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 2015

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

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
7.10. - Network science
7.11. - Performance evaluation
7.2. - Discrete mathematics, combinatorics
7.3. - Operations research, optimization, game theory
7.9. - Graph theory

Other Research Topics and Application Domains:

1. - Life sciences
1.1. - Biology
1.1.1. - Structural biology
1.1.11. - Systems biology
1.1.12. - Synthetic biology
1.1.2. - Molecular biology
1.1.5. - Genetics
1.1.6. - Genomics
1.1.8. - Evolutionnary biology
1.1.9. - Bioinformatics
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
2.3. - Epidemiology

ERABLE is a European joint project-team bringing together researchers from the former Inria project-team Bamboo, researchers in Italy under the banner of the Sapienza University of Rome from the Sapienza, the University of Florence, and the University of Pisa, and researchers in The Netherlands under the banner of CWI from the Free University of Amsterdam and CWI.

1. Members

Research Scientists

Gunnar Klau [CWI, The Netherlands, Researcher]
Marie-France Sagot [Team leader, Inria, Senior Researcher, HdR]
Blerina Sinaimeri [Inria, Junior Researcher, from October 2015]
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]
Christian Gautier [University Lyon I, 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]
Leen Stougie [Free University Amsterdam & CWI, The Netherlands, Full Professor]
Cristina Vieira [University Lyon I, Full Professor, HdR]

Engineers

Christian Baudet [Inria, grant by European Research Council, until April 2015]
Camille Marchet [Inria, funded by ANR, until September 2015]
Maria Francesca Zini [University of Pisa, until December 2015]

PhD Students

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]
Mariana Galvão Ferrarini [Inria, grant by European Research Council, co-supervised by Arnaldo Zaha, Federal University of Rio Grande do Sul, Brazil, and Marie-France Sagot]
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, since March 2015, co-supervised by Giuseppe Italiano, Tor Vergata University pf Rome, Vincent Lacroix, and Marie-France Sagot, since March 2015]
Alice Julien-Laferrière [Inria, grant by FP7 KBBE project, co-supervised by Vincent Lacroix, Marie-France Sagot, and Susana Vinga, IDMEC-IST, Lisbon, Portugal]
Hélène Lopez-Maestre [University Lyon I, co-supervised by Vincent Lacroix and Cristina Vieira]
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, from September 2015, 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]

Post-Doctoral Fellows

Laurent Bulteau [Inria, grant by FP7 KBBE project, until September 2015]
Emilie Chautard [Inria, grant by INSERM, until August 2015]
Ricardo de Andrade Abrantes [Science Without Frontiers, Ministry of Research Brazil, until March 2015]
Susan Higashi [Inria, grant by European Research Council, until May 2015]
Andrea Marino [University of Pisa, since March 2015]
Delphine Parrot [Inria, grant by FP7 KBBE project]
Arne Reimers [CWI, until August 2015]
Gustavo A. T. Sacomoto [Inria, grant by ANR, until February 2015]
Blerina Sinaimeri [Inria, grant by ANR, until September 2015]
Paulo Trenhago [Capes-Cofecub, until March 2015]

Administrative Assistants

Florence Bouheddi [Inria, until December 2015]
Marina Da Graça [Inria, from September 2015]

Others

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 DESCRIPTIONCIDANE 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 visualisator for host-parasite and gene-specie trees evolution..

- Participants: Laurent Bulteau
- Contact: Laurent Bulteau (not Inria), Blerina Sinaimeri (for Inria)
- 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
- 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, Pierluigi Crescenzi, Marie-France Sagot, Blerina Sinaimeri,
- Contact: Christian Baudet (not Inria), Beatrice Donati (not Inria), and Marie-France Sagot
- 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
- 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: https://github.com/AlgoLab/HapCol

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.

- Participants: Vincent Lacroix, Pierre Peterlongo
- Contact: Pierre Peterlongo (EPI GenScale)

5.14. KisSplice & KisSplice2igv7

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. KISSPLICE2IGV is a pipeline that combines the outputs of KISSPLICE to a reference transcriptome (obtained with a full-length transcriptome assembler or a reference database). It provides a visualisation of the events found by KISSPLICE in a longer context using a genome browser (IGV).

- 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: 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. PhEVER

FUNCTIONAL DESCRIPTION

Database of homologous gene families built from the complete genomes of all available viruses, prokaryotes and eukaryotes and aimed at the detection of virus/virus and virus/host lateral gene transfers.

- Participants: Christian Gautier, Vincent Lotteau, Leonor Palmeira, Simon Penel, Chantal Rabourdin-Combe
- Contact: Christian Gautier, Leonor Palmeira (not EPI)
- URL: http://pbil.univ-lyon1.fr/databases/phever

5.25. PepLine

FUNCTIONAL DESCRIPTION

Pipeline for the high-throughput analysis of proteomic data.

- Participant: Jérôme Garin, Alain Viari
- Contact: Alain Viari

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. Smile

FUNCTIONAL DESCRIPTION

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

- Participants: Laurent Marsan, Marie-France Sagot
- Contact: Marie-France Sagot
- URL: Not available

5.30. UniPathway

FUNCTIONAL DESCRIPTION

Database of manually curated pathways developed with the Swiss-Prot group.

- Participants: Eric Coissac, Anne Morgat, Alain Viari
- Contact: Anne Morgat
- URL: http://www.unipathway.org/

5.31. WhatsHap and pWH

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. PWH 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

6. New Results

6.1. General comments

We present in this section the main results obtained in 2015. Some were already in preparation or submitted at the end of 2014. It will be indicated whenever this is the case.

We tried to organise the results following four of 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 concerned by such results. As concerns the Axis "Going towards control", a work is in preparation that fits it. It will be presented in 2016.

We did not indicate here the results on more theoretical aspects of computer science if it did not seem for now that they could be relevant in contexts related to computational biology. Actually, we do believe those on scheduling (by, among others, A. Marchetti-Spaccamela and/or L. Stougie) [3], [38], [39], [10], [31], [23], [44], [43] or even one result related to context-free grammars (by, among others, P. Crescenzi) [11] could in the future become relevant for the life sciences (biology or ecology). However, we preferred for now 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, 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

Genomic / NGS data management
Next-generation sequencing (NGS) technology has led the life sciences into the big data era. Today, sequencing genomes takes little time and cost, but yields terabytes of data to be stored and analysed. The biologists are often exposed to excessively time consuming and error-prone data management and analysis hurdles. We therefore proposed a database management system (DBMS) based approach to accelerate and substantially simplify genome sequence analysis [9]. To that aim, we extended MONETDB, an open-source column-based DBMS (urlhttps://www.monetdb.org), with a BAM module, which enables easy, flexible, and rapid management and analysis of sequence alignment data stored as Sequence Alignment/Map (SAM/BAM) files. The main features of MONETDB/BAM were described using a case study on Ebola virus.

We also designed and realised a knowledge base for collecting, elaborating, and extracting analytical results of genomic, proteomic, biochemical, morphological investigations from animal models of cerebral stroke [45]. Data analysis techniques are tailored to make the data available for processing and correlation, in order to increase the predictive value of the preclinical data, to perform bio-simulation studies, and to support both academic and industrial research in the area of cerebral stroke therapy. The low reliability of animal models in replicating the human disease is one of the most serious problems in the field of medical and pharmaceutical research about stroke. The standard models for the study of ischaemic stroke are often poorly predictive as they simulate only partially the human disease. This work aims therefore at investigating animal models with diseases typically associated with the onset of stroke in human patients. A first statistical analysis of the retrieved information led to the validation of our animal models and suggested a predictive and translational value for parameters related to a specific model. In particular, concerning gene expression data, we applied a data analysis pipeline that initially takes into account an initial set of 64,000 genes and brought down the focus on a few tens of them.

NGS data analysis
The problem of enumerating bubbles with length constraints in directed graphs arises in transcriptomics where the question is to identify all alternative splicing events present in a sample of mRNAs sequenced by RNA?seq. We presented a new algorithm for enumerating bubbles with length constraints in weighted directed graphs [30]. This is the first polynomial delay algorithm for this problem and we showed that in practice, it is faster than previous approaches. This settled one of the main open questions from previous literature. Moreover, the new algorithm allows us to deal with larger instances and possibly detect longer alternative splicing events.

We also developed CIDANE, a novel framework for genome-based transcript reconstruction and quantification from RNA-seq reads [37]. CIDANE assembles transcripts with significantly higher sensitivity and precision than existing tools, while competing in speed with the fastest methods. In addition to reconstructing transcripts *ab initio*, the algorithm also allows to make 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.

SNPs (Single Nucleotide Polymorphisms) are genetic markers used in many areas of biology. Their precise identification is a prerequisite for association studies, which associate genotypes to phenotypes. Methods are currently developed for model species, but rely on the availability of a (good) reference genome, and 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. We proposed a method 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 for sequencing from one individual. This pooling strategy still enables to allelotype loci and to associate them to phenotypes. Using human RNA-seq data, we first compared the performance of our algorithm, KISSSPLICE, with GATK, a well established method that requires a reference genome. We showed that both methods perform similarly in terms of precision and recall. We then validated experimentally the predictions of our method using RNA-seq data from two non-model species. The method can be used for any species to annotate SNPs and to predict their impact on proteins. It can further be used to assess variants that are associated to a particular phenotype within a population, when replicates are provided for each biological condition. This work was submitted at the end of 2015.

Sequence alignment (full genomes or NGS data)
Sequence comparison is a fundamental step in many important tasks related to biology. 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 genome structure is a common phenomenon in nature, a caveat of specialised 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 computed efficiently for circular sequence comparison [41]. Experimental results, using real and synthetic data, demonstrated orders-of-magnitude superiority of our approach in terms of efficiency, while maintaining an accuracy very competitive to the state of the art.

Burrows-Wheeler Transform (BWT) has been successfully used to reduce the memory requirement for sequence alignment. We improved on previous results related to the problem of computing the Burrows-Wheeler Transform (BWT) using small additional space [12]. Our in-place algorithm does not need the explicit storage for the suffix sort array and the output array, as typically required in such previous work. It relies on the combinatorial properties of the BWT, and runs in $O(n^2)$ time in the comparison model using $O(1)$ extra memory cells, apart from the array of n cells storing the n characters of the input text. We then discussed the time-space trade-off when $O(k\sigma k)$ extra memory cells are allowed with σk distinct characters, providing an $O((n^2/k + n)\log?k)$-time algorithm to obtain (and invert) the BWT. In real systems where the alphabet size is a constant, for any arbitrarily small $\epsilon > 0$, the BWT of a text of n bytes can be computed in $O(n\sigma^{-1}\log n)$ time using just σn extra bytes.

Genome assembly problems
The human genome is diploid, which requires assigning heterozygous single nucleotide polymorphisms (SNPs) to the two copies of the genome. The resulting haplotypes, lists of SNPs belonging to each copy, are crucial for downstream analyses in population genetics. Currently, statistical approaches, which are oblivious to direct read information, constitute the state-of-the-art. Haplotype assembly, which addresses phasing directly from sequencing reads, suffers from the fact that sequencing reads of the current generation are too short to serve the purposes of genome-wide phasing. While future-technology sequencing reads will contain sufficient amounts of SNPs per read for phasing, they are also likely to suffer from higher sequencing

error rates. Currently, no haplotype assembly approaches exist that allow for taking both increasing read length and sequencing error information into account. We developed WHATSHAP, the first approach that yields provably optimal solutions to the weighted minimum error correction problem in runtime linear in the number of SNPs [25]. WHATSHAP is a fixed parameter tractable (FPT) approach with coverage as the parameter. We demonstrated that WHATSHAP can handle datasets of coverage up to 20x, and that 15x are generally enough for reliably phasing long reads, even at significantly elevated sequencing error rates. We also find that the switch and flip error rates of the haplotypes we output are favourable when comparing them with state-of-the-art statistical phasers. By using novel combinatorial properties of Minimum Error Correction (MEC) instances, we were then able to provide new results on the fixed-parameter tractability and approximability of MEC [35]. In particular, we showed that MEC is in FPT when parameterised by the number of corrections, and, on "gapless" instances, it is in FPT also when parameterised by the length of the fragments, whereas the result known in the literature forces the reconstruction of complementary haplotypes. We then showed that MEC cannot be approximated within any constant factor while it is approximable within factor $O(\log nm)$ where nm is the size of the input. Finally, we provided a practical 2-approximation algorithm for the Binary MEC, a variant of MEC that has been applied in the framework of clustering binary data. Finally, 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 SNP position and that minimises the overall error-correction score [26]. We performed an experimental analysis, comparing HAPCOL with 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.

Completing the genome sequence of an organism is an important task in comparative, functional and structural genomics. However, this remains a challenging issue from both a computational and an experimental viewpoint. Genome scaffolding (*i.e.* the process of ordering and orientating contigs) of de novo assemblies usually represents the first step in most genome finishing pipelines. We developed MEDUSA (Multi-Draft based Scaffolder), an algorithm for genome scaffolding [6]. MEDUSA exploits information obtained from a set of (draft or closed) genomes from related organisms to determine the correct order and orientation of the contigs. MEDUSA formalises the scaffolding problem by means of a combinatorial optimisation formulation on graphs and implements an efficient constant factor approximation algorithm to solve it. In contrast to currently used scaffolders, it does not require either prior knowledge on the microrganisms dataset under analysis (*e.g.* their phylogenetic relationships) or the availability of paired end read libraries. This makes usability and running time two additional important features of our method. Moreover, benchmarks and tests on real bacterial datasets showed that MEDUSA is highly accurate and, in most cases, outperforms traditional scaffolders. The possibility to use MEDUSA on eukaryotic datasets has also been evaluated, leading to interesting results. medusa/releases.

Genome annotation problems

Repetitive DNA, including transposable elements (TEs), is found throughout eukaryotic genomes. Annotating and assembling the "repeatome" during genome-wide analysis often poses a challenge. To address this problem, we developed DNAPIPETE – a new pipeline that uses a sample of raw genomic reads [20]. It produces precise estimates of repeated DNA content and TE consensus sequences, as well as the relative ages of TE families. We showed that DNAPIPETE performs well using very low coverage sequencing in different genomes, losing accuracy only with old TE families. We applied this pipeline to the genome of the Asian tiger mosquito *Aedes albopictus*, an invasive species of human health interest, for which the genome size is estimated to be over 1 Gbp. Using DNAPIPETE, we showed that this species harbours a large (50% of the genome) and potentially active repeatome with an overall TE class and order composition similar to that of *Aedes aegypti*, the yellow fever mosquito. However, intra-order dynamics showed clear distinctions between the two species, with differences at the TE family level. Our pipeline's ability to manage the repeatome

annotation problem will make it helpful for new or ongoing assembly projects, and our results will benefit future genomic studies of *A. albopictus*.

On another topic, we developed a reliable, robust, and much faster method for the prediction of pre-miRNAs [22]. With this method, we aimed mainly at two goals: efficiency and flexibility. Efficiency was made possible by means of a quadratic algorithm. Since the majority of the predictors use a cubic algorithm to verify the pre-miRNA hairpin structure, they may take too long when the input is large. Flexibility relies on two aspects, the input type and the organism clade. MIRINHO can receive as input both a genome sequence and small RNA sequencing (sRNA-seq) data of both animal and plant species. To change from one clade to another, it suffices to change the lengths of the stem-arms and of the terminal loop. Concerning the prediction of plant miRNAs, because their pre-miRNAs are longer, the methods for extracting the hairpin secondary structure are not as accurate as for shorter sequences. With MIRINHO, we also addressed this problem, which enabled to provide premiRNA secondary structures more similar to the ones in MIRBASE than the other available methods. MIRINHO served also as the basis to two other issues we addressed. The first issue led to the treatment and analysis of sRNA-seq data of *Acyrthosiphon pisum*, the pea aphid. The goal was to identify the miRNAs that are expressed during the four developmental stages of this species, allowing further biological conclusions concerning the regulatory system of such an organism. For this analysis, we developed a whole pipeline, called MIRINHOPIPE, at the end of which MIRINHO was aggregated. A paper is currently being prepared that presents this work.

6.3. Inferring and analysing the networks of molecular elements

Protein structure comparison

We proposed a new distance measure for comparing two protein structures based on their contact map representations [1]. We showed that our novel measure, which we refer to as the maximum contact map overlap (max-CMO) metric, satisfies all properties of a metric on the space of protein representations. Having a metric in that space allows one to avoid pairwise comparisons on the entire database and, thus, to significantly accelerate exploring the protein space compared to no-metric spaces. We showed on a gold standard superfamily classification benchmark set of 6759 proteins that our exact k-nearest neighbour ($k-NN$) scheme classifies up to 224 out of 236 queries correctly and on a larger, extended version of the benchmark with 60; 850 additional structures, up to 1361 out of 1369 queries. Our $k-NN$ classification thus provides a promising approach for the automatic classification of protein structures based on flexible contact map overlap alignments.

Metabolic network analysis

Flux balance analysis (FBA) is one of the most often applied methods on genome-scale metabolic networks. Although FBA uniquely determines the optimal yield, the pathway that achieves this is usually not unique. The analysis of the optimal-yield flux space has been an open challenge. Flux variability analysis is only capturing some properties of the flux space, while elementary mode analysis is intractable due to the enormous number of elementary modes. However, it had been previously found that the space of optimal-yield fluxes decomposes into flux modules. These decompositions allow a much easier but still comprehensive analysis of the optimal-yield flux space. Using the mathematical definition of module introduced by Müller and Bockmayr in 2013, we discovered that flux modularity is rather a local than a global property which opened connections to matroid theory [28]. Specifically, we showed that our modules correspond one-to-one to so-called separators of an appropriate matroid. Employing efficient algorithms developed in matroid theory we are now able to compute the decomposition into modules in a few seconds for genome-scale networks. Using that every module can be represented by one reaction that corresponds to its function, we also presented a method that uses this decomposition to visualise the interplay of modules. We expect the new method to replace flux variability analysis in the pipelines for metabolic networks.

Integrated network analysis

Data on molecular interactions is increasing at a tremendous pace. Since biological functionality primarily operates at the network level, there is a clear need for topology-aware comparison methods. We developed one such method for global network alignment that is fast and robust and can flexibly deal with various scoring schemes taking both node-to-node correspondences as well as network topologies into account [18]. We exploited that network alignment is a special case of the well-studied quadratic assignment problem (QAP). We focused on sparse network alignment, where each node can be mapped only to a typically small subset of nodes in the other network. This corresponds to a QAP instance with a symmetric and sparse weight matrix. We obtained strong upper and lower bounds for the problem by improving a Lagrangian relaxation approach and introduce the open source software tool NATALIE 2.0, a publicly available implementation of our method (https://github.com/ls-cwi/natalie). In an extensive computational study on protein interaction networks for six different species, we found that our new method outperforms alternative established and recent state-of-the-art methods.

Integrative network analysis methods provide robust interpretations of differential high-throughput molecular profile measurements. They are often used in a biomedical context-to generate novel hypotheses about the underlying cellular processes or to derive biomarkers for classification and subtyping. The underlying molecular profiles are frequently measured and validated on animal or cellular models. Therefore the results are not immediately transferable to human. In particular, this is also the case in a study of the recently discovered interleukin-17 producing helper T cells (Th17), which are fundamental for anti-microbial immunity but also known to contribute to autoimmune diseases. We proposed a mathematical model for finding active subnetwork modules that are conserved between two species [19]. These are sets of genes, one for each species, which (1) induce a connected subnetwork in a species-specific interaction network, (2) show overall differential behaviour and (iii) contain a large number of orthologous genes. We proposed a flexible notion of conservation, which turns out to be crucial for the quality of the resulting modules in terms of biological interpretability. We developed an algorithm that finds provably optimal or near-optimal conserved active modules in our model. We applied our algorithm to understand the mechanisms underlying Th17 T cell differentiation in both mouse and human. As a main biological result, we found that the key regulation of Th17 differentiation is conserved between human and mouse.

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

Computationally investigating co-phylogenetic reconstructions and co-evolution
Despite an increasingly vast literature on co-phylogenetic reconstructions for studying host-symbiont associations, understanding the common evolutionary history of such systems remains a problem that is far from being solved. Most algorithms for host-symbiont reconciliation use an event-based model, where the events include in general (a subset of) co-speciation, duplication, loss, and host-switch. All known parsimonious event-based methods then assign a cost to each type of event in order to find a reconstruction of minimum cost. The main problem with this approach is that the cost of the events strongly influences the reconciliation obtained.To deal with this problem, we developed an algorithm, called COALA, for estimating the frequency of the events based on an approximate Bayesian computation approach [4]. The benefits of this method are twofold: (1) it provides more confidence in the set of costs to be used in a reconciliation, and (2) it allows estimation of the frequency of the events in cases where the dataset consists of trees with a large number of taxa. We evaluated our method on simulated and on biological datasets. We showed that in both cases, for the same pair of host and parasite trees, different sets of frequencies for the events lead to equally probable solutions. Moreover, often these solutions differ greatly in terms of the number of inferred events. It appears crucial to take this into account before attempting any further biological interpretation of such reconciliations. More generally, we also showed that the set of frequencies can vary widely depending on the input host and parasite trees. Indiscriminately applying a standard vector of costs may thus not be a good strategy. This work had been indicated as submitted in 2014.

Once such a cost vector has been inferred, one can proceed analysing the possible co-evolution of host-symbiont associations, phylogenetic tree reconciliation is the approach of choice for investigating the co-evolution of sets of organisms such as hosts and parasites. It consists in a mapping between the parasite tree and the host tree using event-based maximum parsimony. Given a cost model for the events, many optimal reconciliations are however possible. Only two algorithms existed that attempted such enumeration; in one case not all possible solutions are produced while in the other not all cost vectors are currently handled. We developed a polynomial-delay algorithm, EUCALYPT, for enumerating all optimal reconciliations that address these two issues [15]. We showed that in general many solutions exist. We gave an example where, for two pairs of host-parasite trees having each less than 41 leaves, the number of solutions is 5120, even when only time-feasible ones are kept. To facilitate their interpretation, those solutions are also classified in terms of how many of each event they contain. The number of different classes of solutions may thus be notably smaller than the number of solutions, yet they may remain high enough, in particular for the cases where losses have cost 0. In fact, depending on the cost vector, both numbers of solutions and of classes thereof may increase considerably (for the same instance, to respectively 4080384 and 275). To further deal with this problem, we introduced and analysed a restricted version where host-switches are allowed to happen only between species that are within some fixed distance along the host tree. This restriction allowed us to reduce the number of time-feasible solutions while preserving the same optimal cost, as well as to find time-feasible solutions with a cost close to the optimal in the cases where no time-feasible solution is found. This work had been indicated as submitted in 2014.

Evolution and metabolic complementation of organisms leaving inside the cells of another (endosymbionts)

Insect cells host many endosymbiotic bacteria, which are in general classified according to their importance for the host: "primary" symbionts are by definition mandatory and synthesise essential nutrients for the insects that feed on poor or unbalanced food sources, while "secondary" symbionts are optional and use mutualistic strategies and/or manipulation of reproduction to invade and persist within insect populations.

Hamiltonella defensa is a secondary endosymbiont that established two distinct associations with phloemophagous insects. In aphids, it protects the host against parasitoid attacks. Its ability to infect many host tissues, notably the hemolymph, could promote its contact with parasitoid eggs. Despite this protective phenotype, the high costs associated with its presence within the host prevent its fixation in the population. In the whitefly *Bemisia tabaci* however, this symbiont is found only in cells specialised in hosting endosymbionts, the bacteriocytes. In these cells, it cohabits with other symbiotic species, such as the primary symbiont *Portiera aleyrodidarum*, a proximity that favours potential exchanges between the two symbionts. It is fixed in populations of *B. tabaci*, which suggests an important role for the consortium, probably nutritious.

We studied the specificities of each of these systems [27]. First, in the bacteriocytes of *B. tabaci*, we identified a partitioning of the synthetic capacities of two endosymbionts, *H. defensa* and *P. aleyrodidarum*, in addition to a potential metabolic complementation between the symbionts and their host for the synthesis of essential amino acids. We proposed a key nutritive role for *H. defensa*, which would indicate a transition to a mandatory status in relation to the host and would explain its fixation in the population.

We also focused on the genomic evolution of the genus *Hamiltonella*, by comparing the strains infecting *B. tabaci* with a strain infecting the aphids [29]. We highlighted the specialisation of the symbionts to their hosts, and found that the genomes of the endosymbionts reflected their respective ecology. The aphid strain thus possesses many virulence factors and is associated with two partners, a bacteriophage and a recombination plasmid. These systems, inactive in the symbiont of *B. tabaci*, are directly related to the protection against and arms race with parasitoids. Conversely, the presumed avirulence of whitefly endosymbionts is consistent with their nutritional phenotype and a transition to a mandatory status to the host.

Finally, we studied the phenomenon of "accelerated mutation rate" in *H. defensa*, compared to its sister species *Regiella insecticola*, which is also a clade of protective endosymbionts of aphids. After excluding the assumption that the transition to the intracellular life occurred independently in the two lineages, we tried to establish a link between these differences in terms of evolvability in the endosymbionts and of their gene contents, particularly for genes involved in ecology and DNA repair. All the results obtained have provided

insight into the evolution of the species *H. defensa*, since the last ancestor to the present species, by establishing a link between bacterial.

These results were part of the PhD of Pierre-Antoine Rollat-Farnier, co-supervised by Laurence Mouton (LBBE, UMR5558), Marie-France Sagot (Inria and LBBE, UMR5558) and Fabrice Vavre (LBBE, UMR5558) and defended on November 24th, 2014. The results had been indicated as submitted in 2014.

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. We performed whole-genome metabolic network reconstructions for these three mycoplasmas and were able to show that overall they have similar metabolic capabilities. The metabolic differences that were observed include a wider range of carbohydrate uptake in *M. hyorhinis*, which in turn may also explain why this species is a widely known contaminant in cell cultures. Moreover, the myo-inositol catabolism is exclusive to *M. hyopneumoniae* and may be an important trait for virulence. However, the most important difference seems to be related to glycerol conversion to dihydroxyacetone-phosphate, which produces toxic hydrogen peroxide. This activity, missing only in *M. flocculare*, may be directly involved in cytotoxicity, as already been described for two lung pathogenic mycoplasmas, namely *Mycoplasma pneumoniae* in human and *Mycoplasma mycoides* subsp. mycoides in ruminants. Metabolomic data suggest that even though these mycoplasmas are extremely similar in terms of their genome and metabolism, different products and reaction rates may be the result of differential expression in each of them. We were able to infer from the reconstructed networks that the lack of pathogenicity of *M. flocculare* if compared to the highly pathogenic *M. hyopneumoniae* may be related to its incapacity to produce cytotoxic hydrogen peroxide. Moreover, the ability of *M. hyorhinis* to grow in diverse sites and even in different hosts may be a reflection of its enhanced and wider carbohydrate uptake. Altogether, the metabolic differences highlighted in silico and in vitro provide important insights to the different levels of pathogenicity observed in each of the studied species.

These results were part of the PhD of Mariana Galvão Ferrarini, co-supervised by Arnaldo Zaha (Federal University of Rio Grande do Sul and Marie-France Sagot (Inria and LBBE, UMR5558). and defended on December 10th, 2015. These results have been submitted to a journal. The PhD manuscript will be made available in HAL in early 2016.

6.5. Cross-fertilising different computational approaches

Tree matching
We considered the following problem related to tree matching, that we called the Tree-Constrained Bipartite Matching problem. Given a bipartite graph $G = (V_1, V_2, E)$ with edge weights $w : E \to R^+$, a rooted tree T_1 on the set V_1 and a rooted tree T_2 on the set V_2, find a maximum weight matching M in G, such that none of the matched nodes is an ancestor of another matched node in either of the trees [8]. This generalisation of the classical bipartite matching problem appears, for example, in the computational analysis of live cell video data. We showed that the problem is APX-hard and thus, unless P = NP, disproved a previous claim that it is solvable in polynomial time. Furthermore, we gave a 2-approximation algorithm based on a combination of the local ratio technique and a careful use of the structure of basic feasible solutions of a natural LP-relaxation, which we also show to have an integrality gap of $2 - o(1)$. We then considered a natural generalisation of the problem, where trees are replaced by partially ordered sets (posets). We showed that the local ratio technique gives a $2k\sigma$- approximation for the k-dimensional matching generalisation of the problem, in which the maximum number of incomparable elements below (or above) any given element in each poset is bounded by σ. We finally gave an almost matching integrality gap example, and an inapproximability result showing that the dependence on σ is most likely unavoidable.

Graph measures

We proposed a new algorithm that computes the radius and the diameter of a weakly connected digraph $G = (V, E)$, by finding bounds through heuristics and improving them until they are validated [5]. Although the worst-case running time is $O(|V||E|)$, we experimentally showed that it performs much better in the case of real-world networks, finding the radius and diameter values after 10-100 BFSs instead of $|V|$ BFSs (independently of the value of |V|), and thus having running time $O(|E|)$ in practice. As far as we know, this is the first algorithm able to compute the diameter of weakly connected digraphs, apart from the naive algorithm, which runs in time $O(|V||E|)$ performing a BFS from each node. In the particular cases of strongly connected directed or connected undirected graphs, we compared our algorithm with known approaches by performing experiments on a dataset composed by several real-world networks of different kinds. These experiments showed that, despite its generality, the new algorithm outperforms all previous methods, both in the radius and in the diameter computation, both in the directed and in the undirected case, both in average running time and in robustness. Finally, as an application example, we used the new algorithm to determine the solvability over time of the "Six Degrees of Kevin Bacon" game, and of the "Six Degrees of Wikipedia" game. As a consequence, we computed for the first time the exact value of the radius and the diameter of the whole Wikipedia digraph.

The closeness and the betweenness centralities are two well-known measures of importance of a vertex within a given complex network. Having high closeness or betweenness centrality can have positive impact on the vertex itself: hence, we considered the problem of determining how much a vertex can increase its centrality by creating a limited amount of new edges incident to it [40]. We first proved that this problem does not admit a polynomial-time approximation scheme (unless P=NP), and we then proposed a simple greedy approximation algorithm (with an almost tight approximation ratio), whose performance is then tested on synthetic graphs and real-world networks.

The (Gromov) hyperbolicity is a topological property of a graph, which has been recently applied in several different contexts, such as the design of routing schemes, network security, computational biology, the analysis of graph algorithms, and the classification of complex networks. Computing the hyperbolicity of a graph can be very time consuming: indeed, the best available algorithm has running-time $O(n^{3.69})$, which is clearly prohibitive for big graphs. We provided a new and more efficient algorithm: although its worst-case complexity is $O(n^4)$, in practice it is much faster, allowing, for the first time, the computation of the hyperbolicity of graphs with up to 200,000 nodes [36]. We experimentally showed that the new algorithm drastically outperforms the best previously available algorithms, by analyzing a big dataset of real-world networks. Finally, we applied the new algorithm to compute the hyperbolicity of random graphs generated with the Erdös-Renyi model, the Chung-Lu model, and the Configuration Model.

Hypergraph problems

It had been previously proved independently and with different techniques that there exists an incremental output polynomial algorithm for the enumeration of the minimal edge dominating sets in graphs, *i.e.*, minimal dominating sets in line graphs. We provided the first polynomial delay and polynomial space algorithm for the problem [42]. We proposed a new technique to enlarge the applicability of Berge's algorithm that is based on skipping hard parts of the enumeration by introducing a new search strategy. The new search strategy is given by a strong use of the structure of line graphs.

We also studied some average properties of hypergraphs and the average complexity of algorithms applied to hypergraphs under different probabilistic models [14]. Our approach is both theoretical and experimental since our goal is to obtain a random model that is able to capture the real-data complexity. Starting from a model that generalizes the Erdös-Renyi model and we obtain asymptotic estimations on the average number of transversals, irredundants and minimal transversals in a random hypergraph. We use those results to obtain an upper bound on the average complexity of algorithms to generate the minimal transversals of a hypergraph. Then we make our random model more complex in order to bring it closer to real-data and identify cases where the average number of minimal transversals is at most polynomial, quasi-polynomial or exponential.

The hypergraph transversal problem has been intensively studied, both from a theoretical and a practical point of view. In particular, its incremental complexity is known to be quasi-polynomial in general and polynomial for bounded hypergraphs. Recent applications in computational biology however require to solve a generalisation of this problem, that we call bi-objective transversal problem. The instance is in this case composed of a pair of hypergraphs (A, B), and the aim is to enumerate minimal sets which hit all the hyperedges of A while intersecting a minimal set of hyperedges of B. We formalised this problem and related it to the enumeration of minimal hitting sets of bundles [32]. We showed cases when under degree or dimension contraints, these problems remain NP-hard, and gave a polynomial algorithm for the case when A has bounded dimension, by building a hypergraph whose transversals are exactly the hitting sets of bundles.

7. Bilateral Contracts and Grants with Industry

7.1. Bilateral Contracts with Industry

From April 2013 to April 2015, N. Pisanti had a 60,000 euros + TVA grant from the private company Galileo Research srl for Scientific Counseling on "New technological Platform for Immunotherapy of cancers with synergic treatment".

8. Partnerships and Cooperations

8.1. National Initiatives

8.1.1. ANR

8.1.1.1. 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/

8.1.1.2. 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-Lafferière, H. Lopez-Maestre, C. Marchet, G. Sacomoto, M.-F. Sagot, B. Sinaimeri
- Type: ANR (2013-2016)
- Web page: http://colibread.inria.fr/

8.1.1.3. ExHyb

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

8.1.1.4. IMetSym

- Title: Immune and Metabolic Control in Intracellular Symbiosis of Insects
- Coordinator: A Heddi
- ERABLE participant(s): H. Charles, S. Colella
- Type: ANR Blanc (2014-2017)
- Web page: Not available

8.1.2. Others

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

8.1.2.1. Exomic

- Title: Functional annotation of the transcriptome at the exon level
- Coordinator: D. Auboeuf (Inserm, Lyon)
- ERABLE participant(s): V. Lacroix, M.-F. Sagot
- Type: INSERM Systems Biology Call (2012-2015)
- Web page: Not available

8.1.2.2. 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/research-units.php

8.1.2.3. 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

8.2. European Initiatives

8.2.1. FP7 & H2020 Projects

8.2.1.1. 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-Laferriére, 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/

8.2.1.2. 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/

8.2.1.3. SWIPE

- Title: Predicting whitefly population outbreaks in changing environments
- Duration: 2012 - 2015
- Coordinator: E. Zchori-Fein
- ERABLE participant(s): F. Vavre
- Web page: Not available

8.2.1.4. SISYPHE

- Title: Species Identity and SYmbiosis Formally and Experimentally explored
- Duration: 2010-2015 (ended March 31st)
- Coordinator: M.-F. Sagot
- BAMBOO participant(s): Whole BAMBOO team
- Type: ERC Advanced Grant
- Web page: http://team.inria.fr/erable/en/older-projects/erc-sisyphe/

8.2.2. Collaborations with Major European Organisations

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

8.3. International Initiatives

8.3.1. Inria International Labs

ERABLE participates in a project within the Inria-Chile CIRIC (Communication and Information Research and Innovation Center) titled "Omics Integrative Sciences". The main objectives of the project are the development and implementation of mathematical and computational methods and the associated computational platforms for the exploration and integration of large sets of heterogeneous omics data and their application to the production of biomarkers and bioidentification systems for important Chilean productive sectors. The project started in 2011 and is coordinated in Chile by Alejandro Maass, Mathomics, University of Chile, Santiago. It is in the context of this project that we are currently hosting the presence of 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.

8.3.2. 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/

8.3.3. Participation in other International Programs

ERABLE is coordinator of a CNRS-UCBL-Inria Laboratoire International Associé (LIA) with the Laboratório Nacional de Computação Científica (LNCC), Petrópolis, Brazil. The LIA has for acronym LIRIO ("Laboratoire International de Recherche en bIOinformatique") and is coordinated by Ana Tereza Vasconcelos from the LNCC and Marie-France Sagot from BAMBOO. 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/bamboo/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". 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. The project started in 2013 for 2 years, and then 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 had a Stic AmSud project accepted in 2015 that will start 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 tumor/normal samples and RNAseq of tumor samples). One of the scientific goals is to decipher whether the HER2/ERBB2 amplification is a driver or passenger event in the course of tumor development.

8.4. International Research Visitors

8.4.1. Visits of International Scientists

In 2015, ERABLE greeted the following International scientists:

- In France: Katharina Huber (University of Warwick, UK), Giuseppe Italiano (Tor Vergata University of Rome, Italy, various visits), Ana Rute Neves and Zeidan (ChR Hansen, Oslo, Danemark), three members of the LIA LIRIO (Arnaldo Zaha from the Federal University of Rio Grande do Sul, Maria Cristina Motta from the Federal University of Rio Grande do Sul, and Ana Tereza Vasconcelos from the LNCC, all in Brazil), Susana Vinga and various members of her team (IDMEC-IST Portugal), Tiziana Calamoneri (Sapienza University of Rome).
- In Italy: David Coudert (Inria Sophia Antipolis, France, to Florence), Alberto Policriti (University of Udine, Italy, to Pisa), Fabio Vandin (University of Southern Danemark to Pisa), Solon Pissis (King's College London UK to Pisa), Costas Iliopoulos (King's College London UK to Pisa), Grzegorz Rozenberg (Leiden University, The Netherlands, and Boulder University of Colorado, USA, to Pisa).
- In The Netherlands: Kirk Pruhs (University of Pittsburgh, USA), Kevin Schewior (Technical University of Berlin, Germany), Paola Bonizzoni, Yuri Pirola and Simone Zaccharia (all from the University of Milano-Bicocca, Italy).

8.4.2. Internships

In 2015, ERABLE greeted on average the following internship students:

- In France: Bastien Sylvere, Master 1 (2 months); Audric Cologne, Master 1 (3 months); Henri Dupoy, Master 1 (2 months); Virginie Jouffret, Master 1 (2 months); Caroline Michaud, Master 2 (6 months); Hong-Phong Pham, Master 2 (5 months); Nabel Sersoub, Master 1 (2 months); Manon Villa, Master 1 (2 months).
- In Italy: Anna Tarsia, Master 2 (Pisa).
- In The Netherlands: Gunnar Klau supervised a couple of MSc and BSc theses.

8.4.3. Visits to International Teams

8.4.3.1. Visits

In 2015, 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 Departement 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 the members of her team (IDMEC-IST Portugal).
- In Italy: visit to Pierre Fraigniaud and Michel Habib at LIAFA, Paris, visit to Solon Pissis and Costas Iliopoulos at King's College London UK.
- In The Netherlands: visit to the Technical University of Berlin, visit to Paola Bonizzoni and her group at the University of Milano-Bicocca.

8.4.3.2. Research stays abroad

Gunnar Klau will be spending 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.

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific events organisation

9.1.1.1. General chair, scientific chair

- Pierluigi Crescenzi was co-chair of the 16th Italian Conference on Theoretical Computer Science, 9-11 September 2015, Florence, Italy.
- 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. He was Chair of the 12th Workshop on Models and Algorithms for Planning and Scheduling Problems (MAPSP), 8-12 June 2015, La Roche-en-Ardenne, Belgium.
- Marie-France Sagot is since 2010 member and since 2014 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.

9.1.1.2. Member of the organising committees

- Leen Stougie was co-organiser with Nikhil Bansal (TU Eindhoven) and Sem Borst (TU Eindhoven) of the *EURANDOM/Networks* workshop Scheduling under Uncertainty, 1-5 June, 2015, EURANDOM, Eindhoven, The Netherlands.
- Fabrice Vavre organised the symposium "Evolution and ecology of trait loss and dependency" at the 15th Congress of the European Society for Evolutionary Biology, Lausanne, August 9-14, 2015 in collaboration with J. Ellers (Amsterdam).

9.1.2. Scientific events selection

9.1.2.1. Member of the conference program committee

- Pierluigi Crescenzi was a member of the program committee for the following international conferences in 2015: 30th IEEE International Parallel & Distributed Processing Symposium (IPDPS), 8th International Conference on Algorithms and Complexity (CIAC).

- Roberto Grossi was a member of the program committee for the following international conferences in 2015: 26th Annual Symposium on Combinatorial Pattern Matching (CPM), 26th International Workshop on Combinatorial Algorithms (IWOCA), ACM-SIAM Symposium on Discrete Algorithms (SODA).
- Alberto Marchetti-Spaccamela was a member of the program committee for the following international conferences in 2015: 40th International Symposium on Mathematical Foundations of Computer Science (MFCS), 14th International Symposium on Experimental Algorithms (SEA).
- Nadia Pisanti was a member of the program committee for the following international conferences in 2015: 11th International Symposium on Bioinformatics Research and Applications (ISBRA), 22nd International Symposium on String Processing and Information REtrieval (SPIRE), 4th International Symposium on Network Enabled Health Informatics, Biomedicine and Bioinformatics (Hi BI BI), 16th Italian Conference on Theoretical Computer Science (ICTCS), 15th Workshop on Algorithms in BIoinformatics (WABI), IEEE Conference on Information Reuse and Integration in Health Informatics (IEEE IRI-HI), 5th 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 2015: 8th International Conference on Algorithms and Complexity (CIAC), 26th Annual Symposium on Combinatorial Pattern Matching (CPM), 17th Portuguese Conference on Artificial Intelligence (EPIA), 23rd Annual European Symposium on Algorithms (ESA), 23rd Annual International Conference on Intelligent Systems in Molecular Biology (ISMB), 40th International Symposium on Mathematical Foundations of Computer Science (MFCS), The Prague Stringology Conference 2015, 19th Annual International Conference on Research in Computational Molecular Biology (RECOMB), 13th RECOMB Satellite Workshop on Comparative Genomics (RECOMB-CG), The 15th Workshop of Algorithms in Bioinformatics (WABI).
- Blerina Sinaimeri was a member of the program committee of the 10th International Workshop on Algorithms and Computation (WALCOM).

9.1.2.2. Reviewer

Besides the above, various other members of ERABLE have been reviewer for other international conferences such as FAW, IWOCA, and WADS.

9.1.3. Journal

9.1.3.1. 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 *Transaction 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*.

9.1.3.2. 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*.

9.1.4. Invited talks

Pierluigi Crescenzi gave three seminars (University of Padova; FSMP, Paris; Istituto Stensen, Florence); Leandro Ishi Soares de Lima gave a talk at the SeqBio Meeting in Paris; Gunnar Klau gave four invited lectures (CS Colloquium of the Heinrich Heine University Düsseldorf, Germany; Bertinoro Computational Biology 2015, Italy; Amsterdam Data Science Seminar, The Netherlands; Integrated Systems Biology Symposium at Maastricht, The Netherlands); Vincent Lacroix made two invited presentations in the context of training meetings (Formation Bioinformatique pour les NGS Montpellier; Formation FC3Bio Lyon); Alberto Marchetti-Spaccamela gave two seminars, one at TU Berlin, Germany, and the other at Laboratorio Nazionale CINI-InfoLife, Certosa di Calci, Italy; Nadia Pisanti gave four seminars (IMT Lucca; Laboratorio Nazionale CINI-InfoLife, Certosa di Calci, Italy; Glaxo-Novartis, Siena, Italy; Internet Festival 2015, Pisa, Italy); Blerina Sinaimeri gave three seminars (LIRMM, Montpellier, France; Tor Vergata University of Rome, Italy; Sapienza University of Rome, Italy); Leen Stougie gave an invited plenary lecture at the MOABI-workshop in Paris; Fabrice Vavre gave two invited talks (Symposium Communication between genomes, and Symposium Mathematical modeling and new methods for dengue control, Paris in both cases) and one seminar (University of Groningen, The Netherlands).

9.1.5. Leadership within the scientific community

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

Leen Stougie is Chairman of the Dutch Network on the Mathematics of Operations Research (Landelijk Netwerk Mathematische Besliskunde (LNMB)) 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/).

9.1.6. Scientific expertise

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

9.1.7. 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; she was in 2015 member of the hiring committee of 2 Researcher positions at the University of Pisa.

Marie-France Sagot was until December 2015 member of the Scientific Advisory Board (“Conseil Scientifique (COS)”) for the Inria Grenoble Rhône-Alpes Research Center.

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.

9.2. Teaching - Supervision - Juries

9.2.1. 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) and at the Insa (M1: Gene Expression, M2: Introduction to Bioinformatics for Biochemists). He teaches 192 hours per year in bioinformatics and statistics.
  - Arnaud Mary taught 110 hours in 2015 as a recently recruited personnel of the University of Lyon (L1: mathematics; L2: bioinformatics; M1: data analysis; M1: computer science, as well as 8h at the M2 in Computer Science at the ENS Lyon).
  - 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:
  - 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.
- In The Netherlands:
  - Gunnar Klau taught 40h (BSc course in Algorithm Engineering) at the Free University of Amsterdam, as well as one day of a two-day summer school for PhD students and postdocs on Biological Network Analysis at Free University Medical Center Amsterdam.

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 80 hours of Applied Mathematics and Bioinformatics at the Department of Biology (undergraduate students); 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); Marie-France Sagot (Senior Inria Researcher) taught 6 hours at the Master 2 in Computer Science at the ENS Lyon; Blerina Sinaimeri (Junior Inria Researcher) taught 8h in Discrete Mathematics at the Master of Modelling and Bioinformatics (BIM), INSA, and Master 1 MIV, University Lyon 1, as well as 10h 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).

9.2.2. Supervision

The following are the PhDs defended in ERABLE in 2015.

- Sandro Andreotti, FU Berlin, February 2015, supervisor: G. Klau.
- Mohammed El-Kebir, VU University Amsterdam, October 2015, supervisor: G. Klau.
- Mariana Galvão Ferrarini, University of Lyon 1, December 2015, supervisors: M.-F. Sagot and A. Zaha.
- Nela Lekic, University of Maastricht, December 2015, supervisor: L. Stougie, R. Peeters, S. Kelk.

9.2.3. Juries

The following are the PhD or HDR juries to which members of ERABLE participated in 2015.

- Gunnar Klau: Reviewer of the PhDs of Timo Maarleveld, VU Amsterdam, The Netherlands; Kasper Dinkla, TU Eindhoven, The Netherlands; Martina Summer-Kutmon, Maastricht University, The Netherlands; Daniel Taliun, University of Bozen, Italy.
- Nadia Pisanti: Reviewer of the PhD of Michele Schimd, University of Padova, Italy.
- Leen Stougie: Member of the PhD committee of Ward Romeijnders, University of Groningen, The Netherlands; Mohammed El-Bekir, Free University of Amsterdam, The Netherlands.
- Fabrice Vavre: Reviewer of the PhDs of S. Gerritsma, Groningen University, The Netherlands, and Z.S. Wong, University of Quennsland, Australia; member of the PhD committee of R. Stalinski, University of Grenoble Alpes, France; reviewer of the HDR of O. Kaltz, University of Montpellier 2, France; member of the HDR committee of F. Dedeine, University of Tours, France.
- Alain Viari: Reviewer of the HDR of Sarah Cohen-Boulakia and member of the HDR committee of Julie Bernauer, both University of Paris 11, France.

9.3. Popularisation

Laura Urbini and Blerina Sinaimeri participated at the Fête de la Science of Inria. The title of the workshop they presented was: “Du passé au présent : explorons l’évolution”. More information may be found at this address: http://www.inria.fr/centre/grenoble/actualites/fete-de-la-science-les-coulisses-du-numerique.

Fabrice Vavre participated to the “La foire aux savoirs”, Les Subsistances, Lyon on the topic of “Le savoir-faire de nos bactéries”, and to the "Université Ouverte" of the University Lyon 1 on the topic of “Quoi de neuf sur nos relations avec les bactéries”?

10. Bibliography

Publications of the year

Articles in International Peer-Reviewed Journals

[1] R. ANDONOV, H. DJIDJEV, G. KLAU, M. BOUDIC-JAMIN, I. WOHLERS. *Automatic Classification of Protein Structure Using the Maximum Contact Map Overlap Metric*, in "Algorithms and Combinatorics", December 2015, vol. 8, n^{o} 4 [*DOI :* 10.3390/A8040850], https://hal.inria.fr/hal-01248543

[2] A. C. AZEVEDO-MARTINS, A. C. L. MACHADO, C. C. KLEIN, L. CIAPINA, L. GONZAGA, A. T. R. VASCONCELOS, M.-F. SAGOT, W. DE SOUZA, M. EINICKER-LAMAS, A. GALINA, M. C. M. MOTTA. *Mitochondrial respiration and genomic analysis provide insight into the influence of the symbiotic bacterium on host trypanosomatid oxygen consumption*, in "Parasitology", 2015, vol. 142, n^{o} 2, pp. 352-362 [*DOI :* 10.1017/S0031182014001139], https://hal.inria.fr/hal-01073754

[3] S. BARUAH, V. BONIFACI, G. D'ANGELO, H. LI, A. MARCHETTI-SPACCAMELA, S. VAN DER STER, L. STOUGIE. *Preemptive Uniprocessor Scheduling of Mixed-Criticality Sporadic Task Systems*, in "Journal of the ACM (JACM)", May 2015, vol. 62, n° 2, 14 p. [*DOI :* 10.1145/2699435], https://hal.inria.fr/hal-01249091

[4] C. BAUDET, B. DONATI, B. SINAIMERI, P. CRESCENZI, C. GAUTIER, C. MATIAS, M.-F. SAGOT. *Cophylogeny Reconstruction via an Approximate Bayesian Computation*, in "Systematic Biology", 2015, vol. 64, n° 3, pp. 416-431 [*DOI :* 10.1093/SYSBIO/SYU129], https://hal.inria.fr/hal-01092972

[5] M. BORASSI, P. CRESCENZI, M. HABIB, W. A. KOSTERS, A. MARINO, F. W. TAKES. *Fast diameter and radius BFS-based computation in (weakly connected) real-world graphs*, in "Journal of Theoretical Computer Science (TCS)", June 2015, vol. 586, pp. 59–80 [*DOI :* 10.1016/J.TCS.2015.02.033], https://hal.inria.fr/hal-01248555

[6] E. BOSI, B. DONATI, M. GALARDINI, S. BRUNETTI, M.-F. SAGOT, P. LIÓ, P. CRESCENZI, R. FANI, M. FONDI. *MeDuSa: A multi-draft based scaffolder*, in "Bioinformatics (Oxford, England)", March 2015, pii: btv171 [Epub ahead of print], https://hal.inria.fr/hal-01139506

[7] L. BULTEAU, G. SACOMOTO, B. SINAIMERI. *Computing an Evolutionary Ordering is Hard*, in "Electronic Notes in Discrete Mathematics", 2015 [*DOI :* 10.1016/J.ENDM.2015.07.043], https://hal.archives-ouvertes.fr/hal-01249259

[8] S. CANZAR, K. ELBASSIONI, G. W. KLAU, J. MESTRE. *On Tree-Constrained Matchings and Generalizations*, in "Algorithmica", January 2015, vol. 71, n° 1 [*DOI :* 10.1007/S00453-013-9785-0], https://hal.inria.fr/hal-01248542

[9] R. CIJVAT, S. MANEGOLD, M. KERSTEN, G. W. KLAU, A. SCHÖNHUTH, T. MARSCHALL, Y. ZHANG. *Genome sequence analysis with MonetDB - A case study on Ebola virus diversity*, in "Datenbank-Spektrum", November 2015, vol. 15, n° 3 [*DOI :* 10.1007/S13222-015-0198-X], https://hal.inria.fr/hal-01248546

[10] J. CORREA, A. MARCHETTI-SPACCAMELA, J. MATUSCHKE, L. STOUGIE, O. SVENSSON, V. VERDUGO, J. VERSCHAE. *Strong LP formulations for scheduling splittable jobs on unrelated machines*, in "Mathematical Programming", December 2015, vol. 154, n° 1-2, pp. 305-328 [*DOI :* 10.1007/S10107-014-0831-8], https://hal.inria.fr/hal-01249090

[11] P. CRESCENZI, D. GILDEA, A. MARINO, G. ROSSI, G. SATTA. *Synchronous context-free grammars and optimal linear parsing strategies*, in "Izvestia Rossiiskoi Akademii Nauk.Teoriya i Systemy Upravleniya / Journal of Computer and Systems Sciences International", November 2015, vol. 81, n° 7, pp. 1333-1356 [*DOI :* 10.1016/J.JCSS.2015.04.003], https://hal.inria.fr/hal-01249108

[12] M. CROCHEMORE, R. GROSSI, J. KÄRKKÄINEN, G. M. LANDAU. *Computing the Burrows–Wheeler transform in place and in small space*, in "Journal of Discrete Algorithms", May 2015, vol. 32, pp. 44-52 [*DOI :* 10.1016/J.JDA.2015.01.004], https://hal.inria.fr/hal-01248855

[13] P. DUMAS, F. LEGEAI, C. LEMAITRE, E. SCAON, M. ORSUCCI, K. LABADIE, S. GIMENEZ, A. L. CLAMENS, H. HENRI, F. VAVRE, J. M. AURY, P. FOURNIER, G. KERGOAT, E. D'ALENÇON. *Spodoptera frugiperda (Lepidoptera: Noctuidae) host-plant variants: two host strains or two distinct species?*, in "Genetica", 2015, vol. 143, n° 3, pp. 305-316 [*DOI :* 10.1007/S10709-015-9829-2], https://hal.archives-ouvertes.fr/hal-01208780

[14] J. DAVID, L. LHOTE, A. MARY, F. RIOULT. *An average study of hypergraphs and their minimal transversals*, in "Journal of Theoretical Computer Science (TCS)", September 2015, vol. 596, pp. 124-141 [*DOI :* 10.1016/J.TCS.2015.06.052], https://hal.archives-ouvertes.fr/hal-01086638

[15] B. DONATI, C. BAUDET, B. SINAIMERI, P. CRESCENZI, M.-F. SAGOT. *EUCALYPT: efficient tree reconciliation enumerator*, in "Algorithms for Molecular Biology", January 2015, vol. 10, n^{o} 1, 11 p. [*DOI :* 10.1186/S13015-014-0031-3], https://hal.inria.fr/hal-01092977

[16] O. DURON, V. NOËL, K. D. MCCOY, M. BONAZZI, K. SIDI-BOUMEDINE, O. MOREL, F. VAVRE, L. ZENNER, E. JOURDAIN, P. DURAND, C. ARNATHAU, F. RENAUD, J.-F. TRAPE, A. S. BIGUEZOTON, J. CREMASCHI, M. DIETRICH, E. LÉGER, A. APPELGREN, M. DUPRAZ, E. GÓMEZ-DÍAZ, G. DIATTA, G.-K. DAYO, H. ADAKAL, S. ZOUNGRANA, L. VIAL, C. CHEVILLON. *The Recent Evolution of a Maternally-Inherited Endosymbiont of Ticks Led to the Emergence of the Q Fever Pathogen, Coxiella burnetii*, in "PLoS Pathogens", May 2015, vol. 11, n^{o} 5 [*DOI :* 10.1371/JOURNAL.PPAT.1004892], https://hal.inria.fr/hal-01250470

[17] M. DYER, L. STOUGIE. *Erratum to: Computational complexity of stochastic programming problems*, in "Mathematical Programming", November 2015, vol. 153, n^{o} 2, pp. 723-725 [*DOI :* 10.1007/S10107-015-0935-9], https://hal.inria.fr/hal-01249093

[18] M. EL-KEBIR, J. HERINGA, G. KLAU. *Natalie 2.0: Sparse Global Network Alignment as a Special Case of Quadratic Assignment*, in "Algorithms and Combinatorics", December 2015, vol. 8, n^{o} 4 [*DOI :* 10.3390/A8041035], https://hal.inria.fr/hal-01248544

[19] M. EL-KEBIR, H. SOUEIDAN, T. HUME, D. BEISSER, M. DITTRICH, T. MÜLLER, G. BLIN, J. HERINGA, M. NIKOLSKI, L. F. A. WESSELS, G. W. KLAU. *xHeinz: an algorithm for mining cross-species network modules under a flexible conservation model*, in "Bioinformatics", 2015, vol. 31, n^{o} 19 [*DOI :* 10.1093/BIOINFORMATICS/BTV316], https://hal.inria.fr/hal-01248545

[20] C. GOUBERT, L. MODOLO, C. VIEIRA, C. VALIENTEMORO, P. MAVINGUI, M. BOULESTEIX. *De novo assembly and annotation of the Asian tiger mosquito (Aedes albopictus) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (Aedes aegypti)*, in "Genome Biology and Evolution", March 2015, vol. 7, n^{o} 4, pp. 1192-205, https://hal.inria.fr/hal-01227710

[21] E. HELLARD, D. FOUCHET, F. VAVRE, D. PONTIER. *Parasite-Parasite Interactions in the Wild: How To Detect Them?*, in "Trends in Parasitology", August 2015, https://hal.inria.fr/hal-01227711

[22] S. HIGASHI, C. FOURNIER, C. GAUTIER, C. GASPIN, M.-F. SAGOT. *Mirinho: An efficient and general plant and animal pre-miRNA predictor for genomic and deep sequencing data*, in "BMC Bioinformatics", 2015, vol. 16, n^{o} 1, 179 p. [*DOI :* 10.1186/S12859-015-0594-0], https://hal.archives-ouvertes.fr/hal-01166487

[23] A. MARCHETTI-SPACCAMELA, C. RUTTEN, S. VAN DER STER, A. WIESE. *Assigning sporadic tasks to unrelated machines*, in "Mathematical Programming", August 2015, vol. 152, n^{o} 1-2, pp. 247-274 [*DOI :* 10.1007/S10107-014-0786-9], https://hal.inria.fr/hal-01249101

[24] L. MOUTON, O. GNANKINÉ, H. HENRI, G. TERRAZ, G. KETOH, T. MARTIN, F. FLEURY, F. VAVRE. *Detection of genetically isolated entities within the Mediterranean species of Bemisia tabaci: new insights into the systematics of this worldwide pest*, in "Pest Management Science", March 2015, vol. 71, n^{o} 3, pp. 452-458 [*DOI :* 10.1002/PS.3834], https://hal.inria.fr/hal-01250472

[25] M. PATTERSON, T. MARSCHALL, N. PISANTI, L. VAN IERSEL, L. STOUGIE, G. W. KLAU, A. SCHÖNHUTH. *WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads*, in "Journal of computational biology : a journal of computational molecular cell biology", May 2015, vol. 22, n° 6, pp. 498-509 [*DOI :* 10.1089/CMB.2014.0157], https://hal.inria.fr/hal-01225988

[26] 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", 2015 [*DOI :* 10.1093/BIOINFORMATICS/BTV495], https://hal.inria.fr/hal-01225984

[27] Q. RAO, P.-A. ROLLAT-FARNIER, D.-T. ZHU, D. SANTOS-GARCIA, F. J. SILVA, A. MOYA, A. LATORRE, C. C. KLEIN, F. VAVRE, M.-F. SAGOT, S.-S. LIU, L. MOUTON, X.-W. WANG. *Genome reduction and potential metabolic complementation of the dual endosymbionts in the whitefly Bemisia tabaci*, in "BMC Genomics", December 2015, vol. 16, 226 p. [*DOI :* 10.1186/S12864-015-1379-6], https://hal.inria.fr/hal-01139508

[28] A. C. REIMERS, F. J. BRUGGEMAN, B. G. OLIVIER, L. STOUGIE. *Fast Flux Module Detection Using Matroid Theory*, in "Journal of Computational Biology", May 2015, vol. 22, n° 5 [*DOI :* 10.1089/CMB.2014.0141], https://hal.inria.fr/hal-01227722

[29] P.-A. ROLLAT-FARNIER, D. SANTOS-GARCIA, Q. RAO, M.-F. SAGOT, F. J. SILVA, H. HENRI, E. ZCHORI-FEIN, A. LATORRE, A. MOYA, V. BARBE, S.-S. LIU, X.-W. WANG, F. VAVRE, L. MOUTON. *Two Host Clades, Two Bacterial Arsenals: Evolution through Gene Losses in Facultative Endosymbionts*, in "Genome Biology and Evolution", 2015, vol. 7, n° 3, pp. 839-855, https://hal.inria.fr/hal-01139507

[30] G. SACOMOTO, V. LACROIX, M. SAGOT. *A polynomial delay algorithm for the enumeration of bubbles with length constraints in directed graphs*, in "Algorithms for Molecular Biology", 2015, 10 p. [*DOI :* 10.1186/S13015-015-0046-4], https://hal.inria.fr/hal-01170399

[31] F. SCHALEKAMP, R. SITTERS, S. VAN DER STER, L. STOUGIE, V. VERDUGO, A. VAN ZUYLEN. *Split scheduling with uniform setup times*, in "Journal of Scheduling", April 2015, vol. 18, n° 2, pp. 119-129 [*DOI :* 10.1007/S10951-014-0370-4], https://hal.inria.fr/hal-01249095

International Conferences with Proceedings

[32] R. ANDRADE, E. BIRMELÉ, A. MARY, T. PICCHETTI, M.-F. SAGOT. *Incremental complexity of a bi-objective hypergraph transversal problem*, in "Fundamentals of Computation Theory (FCT2015)", Gdansk, Poland, Fundamentals of Computation Theory (FCT2015), 2015, vol. 9210, https://hal.archives-ouvertes.fr/hal-01149392

[33] S. K. BARUAH, V. BONIFACI, A. MARCHETTI-SPACCAMELA. *The Global EDF Scheduling of Systems of Conditional Sporadic DAG Tasks*, in "ECRTS 2015 - Euromicro Conference on Real-Time Systems", Lund, Sweden, July 2015, pp. 222-231 [*DOI :* 10.1109/ECRTS.2015.27], https://hal.inria.fr/hal-01249105

[34] L. BETTINI, P. CRESCENZI. *Java-Meets Eclipse - An IDE for Teaching Java Following the Object-later Approach*, in "International Conference on Software Paradigm Trend (ICSOFT-PT)", Colmar, France, September 2015, pp. 31-42 [*DOI :* 10.5220/0005512600310042], https://hal.inria.fr/hal-01249109

[35] P. BONIZZONI, R. DONDI, G. W. KLAU, Y. PIROLA, N. PISANTI, S. ZACCARIA. *On the Fixed Parameter Tractability and Approximability of the Minimum Error Correction Problem*, in "26th Annual Symposium

on Combinatorial Pattern Matching (CPM)", Ischia, Italy, June 2015 [*DOI :* 10.1007/978-3-319-19929-0], https://hal.inria.fr/hal-01246260

[36] M. BORASSI, D. COUDERT, P. CRESCENZI, A. MARINO. *On Computing the Hyperbolicity of Real-World Graphs*, in "23rd Annual European Symposium on Algorithms (ESA)", Patras, Greece, Lecture Notes in Computer Science, Springer, September 2015, vol. 9294, pp. 215-226 [*DOI :* 10.1007/978-3-662-48350-3_19], https://hal.inria.fr/hal-01199860

[37] S. CANZAR, A. SANDRO, D. WEESE, R. KNUT, K. GUNNAR. *CIDANE: Comprehensive Isoform Discovery and Abundance Estimation*, in "Research in Computational Molecular Biology (RECOMB)", Warsaw, Poland, April 2015, vol. 9029, pp. 60-84 [*DOI :* 10.1007/978-3-319-16706-0_8], https://hal.inria.fr/hal-01248849

[38] L. CHEN, N. MEGOW, R. RISCHKE, L. STOUGIE. *Stochastic and Robust Scheduling in the Cloud*, in "Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM)", Princeton, United States, August 2015, pp. 175-186 [*DOI :* 10.4230/LIPICS.APPROX-RANDOM.2015.175], https://hal.inria.fr/hal-01249100

[39] L. CHEN, N. MEGOW, R. RISCHKE, L. STOUGIE, J. VERSCHAE. *Optimal Algorithms and a PTAS for Cost-Aware Scheduling*, in "Mathematical Foundations of Computer Science (MFCS)", Milan, Italy, August 2015, vol. 9235, pp. 211-222 [*DOI :* 10.1007/978-3-662-48054-0_18], https://hal.inria.fr/hal-01249098

[40] P. CRESCENZI, G. D'ANGELO, S. LORENZO, V. YLLKA. *Greedily Improving Our Own Centrality in A Network*, in "Experimental Algorithms - SEA", Paris, France, June 2015, vol. 9125, pp. 43-55 [*DOI :* 10.1007/978-3-319-20086-6_4], https://hal.inria.fr/hal-01248558

[41] R. GROSSI, C. S. ILIOPOULOS, R. MERCAS, N. PISANTI, S. PISSIS, A. RETHA, F. VAYANI. *Circular Sequence Comparison with q-grams*, in "15th International Conference on Algorithms in BioInformatics (WABI)", Atlanta, United States, September 2015 [*DOI :* 10.1007/978-3-662-48221-6_15], https://hal.inria.fr/hal-01246271

[42] M. M. KANTÉ, V. LIMOUZY, A. MARY, L. NOURINE, T. UNO. *Polynomial Delay Algorithm for Listing Minimal Edge Dominating Sets in Graphs*, in "Algorithms and Data Structures (WADS)", Victoria, Canada, August 2015, vol. 9214, pp. 446-457 [*DOI :* 10.1007/978-3-319-21840-3_37], https://hal.inria.fr/hal-01248851

[43] A. MELANI, M. BERTOGNA, V. BONIFACI, A. MARCHETTI-SPACCAMELA, G. BUTTAZZO. *Memory-processor co-scheduling in fixed priority systems*, in "International Conference on Real Time and Networks Systems (RTNS)", Lille, France, November 2015, pp. 87-96 [*DOI :* 10.1145/2834848.2834854], https://hal.inria.fr/hal-01249107

[44] A. MELANI, V. BONIFACI, M. BERTOGNA, A. MARCHETTI-SPACCAMELA, G. BUTTAZZO. *Response-Time Analysis of Conditional DAG Tasks in Multiprocessor Systems*, in "ECRTS 2015 - Euromicro Conference on Real-Time Systems", Lund, Sweden, July 2015, pp. 211-221 [*DOI :* 10.1109/ECRTS.2015.26], https://hal.inria.fr/hal-01249103

[45] M. F. ZINI, N. PISANTI, E. BIASCI, A. PODDA, V. MEY, F. PIRAS, G. L'ABBATE, S. MARINI, D. FRATTA, S. BONARETTI, S. TRASCIATTI. *Preclinical Tests for Cerebral Stroke*, in "The 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)", Paris, France, August 2015 [*DOI :* 10.1145/2808797.2808816], https://hal.inria.fr/hal-01246340