2.1.2. Colib'read

- Title: Methods for efficient detection and visualization of biological information from non assembled NGS data
- Coordinator: P. Peterlongo
- ERABLE participant(s): V. Lacroix, L. I. S. de Lima, A. Julien-Laferrière, H. Lopez-Maestre, C. Marchet, G. Sacomoto, M.-F. Sagot, B. Sinaimeri
- Type: ANR (2013-2016)
- Web page: http://colibread.inria.fr/

7.2.1.3. ExHyb

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

7.2.1.4. 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/

7.2.1.5. 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

7.2.2. Others

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

7.2.2.1. Amanda

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

7.2.2.2. Effets de l'environnement sur la stabilité des éléments transposables

- Title: Effets de l'environnement sur la stabilité des éléments transposables
- Coordinator: C. Vieira
- ERABLE participant(s): C. Vieira
- Type: Fondation pour la Recherche Médicale (FRM) (2014-2016)
- Web page: Not available

7.2.2.3. QualiBioConsensus

- Title: Qualité des classements consensuels de données biologiques massives
- Coordinator: S. Cohen-Boulakia
- ERABLE participant(s): L. Bulteau (external collaborator of ERABLE)
- Type: Défi Mastodons (2016)
- Web page: Not available

7.3. European Initiatives

7.3.1. FP7 & H2020 Projects

7.3.1.1. BacHBerry

Title: BACterial Hosts for production of Bioactive phenolics from bERRY fruits
Duration: November 2013 - October 2016
Coordinator: Jochen Förster, DTU Danemark
ERABLE participant(s): R. Andrade, L. Bulteau, A. Julien-Laferrière, V. Lacroix, A. Marchetti-Spaccamela, A. Mary, D. Parrot, M.-F. Sagot, L. Stougie, A. Viari, M. Wannagat
Type: FP7 - KBBE
Web page: http://www.bachberry.eu/

7.3.1.2. 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: http://www.microwine.eu/

7.3.2. Collaborations in European Programs, Except FP7 & H2020

7.3.2.1. Combinatorics of co-evolution

- Title: The combinatorics of co-evolution
- Duration: 2015 - 2017
- Coordinator: Katharina Huber, University of Warwick, UK
- ERABLE participant(s): M.-F. Sagot, B. Sinaimeri
- Type: The Royal Society
- Web page: not available

7.3.3. 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.

7.4. International Initiatives

7.4.1. 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 are currently hosting Alex di Genova in ERABLE as a PhD sandwich student (for 18 to 24 months). Alex is co-supervised by Alejandro Maass and by Eric Goles from the University Adolfo Ibañez, Santiago, Chile.

7.4.2. Inria Associate Teams Not Involved in an Inria International Labs

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/

7.4.3. 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 coordinates another project with Brazil. This is a CAPES-COFECUB project titled: "Multidisciplinary Approach to the Study of the Biodiversity, Interactions and Metabolism of the Microbial Ecosystem of Swines", and its acronym MICO. The coordinators are M.-F. Sagot (France) and A. T. Vasconcelos (LNCC, Brazil) with also the participation of Arnaldo Zaha (Federal University of Rio Grande do Sul, Brazil). The project started in 2013 for 2 years, and was renewed for 2 more years starting from 2015. The main objective of this project is to experimentally and mathematically explore the biodiversity of the bacterial organisms living in the respiratory tract of swines, many of which are pathogenic. This project is strongly linked to the LIA LIRIO. More information on it may be found at this address: http://team.inria.fr/erable/en/cnrs-lia-laboratoire-international-associe-lirio/associated-projects/#CAPES-COFECUB_Microbial_Ecosystem_of_Swines.

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 preliminary web page for MAIA is available at this address: http://team.inria.fr/erable/en/projects/maia/.

Finally, we would like to mention the participation of one member of ERABLE (Alain Viari) in the Breast Cancer French Working Group of the International Cancer Genome Consortium (ICGC, https://icgc.org) led by the Institut National du Cancer (INCa, http://www.e-cancer.fr/Professionnels-de-la-recherche/Innovations/Les-progres-de-la-genomique/ICGC-France). This project was initiated by Pr. Gilles Thomas who passed away in 2014. Alain took the head of the bioinformatics platform located at the Centre Léon Bérard. The project aims at the genomic characterisation of 75 HER2-amplified breast cancers by using high-throughput sequencing (whole genome of paired tumour/normal samples and RNAseq of tumour samples). One of the scientific goals is to decipher whether the HER2/ERBB2 amplification is a driver or a passenger event in the course of tumour development.

7.5. International Research Visitors

7.5.1. Visits of International Scientists

In 2016, ERABLE greeted the following International scientists:

- In France: Katharina Huber and Vincent Moulton (University of Warwick, UK), Giuseppe Italiano (Tor Vergata University of Rome, Italy, various visits), Ana Rute Neves and Thomas Janzen (ChR Hansen, Oslo, Danemark), two members of the LIA LIRIO (Arnaldo Zaha from the Federal University of Rio Grande do Sul, and Ana Tereza Vasconcelos from the LNCC, both in Brazil), Susana Vinga and various members of her team (IDMEC-IST Portugal), Tiziana Calamoneri (Sapienza University of Rome).
- In Italy: Costas Iliopoulos and Solon Pissis (King's College, London, UK).

7.5.2. Internships

In 2016, ERABLE greeted the following internship students:

- In France: Audric Cologne, Master 2 (6 months); Irene Ziska, Master Free University Berlin (2 months), Louis Duchemin Master 1 (5 months).

7.5.3. Visits to International Teams

7.5.3.1. Visits

In 2016, members of ERABLE visited the following International teams:

- In France: Giuseppe Italiano (Tor Vergata University of Rome), visit to members of the LIA LIRIO at the LNCC in Brazil, visit to the Department of Computer Science of the University of São Paulo and to members of the TecSinapse company in Brazil, Tiziana Calamoneri (La Sapienza University of Rome), Susana Vinga and members of her team (IDMEC-IST Portugal), Rafaelle Giancarlo (Palermo University, Italy).
- In Italy: Costas Iliopoulos (King's College, London, UK), Luís Russo (INESC-IST, Lisbon, Portugal), Paola Bonizzoni (Milan-Bicocca, Italy), Rafaelle Giancarlo (Palermo University, Italy).

7.5.3.2. Research stays abroad

Gunnar Klau spent 9 months starting from November 2015 at the Center for Computational Molecular Biology at Brown University, USA, visiting notably Benjamin Raphael, Director of the Center.

8. Dissemination

8.1. Promoting Scientific Activities

8.1.1. Scientific events organisation

8.1.1.1. General chair, scientific chair

- Alberto Marchetti-Spaccamela is member of the Steering committee of WG, Workshop on Graph Theoretic Concepts in Computer Science, and of ATMOS, Workshop on Algorithmic Approaches for Transportation Modeling, Optimization, and Systems.
- Marie-France Sagot was from 2010 to 2016 member and from 2014 to 2016 Chair of the Steering Committee of the International Conference *LATIN* (http://www.latintcs.org/). She is member of the Steering Committee of the *European Conference on Computational Biology* (*ECCB*) since 2002 and of the International Symposium on Bioinformatics Research and Applications (ISBRA) since 2008.

8.1.1.2. Member of the organising committees

- Leen Stougie was co-organiser, together with Neil Olver and René Sitters of the 8th Workshop on Flexible Network Design, July 4-8, 2016, Vrije Universiteit, Amsterdam, The Netherlands.

8.1.2. Scientific events selection

8.1.2.1. Member of the conference program committee

- Laurent Bulteau (external collaborator of ERABLE) was a member of the program committee for the following international conferences in 2016: 41st International Symposium on Mathematical Foundations of Computer Science (MFCS 2016) and 11th International Conference on Algorithmic Aspects of Information and Management (AAIM).
- Pierluigi Crescenzi was a member of the program committee for the following international conferences in 2016: 17th Italian Conference on Theoretical Computer Science (ICTCS), 31st IEEE International Parallel & Distributed Processing Symposium (IPDPS).
- Roberto Grossi was a member of the program committee for the following international conferences in 2016: 12th Latin American Theoretical Informatics Symposium (LATIN), 27th International Workshop on Combinatorial Algorithms (IWOCA), 8th International Conference on Fun with Algorithms (FUN).
- Alberto Marchetti-Spaccamela was a member of the program committee for the following international conference in 2016: 15th International Symposium on Experimental Algorithms (SEA).
- Nadia Pisanti was a member of the program committee for the following international conferences in 2016: 10th International Workshop on Algorithms and Computation (WALCOM), 5th International Symposium on Network Enabled Health Informatics, Biomedicine and Bioinformatics (Hi BI BI), 5th RECOMB Satellite Workshop on Computational Cancer Biology (RECOMB-CCB), 16th Workshop on Algorithms in BIoinformatics (WABI), 6th RECOMB Satellite Workshop on Massively Parallel Sequencing (RECOMB-Seq), 12th International Symposium on Bioinformatics Research and Applications (ISBRA), 7th International Conference on Information Technology on Bio- and Medical-Informatics (ITBAM), 6th IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS).
- Marie-France Sagot was a member of the program committee for the following international conferences in 2016: Intelligent Systems for Molecular Biology (ISMB), Prague Stringology Conference 2016, 11th International Conference on Algorithmic Aspects of Information and Management (AAIM), 20th Annual International Conference on Research in Computational Molecular Biology (RECOMB), 14th RECOMB Satellite Workshop on Comparative Genomics (RECOMB-CG), 16th Workshop of Algorithms in Bioinformatics (WABI).
- Leen Stougie was a member of the program committee of the following conferences in 2016: 18th Conference on Integer Programming and Combinatorial Optimization (IPCO), Workshop on Approximation and Online Algorithms (WAOA).

8.1.2.2. Reviewer

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

8.1.3. Journal

8.1.3.1. Member of the editorial board

- 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 *RAIRO – Theoretical Informatics and Applications – Informatique Théorique et Applications*.
- Alberto Marchetti-Spaccamela is member of the Editorial Board of *Theoretical Computer Science* and *Transactions on Algorithms Engineering*.
- Nadia Pisanti is since 2012 member of Editorial Board of *International Journal of Computer Science and Application (IJCSA)*.
- Marie-France Sagot is member of the Editorial Board of *Lecture Notes in Bioinformatics* (subseries of *Lectures Notes in Computer Science*), *Journal of Discrete Algorithms*, *BMC Bioinformatics*, and *BMC Algorithms for Molecular Biology*.
- Leen Stougie is member of the Editorial Board of *Transactions on Algorithms Engineering* since 2010, *Surveys in Operations Research and Management Science* since 2011, and *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*.

8.1.3.2. Reviewer for Journals

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*.

8.1.4. Invited talks

Clara Benoit-Pilven gave a lecture (Workshop Colib'read, November 7-8, 2016).

Roberto Grossi gave an invited talk (Eleventh International Conference on Algorithmic Aspects in Information and Management (AAIM), Bergamo, Italy, July 18-20, 2016).

Gunnar Klau gave an invited talk (Simon's Institute for the Theory of Computing, Workshop on Network Biology, April 11-15, 2016).

Vincent Lacroix gave a lecture+demonstration (Workshop Colib'read, November 7-8, 2016). He also made an invited presentation in the context of a CNRS training meeting (Formation Bioinformatique pour les NGS Montpellier, March 24, 2016)

Hélène Lopez-Maestre gave a lecture (Workshop Colib'read, November 7-8, 2016).

Nadia Pisanti gave two invited talks (Data Driven Innovation Open Summit, Rome, May 20-21, 2016; Conference Mathematical Foundations in Bioinformatics (MatBio), London, UK, July 20, 2016).

Marie-France Sagot gave two invited talks (German Conference on Bioinformatics (GCB), Berlin, Germany, Sept 12-15, 2016; First Workshop on Enumeration Problems and Applications (WEPA), Aubière, France, Nov 21-22, 2016).

Blerina Sinaimeri gave a talk (University of Palermo, Sept 27, 2016).

Leen Stougie gave two lectures (Graduate School on Methods for Discrete Structures, Free University Berlin, 25 April 2016; 7th Cargese Workshop on Combinatorial Optimization, October 9-14, 2016).

Laura Urbini gave a lecture (Warwick University, Computer Science Department, November 10, 2016).

8.1.5. Leadership within the scientific community

Alberto Marchetti-Spaccamela is Member of the Council of EATCS, the European Association for Theoretical Computer Science.

Leen Stougie was Chairman of the Dutch Network on the Mathematics of Operations Research (Landelijk Netwerk Mathematische Besliskunde (LNMB)) from February 2011 to January 2016. From February 2016, he is Member of the general board of the LNMB. He is also Chairman Program Committee Econometrics and OR, VU Amsterdam and Member of the Board of the research school ABRI-VU, Amsterdam.

Cristina Vieira is director of the GDRE "Comparative genomics" since the latter was renewed in 2010.

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

8.1.6. 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. She was also a Panel Member for the ERC.

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

8.1.7. Research administration

Hubert Charles is director of the Biosciences Department of the Insa-Lyon.

Alberto Marchetti-Spaccamela is Director of the Department of Computer, Control, and Management Engineering Antonio Ruberti at Sapienza University of Rome, Italy.

Nadia Pisanti is since 2013 member of the Board of the Regional PhD School of Computer Science at the University of Pisa, Italy.

Alain Viari is since 2012 Deputy Scientific Director at Inria responsible for the domain "Digital Health, Biology and Earth". He thus represents Inria at several national instances related to Life Sciences, Health and Environment.

8.2. Teaching - Supervision - Juries

8.2.1. Teaching

Most of the members of ERABLE are Assistant / Associate or Full Professors and as such have a heavy load of teaching. Depending on the country, this represents between 150 to 192 hours in front of a class plus the additional work of preparing the courses and exams, and of correcting the latter. Many are also responsible for some of the university courses at the undergraduate or graduate levels.

More in detail:

- In France:
  - Hubert Charles is responsible for the Master of Modelling and Bioinformatics (BIM) at the Insa of Lyon. He teaches 192 hours per year in statistics and biology.
  - Pierluigi Crescenzi taught 120h (72h of Programming in Java for the undergraduate program in Computer Science and 48 of Distributed Algorithms for the Master in Computer Science) at the University of Florence.
  - Vincent Lacroix is responsible for several courses both at the University (L2: Bioinformatics, L3: Advanced Bioinformatics, M1: Methods for Genomics, M1: Methods for Transcriptomics, M1: Projects, M2: Bioethics) and at the Insa (M1: Gene Expression). He teaches 192 hours per year in bioinformatics and statistics.
  - Arnaud Mary taught 109+115 hours in 2016 at the University of Lyon 1 (L1: mathematics; L2: bioinformatics; M1: data analysis; M1: computer science).

  - Cristina Vieira is responsible for the Evolutionary Genetics and Genomics academic career of the Master Ecosciences-Microbiology. She was awarded an IUF (Institut Universitaire de France) distinction and teaches genetics 64 hours per year at the University and ENS Lyon.
- In Italy: Nadia Pisanti taught a total of 104 hours (L1: algorithms and programming; M2: algorithms for bioinformatics).
  - Alberto Marchetti-Spaccamela taught 60 hours of Computing Models (undergraduate class) and 30 hours of Privacy in the electronic society (master class) at Sapienza University of Rome.
  - Nadia Pisanti taught 72h (24h of Programming in C for the undergraduate program in Computer Science and 42 of Algorithms for Bioinformatics for the Master in Computer Science) at the University of Pisa.

Inria or CNRS Junior and senior researchers as well as PhD students and postdocs are also involved in teaching. Notably Alice Julien-Laferrière (PhD student) taught 6 hours (Jury PEL 4); Hélène Lopez-Maestre (PhD student) and Laura Urbini (PhD student) taught each 64 hours of Mathematics and Statistics at the Department of Biology (undergraduate students); Blerina Sinaimeri (Junior Inria Researcher) taught 12h in Discrete Mathematics at the Master of Modelling and Bioinformatics (BIM), INSA, University Lyon 1, as well as 24h at the Master 2 in Computer Science at the ENS Lyon; Fabrice Vavre taught 25h on symbiosis (L3, M1, M2, University Lyon 1, ENS Lyon, University of Poitiers).

Roberto Grossi also participated this year to the Olympiads in Informatics.

8.2.2. Supervision

The following are the PhDs defended in ERABLE in 2016.

- Martin Wannagat, University of Lyon 1, June 2016, supervisors: M.-F. Sagot, A. Marchetti-Spaccamela, L. Stougie.
- Alice Julien-Laferrière, University of Lyon 1, December 2016, supervisors: M.-F. Sagot, V. Lacroix, S. Vinga.

8.2.3. Juries

The following are the PhD or HDR juries to which members of ERABLE participated in 2016.

- Gunnar Klau: Reviewer of the PhD of Arnon Mazza, Tel Aviv University, and of the HDR of Pierre Peterlongo, University of Rennes.
- Marie-France Sagot: Reviewer of the PhD of Yoann Dufresne, University of Lille 1, France.
- Blerina Sinaimeri: Reviewer of the PhD of Nilakantha Paudel, University of Rome, Italy.

8.3. Popularisation

Gunnar Klau wrote an article for ERCIM News about Networks in Biology (http://ercim-news.ercim.eu/en104/special/networks-to-the-rescue-from-big-omics-data-to-targeted-hypotheses).

Roberto Grossi participated to the Pisa CoderDojo (https://www.unipi.it/index.php/news/item/7983-toscana-dojocon-2016-una-giornata-per-la-programmazione-digitale).

Marie-France Sagot, together with Roeland Merks from the CWI, co-edited a special Theme for ERCIM News on Tackling Big Data in the Life Sciences (http://ercim-news.ercim.eu/images/stories/EN104/EN104-web.pdf).

Blerina Sinaimeri participated to the “Conférences ISN et enseignement 2016” organised by Inria, 27 April 2016, to "Les journées nationales de l’APMEP: À la lumière des mathématiques", October 2016, and to “Filles et informatique : une équation lumineuse !” (Nov. 28. 2016, Lyon, France).

Fabrice Vavre participated in a television program on Arte, on the topic of “Ces microbes qui nous gouvernent”.

9. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] A. JULIEN-LAFERRIÈRE. *Models and algorithms applied to metabolism: From revealing the responses to perturbations towards the design of microbial consortia* , Université Lyon 1 - Claude Bernard, December 2016, https://hal.inria.fr/tel-01394113

[2] M. WANNAGAT. *Study of the evolution of symbiosis at the metabolic level using models from game theory and economics*, Université de Lyon, July 2016, https://hal.inria.fr/tel-01394107

Articles in International Peer-Reviewed Journals

[3] R. ANDRADE, M. WANNAGAT, C. C. KLEIN, V. ACUÑA, A. MARCHETTI-SPACCAMELA, P. V. MILREU, L. STOUGIE, M.-F. SAGOT. *Enumeration of minimal stoichiometric precursor sets in metabolic networks*, in "Algorithms for Molecular Biology", December 2016, vol. 11, n^{o} 1, 25 p. [*DOI :* 10.1186/s13015-016-0087-3], https://hal.inria.fr/hal-01368653

[4] P. BAA-PUYOULET, N. PARISOT, G. FEBVAY, J. HUERTA-CEPAS, A. F. VELLOZO, T. GABALDON, F. CALEVRO, H. CHARLES, S. COLELLA. *ArthropodaCyc: a CycADS powered collection of BioCyc databases to analyse and compare metabolism of arthropods*, in "Database - The journal of Biological Databases and Curation", 2016 [*DOI :* 10.1093/DATABASE/BAW081], https://hal.inria.fr/hal-01352558

[5] P. BONIZZONI, R. DONDI, G. W. KLAU, Y. PIROLA, N. PISANTI, S. ZACCARIA. *On the Minimum Error Correction Problem for Haplotype Assembly in Diploid and Polyploid Genomes*, in "Journal of Computational Biology", 2016, vol. 23, n^{o} 9, pp. 718 - 736 [*DOI :* 10.1089/CMB.2015.0220], https://hal.inria.fr/hal-01388448

[6] M. BORASSI, P. CRESCENZI, M. HABIB. *Into the Square: On the Complexity of Some Quadratic-time Solvable Problems*, in "Electronic Notes in Theoretical Computer Science", 2016, vol. 322, pp. 51-67 [*DOI :* 10.1016/J.ENTCS.2016.03.005], https://hal.inria.fr/hal-01390131

[7] T. CALAMONERI, B. SINAIMERI. *Pairwise Compatibility Graphs: A Survey*, in "SIAM Review", 2016, vol. 58, n^{o} 3, pp. 445 - 460 [*DOI :* 10.1137/140978053], https://hal.inria.fr/hal-01388533

[8] S. CANZAR, S. ANDREOTTI, D. W. WEESE, K. REINERT, G. W. KLAU. *CIDANE: comprehensive isoform discovery and abundance estimation*, in "Genome Biology", 2016, vol. 17, 16 p. [*DOI :* 10.1186/s13059-015-0865-0], https://hal.inria.fr/hal-01397539

[9] A. CLEMENTI, P. CRESCENZI, C. DOERR, P. FRAIGNIAUD, F. PASQUALE, R. SILVESTRI. *Rumor Spreading in Random Evolving Graphs*, in "Random Structures and Algorithms", March 2016, vol. 48, n^{o} 2, pp. 290-312 [*DOI :* 10.1002/RSA.20586], https://hal.inria.fr/hal-01390133

[10] P. CRESCENZI, G. D'ANGELO, L. SEVERINI, Y. VELAJ. *Greedily Improving Our Own Closeness Centrality in a Network*, in "ACM Transactions on Knowledge Discovery from Data (TKDD)", August 2016, vol. 11, n° 1, pp. 1-32 [*DOI :* 10.1145/2953882], https://hal.inria.fr/hal-01390134

[11] M. CROCHEMORE, C. EPIFANIO, R. GROSSI, F. MIGNOSI. *Linear-size suffix tries*, in "Theoretical Computer Science", 2016, vol. 638, pp. 171 - 178 [*DOI :* 10.1016/J.TCS.2016.04.002], https://hal.inria.fr/hal-01388452

[12] A. FERRARI, V. VINCENT-SALOMON, X. PIVOT, A.-S. SERTIER, T. THOMAS, L. TONON, S. BOYAULT, E. MULUGETA, I. TREILLEUX, G. MACGROGAN, L. ARNOULD, J. KIELBASSA, V. LE TEXIER, H. BLANCHÉ, J.-F. DELEUZE, J. JACQUEMIER, M.-C. MATHIEU, F. PENAULT-LLORCA, F. BIBEAU, M. MARIANI, C. MANNINA, J.-Y. PIERGA, O. TRÉDAN, H. BONNEFOI, G. ROMIEU, P. FUMOLEAU, S. DELALOGE, M. RIOS, J.-M. FERRERO, C. TARPIN, C. BOUTEILLE, F. CALVO, I. G. GUT, M. GUT, S. MARTIN, S. NIK-ZAINAL, M. R. STRATTON, I. PAUPORTÉ, P. SAINTIGNY, D. BIRNBAUM, A. VIARI, G. THOMAS. *A whole-genome sequence and transcriptome perspective on HER2-positive breast cancers*, in "Nature Communications", 2016, vol. 7 [*DOI :* 10.1038/NCOMMS12222], https://hal.inria.fr/hal-01388446

[13] M. G. FERRARINI, F. M. SIQUEIRA, S. G. MUCHA, T. L. PALAMA, É. JOBARD, B. ELENA-HERRMANN, A. T. RIBEIRO DE VASCONCELOS, F. TARDY, I. S. SCHRANK, A. ZAHA, M.-F. SAGOT. *Insights on the virulence of swine respiratory tract mycoplasmas through genome-scale metabolic modeling*, in "BMC Genomics", December 2016, vol. 17, n° 1, 353 p. [*DOI :* 10.1186/S12864-016-2644-Z], https://hal.inria.fr/hal-01315893

[14] F. FREYERMUTH, F. RAU, Y. KOKUNAI, T. LINKE, C. SELLIER, M. NAKAMORI, Y. KINO, L. ARANDEL, A. JOLLET, C. THIBAULT, M. PHILIPPS, S. VICAIRE, B. JOST, B. UDD, J. W. DAY, D. DUBOC, K. WAHBI, T. MATSUMURA, H. FUJIMURA, H. MOCHIZUKI, F. DERYCKERE, T. KIMURA, N. NUKINA, S. ISHIURA, V. LACROIX, A. CAMPAN-FOURNIER, V. NAVRATIL, E. CHAUTARD, D. AUBOEUF, M. HORIE, K. IMOTO, K.-Y. LEE, M. S. SWANSON, A. L. DE MUNAIN, S. INADA, H. ITOH, K. NAKAZAWA, T. ASHIHARA, E. WANG, T. ZIMMER, D. FURLING, M. P. TAKAHASHI, N. CHARLET-BERGUERAND. *Splicing misregulation of SCN5A contributes to cardiac-conduction delay and heart arrhythmia in myotonic dystrophy*, in "Nature Communications", 2016, vol. 7 [*DOI :* 10.1038/NCOMMS11067], https://hal.inria.fr/hal-01388496

[15] R. GROSSI, C. S. ILIOPOULOS, R. MERCAS, N. PISANTI, S. P. PISSIS, A. RETHA, F. VAYANI. *Circular sequence comparison: algorithms and applications*, in "Algorithms for Molecular Biology", 2016, vol. 11, n° 1 [*DOI :* 10.1186/S13015-016-0076-6], https://hal.inria.fr/hal-01388449

[16] A. JULIEN-LAFERRIÈRE, L. BULTEAU, D. PARROT, A. MARCHETTI-SPACCAMELA, L. STOUGIE, S. VINGA, A. MARY, M.-F. SAGOT. *A Combinatorial Algorithm for Microbial Consortia Synthetic Design*, in "Scientific Reports", July 2016 [*DOI :* 10.1038/SREP29182], https://hal.archives-ouvertes.fr/hal-01344296

[17] Y. LE BRAS, O. COLLIN, C. MONJEAUD, V. LACROIX, E. RIVALS, C. LEMAITRE, V. MIELE, G. SACOMOTO, C. MARCHET, B. CAZAUX, A. ZINE EL AABIDINE, L. SALMELA, S. ALVES-CARVALHO, A. ANDRIEUX, R. URICARU, P. PETERLONGO. *Colib'read on galaxy: a tools suite dedicated to biological information extraction from raw NGS reads*, in "GigaScience", February 2016, vol. 5, n° 1 [*DOI :* 10.1186/S13742-015-0105-2], https://hal.inria.fr/hal-01280238

[18] H. LOPEZ-MAESTRE, L. BRINZA, C. MARCHET, J. KIELBASSA, S. BASTIEN, M. BOUTIGNY, D. MONNIN, A. EL FILALI, C. M. CARARETO, C. VIEIRA, F. PICARD, N. KREMER, F. VAVRE, M.-F. SAGOT, V. LACROIX. *SNP calling from RNA-seq data without a reference genome: identification, quan-*

tification, differential analysis and impact on the protein sequence, in "Nucleic Acids Research", 2016 [*DOI :* 10.1093/NAR/GKW655], https://hal.inria.fr/hal-01352586

[19] T. MARSCHALL, M. MARZ, T. ABEEL, L. DIJKSTRA, B. E. DUTILH, A. GHAFFAARI, P. KERSEY, W. P. KLOOSTERMAN, V. MAKINEN, A. M. NOVAK, B. PATEN, D. PORUBSKY, E. RIVALS, C. ALKAN, J. A. BAAIJENS, P. I. W. D. BAKKER, V. BOEVA, R. J. P. BONNAL, F. CHIAROMONTE, R. CHIKHI, F. D. CICCARELLI, R. CIJVAT, E. DATEMA, C. M. V. DUIJN, E. E. EICHLER, C. ERNST, E. ESKIN, E. GARRISON, M. EL-KEBIR, G. W. KLAU, J. O. KORBEL, E.-W. LAMEIJER, B. LANGMEAD, M. MARTIN, P. MEDVEDEV, J. C. MU, P. NEERINCX, K. OUWENS, P. PETERLONGO, N. PISANTI, S. RAHMANN, B. RAPHAEL, K. REINERT, D. D. RIDDER, J. D. RIDDER, M. SCHLESNER, O. SCHULZ-TRIEGLAFF, A. D. SANDERS, S. SHEIKHIZADEH, C. SHNEIDER, S. SMIT, D. VALENZUELA, J. WANG, L. WESSELS, Y. ZHANG, V. GURYEV, F. VANDIN, K. YE, A. SCHÖNHUTH. *Computational pan-genomics: status, promises and challenges*, in "Briefings in Bioinformatics", October 2016 [*DOI :* 10.1093/BIB/BBW089], https://hal.inria.fr/hal-01390478

[20] A. MAY, B. W. BRANDT, M. EL-KEBIR, G. W. KLAU, E. ZAURA, W. CRIELAARD, J. HERINGA, S. ABELN. *metaModules identifies key functional subnetworks in microbiome-related disease*, in "Bioinformatics", 2016, vol. 32, n^{o} 11, pp. 1678 - 1685 [*DOI :* 10.1093/BIOINFORMATICS/BTV526], https://hal.inria.fr/hal-01388508

[21] S. NIK-ZAINAL, H. R. DAVIES, J. STAAF, M. RAMAKRISHNA, D. GLODZIK, X. ZOU, I. MARTINCORENA, L. B. ALEXANDROV, S. MARTIN, D. C. WEDGE, P. VAN LOO, Y. S. JU, M. SMID, A. B. BRINKMAN, S. MORGANELLA, M. R. AURE, O. C. LINGJÆRDE, A. LANGERØD, M. RINGNÉR, S.-M. AHN, S. BOYAULT, J. E. BROCK, A. BROEKS, A. BUTLER, C. DESMEDT, L. DIRIX, S. DRONOV, A. FATIMA, J. A. FOEKENS, M. GERSTUNG, G. K. J. HOOIJER, S. J. JANG, D. R. JONES, H.-Y. KIM, T. A. KING, S. KRISHNAMURTHY, H. J. LEE, J.-Y. LEE, Y. LI, S. MCLAREN, A. MENZIES, V. MUSTONEN, S. O'MEARA, I. PAUPORTÉ, X. PIVOT, C. A. PURDIE, K. RAINE, K. RAMAKRISHNAN, F. G. RODRÍGUEZ-GONZÁLEZ, G. ROMIEU, A. M. SIEUWERTS, P. SIMPSON, R. SHEPHERD, L. STEBBINGS, O. A. STEFANSSON, J. TEAGUE, S. TOMMASI, I. TREILLEUX, G. G. VAN DEN EYNDEN, P. VERMEULEN, A. VINCENT-SALOMON, L. YATES, C. CALDAS, L. V. VEER, A. TUTT, S. KNAPPSKOG, B. K. T. TAN, J. JONKERS, Å. BORG, N. T. UENO, C. SOTIRIOU, A. VIARI, P. A. FUTREAL, P. J. CAMPBELL, P. N. SPAN, S. VAN LAERE, S. R. LAKHANI, J. E. EYFJORD, 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, G. THOMAS, M. R. STRATTON. *Landscape of somatic mutations in 560 breast cancer whole-genome sequences*, in "Nature", 2016, vol. 534, n^{o} 7605, pp. 47 - 54 [*DOI :* 10.1038/NATURE17676], https://hal.inria.fr/hal-01388447

[22] Y. PIROLA, S. ZACCARIA, R. DONDI, G. W. KLAU, N. PISANTI, P. BONIZZONI. *HapCol: accurate and memory-efficient haplotype assembly from long reads*, in "Bioinformatics", 2016 [*DOI :* 10.1093/BIOINFORMATICS/BTV495], https://hal.inria.fr/hal-01225984

[23] G. RODRIGUES GALVAO, C. BAUDET, Z. DIAS. *Sorting Circular Permutations by Super Short Reversals*, in "IEEE/ACM Transactions on Computational Biology and Bioinformatics", January 2016 [*DOI :* 10.1109/TCBB.2016.2515594], https://hal.inria.fr/hal-01317003

[24] V. ROMERO-SORIANO, N. BURLET, D. VELA, A. FONTDEVILA, C. VIEIRA, M. P. GARCIÁ GUERREIRO. *Drosophila Females Undergo Genome Expansion after Interspecific Hybridization*, in "Genome Biology and Evolution", February 2016, vol. 8, n^{o} 3, pp. 556-561 [*DOI :* 10.1093/GBE/EVW024], https://hal.inria.fr/hal-01352572

[25] P. SIMONET, G. DUPORT, K. GAGET, M. WEISS-GAYET, S. COLELLA, G. FEBVAY, H. CHARLES, J. VIÑUELAS, A. HEDDI, F. CALEVRO. *Direct flow cytometry measurements reveal a fine-tuning of symbiotic cell dynamics according to the host developmental needs in aphid symbiosis*, in "Scientific Reports", 2016, vol. 6, 19967 p. [*DOI :* 10.1038/SREP19967], https://hal.inria.fr/hal-01352561

[26] P. SIMONET, K. GAGET, N. PARISOT, G. DUPORT, M. REY, G. FEBVAY, H. CHARLES, P. CALLAERTS, S. COLELLA, F. CALEVRO. *Disruption of phenylalanine hydroxylase reduces adult lifespan and fecundity, and impairs embryonic development in parthenogenetic pea aphids*, in "Scientific Reports", 2016, vol. 6 [*DOI :* 10.1038/SREP34321], https://hal.inria.fr/hal-01388523

[27] F. M. SIQUEIRA, G. LOSS DE MORAIS, S. HIGASHI, L. SCHERER BEIER, G. MERKER BREYER, C. PADOAN DE SÁ GODINHO, M.-F. SAGOT, I. SILVEIRA SCHRANK, A. ZAHA, A. T. RIBEIRO DE VASCONCELOS. *Mycoplasma non-coding RNA: identification of small RNAs and targets*, in "BMC Genomics", 2016, vol. 23, pp. 1289 - 26 [*DOI :* 10.1186/S12864-016-3061-Z], https://hal.inria.fr/hal-01393122

[28] M. SMID, F. G. RODRÍGUEZ-GONZÁLEZ, A. M. SIEUWERTS, R. SALGADO, W. J. C. PRAGER-VAN DER SMISSEN, M. V. D. VLUGT-DAANE, A. VAN GALEN, S. NIK-ZAINAL, J. STAAF, A. B. BRINKMAN, M. J. VAN DE VIJVER, A. L. RICHARDSON, A. FATIMA, K. BERENTSEN, A. BUTLER, S. MARTIN, H. R. DAVIES, R. DEBETS, M. E. M.-V. GELDER, C. H. M. VAN DEURZEN, G. MACGROGAN, G. G. G. M. VAN DEN EYNDEN, C. PURDIE, A. M. THOMPSON, C. CALDAS, P. N. SPAN, P. T. SIMPSON, S. R. LAKHANI, S. VAN LAERE, C. DESMEDT, M. RINGNÉR, S. TOMMASI, J. EYFORD, A. BROEKS, A. VINCENT-SALOMON, P. A. FUTREAL, S. KNAPPSKOG, T. A. KING, G. THOMAS, A. VIARI, A. LANGERØD, A.-L. BØRRESEN-DALE, E. BIRNEY, H. G. STUNNENBERG, M. STRATTON, J. A. FOEKENS, J. W. M. MARTENS. *Breast cancer genome and transcriptome integration implicates specific mutational signatures with immune cell infiltration*, in "Nature Communications", 2016, vol. 7 [*DOI :* 10.1038/NCOMMS12910], https://hal.inria.fr/hal-01388445

[29] L. URBINI, C. MATIAS, M.-F. SAGOT, B. SINAIMERI. *Robustness of the Parsimonious Reconciliation Method in Cophylogeny*, in "Springer - Lecture Notes in Computer Science (LNCS)", June 2016, vol. 9702, 12 p. [*DOI :* 10.1007/978-3-319-38827-4_10], https://hal.inria.fr/hal-01349773

[30] A. VERÍSSIMO, A. L. OLIVEIRA, M.-F. SAGOT, S. VINGA. *DegreeCox – a network-based regularization method for survival analysis*, in "BMC Bioinformatics", December 2016, vol. 17, n^{o} Suppl 16, pp. 109-121 [*DOI :* 10.1186/s12859-016-1310-4], https://hal.inria.fr/hal-01415968

International Conferences with Proceedings

[31] E. BERGAMINI, M. BORASSI, P. CRESCENZI, A. MARINO, H. MEYERHENKE. *Computing Top-k Closeness Centrality Faster in Unweighted Graphs*, in "Eighteenth Workshop on Algorithm Engineering and Experiments, ALENEX 2016", Arlington, United States, 2016 [*DOI :* 10.1137/1.9781611974317.6], https://hal.inria.fr/hal-01390137

[32] V. BONIFACI, B. BRANDENBURG, G. D'ANGELO, A. MARCHETTI-SPACCAMELA. *Multiprocessor Real-Time Scheduling with Hierarchical Processor Affinities*, in "28th Euromicro Conference on Real-Time Systems, ECRTS 2016", Toulouse, France, IEEE Computer Society, July 2016, pp. 237-247 [*DOI :* 10.1109/ECRTS.2016.24], https://hal.inria.fr/hal-01397807

[33] R. BRUNI, A. MARCHETTI-SPACCAMELA, S. K. BARUAH, V. BONIFACI. *ILP-Based Approaches to Partitioning Recurrent Workloads Upon Heterogeneous Multiprocessors ILP-based approaches to partitioning recurrent workloads upon heterogeneous multiprocessors*, in "28th Euromicro Conference on

Real-Time Systems, ECRTS 2016", Toulouse, France, IEEE Computer Society, July 2016, pp. 215-225 [*DOI :* 10.1109/ECRTS.2016.10], https://hal.inria.fr/hal-01397810

[34] K. BÖHMOVÁ, M. MIHALÁK, T. PRÖGER, G. SACOMOTO, M.-F. SAGOT, P. WIDMAYER. *Computing and Listing st-Paths in Subway Networks*, in "CSR 2016 - 11th International Computer Science Symposium in Russia", St. Petersburg, Russia, A. S. KULIKOV, G. J. WOEGINGER (editors), Lecture Notes in Computer Science, Springer, June 2016, vol. 9691, pp. 102-116 [*DOI :* 10.1007/978-3-319-34171-2_8], https://hal.inria.fr/hal-01348869

[35] T. CALAMONERI, M. GASTALDELLO, A. MARY, B. SINAIMERI, M.-F. SAGOT. *On Maximal Chain Subgraphs and Covers of Bipartite Graphs*, in "Combinatorial Algorithms - 27th International Workshop, IWOCA 2016", Helsinki, Finland, Lecture Notes in Computer Science, 2016, vol. 9843, pp. 137-150 [*DOI :* 10.1007/978-3-319-44543-4_11], https://hal.inria.fr/hal-01388546

[36] F. CAMBI, P. CRESCENZI, L. PAGLI. *Analyzing and Comparing On-Line News Sources via (Two-Layer) Incremental Clustering*, in "8th International Conference on Fun with Algorithms, FUN 2016", La Maddalena, Italy, June 2016 [*DOI :* 10.4230/LIPICS.FUN.2016.9], https://hal.inria.fr/hal-01390139

[37] A. CONTE, R. GROSSI, A. MARINO. *Clique covering of large real-world networks*, in "31st Annual ACM Symposium on Applied Computing (SAC 2016)", Pisa, Italy, ACM, April 2016 [*DOI :* 10.1145/2851613.2851816], https://hal.inria.fr/hal-01388477

[38] A. CONTE, R. GROSSI, A. MARINO, R. RIZZI. *Listing Acyclic Orientations of Graphs with Single and Multiple Sources*, in "LATIN 2016: Theoretical Informatics - 12th Latin American Symposium", Ensenada, Mexico, E. KRANAKIS, G. NAVARRO, E. CHÁVEZ (editors), Lecture Notes in Computer Science, Springer, April 2016, vol. 9644, pp. 319-333 [*DOI :* 10.1007/978-3-662-49529-2_24], https://hal.inria.fr/hal-01388470

[39] A. CONTE, R. GROSSI, A. MARINO, R. RIZZI, L. VERSARI. *Directing Road Networks by Listing Strong Orientations*, in "Combinatorial Algorithms - 27th International Workshop, IWOCA 2016", Helsinki, Finland, V. MÄKINEN, S. J. PUGLISI, L. SALMELA (editors), Lecture Notes in Computer Science, Springer, August 2016, vol. 9843 [*DOI :* 10.1007/978-3-319-44543-4_7], https://hal.inria.fr/hal-01388476

[40] A. CONTE, R. GROSSI, A. MARINO, L. VERSARI. *Sublinear-Space Bounded-Delay Enumeration for Massive Network Analytics: Maximal Cliques* *, in "43rd International Colloquium on Automata, Languages, and Programming (ICALP 2016)", Schloss Dagstuhl, Germany, July 2016, vol. 148, pp. 1 - 148 [*DOI :* 10.4230/LIPICS.ICALP.2016.148], https://hal.inria.fr/hal-01388461

[41] A. MARY, Y. STROZECKI. *Efficient Enumeration of Solutions Produced by Closure Operations*, in "33rd Symposium on Theoretical Aspects of Computer Science, STACS 2016", Orléans, France, LIPIcs, February 2016, vol. 47 [*DOI :* 10.4230/LIPICS.STACS.2016.52], https://hal.inria.fr/hal-01388505

[42] C. MASSIMO, R. GROSSI, R. RIZZI. *New Bounds for Approximating Extremal Distances in Undirected Graphs*, in "Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016", Arlington, United States, SIAM, January 2016 [*DOI :* 10.1137/1.9781611974331.CH27], https://hal.inria.fr/hal-01388484

Other Publications

[43] M. G. FERRARINI, F. M. SIQUEIRA, S. G. MUCHA, T. L. PALAMA, E. E. JOBARD, B. E. HERRMANN, A. T. RIBEIRO DE VASCONCELOS, F. F. TARDY, I. S. SCHRANK, A. ZAHA, M.-F. SAGOT. *Metabolic Investigation of the Mycoplasmas from the Swine Respiratory Tract*, September 2016, JOBIM 2016, Poster [*DOI :* 10.1186/s12864-016-2644-z], https://hal.inria.fr/hal-01394118

[44] A. JULIEN-LAFERRIÈRE, L. BULTEAU, D. PARROT, A. MARCHETTI-SPACCAMELA, L. STOUGIE, S. VINGA, A. MARY, M.-F. SAGOT. *MultiPus: Conception de communautés microbiennes pour la production de composés d'intérêt*, June 2016, Jobim, Poster, https://hal.inria.fr/hal-01394119