IN PARTNERSHIP WITH:
CNRS

Institut national des sciences appliquées de Lyon

Université Claude Bernard (Lyon 1)

Activity Report 2014

Project-Team BAMBOO

An algorithmic view on genomes, cells, and environments

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

RESEARCH CENTER
Grenoble - Rhône-Alpes

THEME
Computational Biology

Table of contents

Project-Team BAMBOO

Keywords: Computational Biology, Systems Biology, Analysis Of Algorithms, Genomics, Network Modeling

BAMBOO will cease to exist at the end of 2014. In 2015, the team will be re-created, first as an Inria "Center team" and then as an "Inria project-team" called ERABLE. ERABLE will be a European Inria team gathering the current members of BAMBOO, together with four researchers in Italy under the banner of the University of Rome La Sapienza (Alberto Marchetti-Spaccamela from La Sapienza, Pierluigi Crescenzi from the University of Florence, Roberto Grossi and Nadia Pisanti from the University of Pisa), and two researchers in the Netherlands under the banner of the CWI (Leen Stougie from the Free University of Amsterdam and the CWI, Gunnar Klau from the CWI).

Creation of the Team: 2009 January 01*, updated into Project-Team:* 2012 January 01, end of the Project-Team: 2014 December 31.

1. Members

Research Scientists

Marie-France Sagot [Team leader, Inria, Senior Researcher, HdR]
Fabrice Vavre [CNRS, Researcher, HdR]
Alain Viari [Inria, Senior Researcher & Deputy Scientific Director for ICST for Life and Environmental Sciences at Inria]

Faculty Members

Hubert Charles [INSA Lyon, Associate Professor, HdR]
Christian Gautier [Univ. Lyon I, Professor, HdR]
Vincent Lacroix [Univ. Lyon I, Associate Professor]
Arnaud Mary [Univ. Lyon I, Associate Professor, from Sep 2014]
Cristina Vieira [Univ. Lyon I, Full Professor, HdR]

Engineer

Camille Marchet [Inria]

PhD Students

Beatrice Donati [Inria, until Nov 2014, PhD defended Nov 12, 2014]
Pierre-Antoine Farnier [Inria, until Jul 2014, PhD defended Nov 24, 2014]
Mariana Galvao Ferrarini [Inria, granted by European Research Council]
Susan Higashi [Inria, granted by European Research Council until Nov 2014, PhD defended Nov 26, 2014]
Alice Julien-Laferrière [Inria, granted by European Research Council]
Scheila Mucha [Capes-Cofecub, Brazilian Sandwich PhD for one year starting from Jul 2014]
Gustavo A. T. Sacomoto [Inria, granted by European Research Council, until Feb 2014]
Laura Urbini [Inria, from Oct 2014]
Martin Wannagat [Inria, granted by European Research Council]

Post-Doctoral Fellows

Christian Baudet [Inria, granted by European Research Council]
Laurent Bulteau [Inria, from Oct 2014, granted by FP7 KBBE project]
Emilie Chautard [Inria, granted by INSERM]
Ricardo de Andrade Abrantes [Science Without Frontiers, Ministry of Research Brazil, until March 2015]
Susan Higashi [Inria, granted by European Research Council from Dec 2014]
Delphine Parrot [Inria, granted by FP7 KBBE project, from Nov 2014]
Paulo Trenhago [Capes-Cofecub, for one year starting from Apr 2014]
Gustavo A. T. Sacomoto [Inria, granted by European Research Council]

Blerina Sinaimeri [Inria, granted by European Research Council]

Visiting Scientists

Leandro Ishi Soares de Lima [from Sep to Dec 2014]
Maria Cristina Machado Motta [May 2014, Nov 2014]
Franciele Siqueira [May 2014]
Ana Tereza Vasconcelos [May 2014, Dec 2014]
Arnaldo Zaha [Dec 2014]

Administrative Assistant

Florence Bouheddi [Inria]

Others

Stefano Colella [INRA, Researcher, external collaborator]
Pierluigi Crescenzi [University of Florence, Italy, Full Professor, external collaborator]
Roberto Grossi [University of Pisa, Italy, Full Professor, external collaborator]
Laurent Jacob [CNRS & LBBE, Researcher, external collaborator]
Gunnar Klau [CWI, The Netherlands, Researcher, external collaborator]
Alberto Marchetti-Spaccamela [University or Rome La Sapienza, Italy, Full Professor, external collaborator]
Vincent Miele [CNRS & LBBE, Research engineer, external collaborator]
Anne Morgat [SIB Geneva, Researcher, external collaborator]
Nadia Pisanti [University of Pisa, Italy, Assistant Professor, external collaborator]
Maria Puig Lombardi [Inria, Traineeship Master 1, from Jun 2014 until Aug 2014]
Leen Stougie [Free University Amsterdam & CWI, The Netherlands, Full Professor, external collaborator]
Laura Urbini [UCBL, Traineeship Master 2, from Jan to Sept 2014]
Ana Tereza Vasconcelos [LNCC Brazil, Researcher, external collaborator, co-responsible for LIA LIRIO]

2. Overall Objectives

2.1. Overall Objectives

The study of symbiosis and of biological interactions more in general is the motivation for the work conducted within BAMBOO, but runs in parallel with another important objective. This concerns to (re)visit classical combinatorial (mainly counting / enumerating) and algorithmic problems on strings and (hyper)graphs, and to explore the new variants / original combinatorial and algorithmic problems that are raised by the main areas of application of this project. As the objectives of these formal methods are motivated by biological questions, they are briefly described together with those questions in the next section.

3. Research Program

3.1. Symbiosis

The study we propose to do on symbiosis decomposes into four main parts - (1) genetic dialog, (2) metabolic dialog, (3) symbiotic dialog and genome evolution, and (4) symbiotic dynamics - that are however strongly interrelated, and the study of such interrelations will represent an important part of our work. Another biological objective, larger and which we hope within the ERC project SISYPHE just to sketch for a longer term investigation, will aim at getting at a better grasp of species identity and of a number of identity-related concepts. We now briefly indicate the main points that have started been investigated or should be investigated in the next five years.

Genetic dialog

We plan to study the genetic dialog at the regulation level between symbiont and host by addressing the following mathematical and algorithmic issues:

1. model and identify all small RNAs from the bacterium and the host which may be involved in the genetic dialog between the two, and model/identify the targets of such small RNAs;
2. infer selected parts of the regulatory network of both symbiont and host (this will enable to treat the next point) using all available information;
3. explore at both the computational and experimental levels the complementarity of the two networks, and revisit at a network level the question of a regulatory response of the symbiont to its host's demand;
4. compare the complementarities observed between pairs of networks (the host's and the symbiont's); such complementarities will presumably vary with the different types of host-symbiont relationships considered, and of course with the information the networks model (structural or dynamic); Along the way, it may become important at some point to address also the issue of transposable elements (abbreviated into TEs, that are genes which can jump spontaneously from one site to another in a genome following or not a duplication event). It is increasingly believed that TEs play a role in the regulation of the expression of the genes in eukaryotic genomes. The same role in symbionts, and in the host-symbiont dialog has been less or not explored. This requires to address the following additional task:
5. accurately and systematically detect all transposable elements (*i.e.* genes which can jump spontaneously from one site to another in a genome following or not a duplication event) and assess their implication in their own regulation and that of their host genome (the new sequencing technologies should facilitate this task as well as other data expression analyses, if we are able to master the computational problem of analysing the flow of data they generate: fragment indexing, mapping and assembly);
6. where possible, obtain data enabling to infer the PPI (Protein-Protein Interaction) for hosts and symbionts, and at the host-symbiont interface and analyse the PPI networks obtained and how they interact.

Initial algorithmic and statistical approaches for the first two items above are under way and are sustained by a well-established expertise of the team on sequence and microarray bioinformatic analysis. Both problems are however notoriously hard because of the high level of missing data and noise, and of our relative lack of knowledge of what could be the key elements of genetic regulation, such as small and micro RNAs.

We also plan to establish the complete repertoire of transcription factors of the interacting partners (with possible exchanges between them) at both the computational and experimental levels. Comparative biology (search by sequence homology of known regulators), 3D-structural modelling of putative domains interacting with the DNA molecule, regulatory domains conserved in the upstream region of coding DNA are among classical and routinely used methods to search for putative regulatory proteins and elements in the genomes. Experimentally, the BiaCore (using the surface plasmon resonance principle) and ChIP-Seq (using chromatin precipitation coupled with high-throughput sequencing from Solexa) techniques offer powerful tools to capture all the protein-DNA interactions corresponding to a specific putative regulator. However, these techniques have not been evaluated in the context of interacting partners making this task an interesting challenge.

Metabolic dialog

Our main plan for this part, where we have already many results, some obtained this last year, is to:

1. continue with and improve our work on reconstructing the metabolic networks of organisms with sequenced genomes, taking in particular care to cover as much as possible the different types of hosts and symbionts in interaction;
2. refine the network reconstructions by using flux balance analysis which will in turn require addressing the next item;
3. improve our capacity to efficiently compute fluxes and do flux balance analysis; current algorithms can handle only relatively small networks;

4. analyse and compare the networks in terms of their general structural, quantitative and dynamic characteristics;
5. develop models and algorithms to compare different types of metabolic interfaces which will imply being able, by a joint computational and experimental approach, to determine what is transported across interacting metabolisms;
6. define what would be a good null hypothesis to test the statistical significance, and therefore possible biological relevance of the characteristics observed when analysing or comparing (random network problem, a mostly open issue despite the various models available);
7. use the results from item 5, that is indications on the precursors of a bacterial metabolism that are key players in the dialog with the metabolism of the host, to revisit the genetic regulation dialog between symbiont and host.

Computational results from the last item will be complemented with experiments to help understand what is transported from the host to the symbiont and how what is transported may be related with the genetic dialog between the two organisms (items 5 and 6).

Great care will also be taken in all cases (metabolism- or regulation-only, or both together) to consider the situations, rather common, where more than two partners are involved in a symbiosis, that is when there are secondary symbionts of a same host.

The first five items above have started being computationally explored by our team, as has the last item including experimentally. Some algorithmic proofs-of-concept, notably as concerns structural, flux, precursor and chemical organisation studies (see some of the publications of the last year and this one), have been established but much more work is necessary. The main difficulties with items 3 and 4 are of two sorts. The first one is a modelling issue: what are the best models for analysing and comparing two or more networks? This will greatly depend on the biological question put, whether evolutionary or functional, structural or physiologic, besides being a choice that should be motivated by the extent and quality of the data available. The second sort of difficulty ,which also applies to other items notably (item 2), is computational. Most of the problems related with analysing and specially comparing are known to be hard but many issues remain open. The question of a good random model (item 6) is also largely open.

Symbiotic dialog and genome evolution

Genomes are not static. Genes may get duplicated, sometimes the duplication affects the whole genome, or genes can transpose, while whole genomic segments can be reversed or deleted. Deletions are indeed one of the most common events observed for some symbionts. Genetic material may also be transferred across sub-species or species (lateral transfer), thus leading to the insertion of new elements in a genome. Finally, parts of a genome may be amplified through, for instance, slippage during DNA replication resulting in the multiplication of the copies of a repeat that appear tandemly arrayed along a genome. Tandem repeats, and other types of short or long repetitions are also believed to play a role in the generation of new genomic rearrangements although whether they are always the cause or consequence of the genome break and gene order change remains a disputed issue.

Work on this part will involve the following items:

1. extend the theoretical work done in the past years (rearrangement distance, rearrangement scenarios enumeration) to deal with different types of rearrangements and explore various types of biological constraints;
2. develop good random models (a largely open question despite some initial work in the area) for rearrangement distances and scenarios under a certain model, i.e. type of rearrangement operation(s) and of constraint(s), to assess whether the distances / scenarios observed have statistically notable characteristics;
3. extensively use the method(s) developed to investigate the rearrangement histories for the families of symbionts whose genomes have been sequenced and sufficiently annotated;

4. investigate the correlation of such histories with the repeats content and distribution along the genomes;
5. use the results of the above analyses together with a natural selection criterion to revisit the optimality model of rearrangement dynamics;
6. extend such model to deal with eukaryotic (multi-chromosomal) genomes;
7. at the interface host-symbiont, investigate the relation between the rearrangement histories in hosts and symbionts and the various types of symbiotic relationships observed in nature;
8. map such histories and their relation with the genetic and metabolic networks of hosts and symbionts, separately and at the interface;
9. develop methods to identify and quantify rearrangement events from NGS data.

Symbiotic dynamics
In order to understand the evolutionary consequences of symbiotic relations and their long term trajectories, one should be able to assess how tight is the association between symbionts and their hosts.

The main questions we would like to address are:

1. how often are symbionts horizontally transferred among branches of the host phylogenetic tree?
2. how long do parasites persist inside their host following the invasion of a new lineage?
3. what processes underlie this dynamic gain/loss equilibrium?

Mathematically, these questions have been traditionally addressed by co-phylogenetic methods, that is by comparing the evolutionary histories of hosts and parasites as represented in phylogenetic trees.

Currently available co-phylogenetic algorithms present various types of limitations as suggested in recent surveys. This may seriously compromise their interpretation with a view to understanding the evolutionary dynamics of parasites in communities. A few examples of limitations are the (often wrong) assumption made that the same rates of loss and gain of parasite infection apply for every host taxonomic group, and the fact that the possibility of multi-infections is not considered. In the latter case, exchange of genetic material between different parasites of a same host could further scramble the co-evolutionary signal. We therefore plan to:

1. better formalise the problem and the different simplifications that could be made, or inversely, should be avoided in the co-phylogeny studies; examples of the latter are the possibility of multi-infections, differential rate of loss and gain of infection depending on the host taxonomic group and geographic distance between hosts, etc., and propose better co-phylogenetic algorithms;
2. elaborate series of simulated data that will enable to (i) get a better grasp of the effect of the different parameters of the problem and, more practically, (ii) evaluate the performance of the method(s) that exist or are proposed (see next item);
3. apply the new methods to address the three questions above.

3.2. Intracellular interactions

The interactions of a symbiont with others sharing a same host, or with a symbiont and the cell of its host in the case of endosymbionts (organism that lives within the body or cells of another) are special, perhaps more complex cases of intracellular interactions that may concern different types of genetic elements, from organelles to whole chromosomes. The spatial arrangement of those genetic elements inside the nucleus of a cell is believed to be important both for gene expression and exchanges of genetic material between chromosomes. This question goes beyond the symbiosis one and has been investigated in the team in the last few years. Work on this will continue in future and concern developing algorithmic and statistical methods to analyse the interaction data that is starting to become available, in particular using NGS methods, in order to arrive at a better understanding of transcription, regulation both classical and epigenetic (inherited changes in phenotype or gene expression caused by mechanisms other than changes in the underlying DNA sequence), alternative splicing and trans-splicing phenomena, as well as study the possible interactions between an eukaryotic cell and its organelles or other cytoplasmic structures.

4. Application Domains

4.1. Domain

The main area of application of BAMBOO is biology, with a special focus on symbiosis (ERC project) and on intracellular interactions.

5. New Software and Platforms

5.1. AcypiCyc

Participants: Hubert Charles [EPI], Patrice Baa Puyoule [Contact, Patrice.Baa-Puyoulet@lyon.inra.fr], Stefano Colella [Contact, stefano.colella@lyon.inra.fr], Ludovic Cottret, Marie-France Sagot [EPI], Augusto Vellozo [Contact, augusto@cycadsys.org], Amélie Véron.

Database of the metabolic network of *Acyrthosiphon pisum*.
http://acypicyc.cycadsys.org/

5.2. AlViE

Participants: Pierluigi Crescenzi [Contact, pierluigi.crescenzi@unifi.it, ext. member EPI], Giorgio Gambosi, Roberto Grossi, Carlo Nocentini, Tommaso Papini, Walter Verdese.

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

5.3. Cassis

Participants: Christian Baudet [EPI, Contact, christian.baudet@inria.fr], Christian Gautier [EPI], Claire Lemaitre [Contact, claire.lemaitre@inria.fr], Marie-France Sagot [EPI], Eric Tannier.

Algorithm for precisely detecting genomic rearrangement breakpoints.
http://pbil.univ-lyon1.fr/software/Cassis/

5.4. Coala

Participants: Christian Baudet [EPI, Contact, christian.baudet@inria.fr], Pielrluigi Crescenzi, Bea Donati [EPI, Contact, bea.donati@inria.fr], Christian Gautier [EPI], Catherine Matias, Blerina Sinaimeri [EPI, Contact, blerina.sinaimeri@inria.fr], Marie-France Sagot [EPI, Contact, marie-france.sagot@inria.fr].

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).
http://coala.gforge.inria.fr/

5.5. C3Part & Isofun

Participants: Frédéric Boyer, Yves-Pol Deniélou, Anne Morgat [EPI, ext. member], Marie-France Sagot [EPI], Alain Viari [EPI, Contact, alain.viari@inria.fr].

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.http://www.inrialpes.fr/helix/people/viari/lxgraph/index.html

5.6. CycADS

Participants: Hubert Charles [EPI], Patrice Baa Puyoule [Contact, Patrice.Baa-Puyoulet@lyon.inra.fr], Stefano Colella [Contact, stefano.colella@lyon.inra.fr], Ludovic Cottret, Marie-France Sagot [EPI], Augusto Vellozo [Contact, augusto@cycadsys.org].

Cyc annotation database system.
http://www.cycadsys.org/

5.7. Eucalypt

Participants: Christian Baudet [EPI, Contact, christian.baudet@inria.fr], Pielrluigi Crescenzi, Bea Donati [Contact, bea.donati@inria.fr], Blerina Sinaimeri, Marie-France Sagot [EPI].

Algorithm for enumerating all optimal (possibly time-unfeasible) mappings of a parasite tree unto a host tree.
http://eucalypt.gforge.inria.fr/

5.8. Gobbolino & Touché

Participants: Vicente Acuña [EPI], Etienne Birmelé, Ludovic Cottret, Pierluigi Crescenzi, Fabien Jourdan, Vincent Lacroix, Alberto Marchetti-Spaccamela [EPI, ext. member], Andrea Marino, Paulo Vieira Milreu [EPI, Contact, pvmilreu@gmail.com], Marie-France Sagot [EPI, Contact, marie-france.sagot@inria.fr], Leen Stougie [EPI, ext. member].

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).
http://gforge.inria.fr/projects/gobbolino

5.9. KisSNP

Participants: Vincent Lacroix [EPI], Pierre Peterlongo [Contact, pierre.peterlongo@inria.fr], Nadia Pisanti, Marie-France Sagot [EPI], Nicolas Schnel.

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 BAMBOO and the GenScale Inria Team at Rennes (contact: pierre.peterlongo@inria.fr).
http://alcovna.genouest.org/kissnp/, http://colibread.inria.fr/software/discosnp/

5.10. KisSplice & KisSplice2igv7

Participants: Lilia Brinza [EPI], Alice Julien-Laferrière [EPI], Janice Kielbassa, Vincent Lacroix [Contact, EPI], Camille Marchet [EPI], Vincent Miele, Gustavo Sacomoto [EPI], Marie-France Sagot [EPI].

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).
http://kissplice.prabi.fr/

5.11. kissDE

Participants: Lilia Brinza [EPI], Janice Kielbassa, Vincent Lacroix [Contact, EPI], Camille Marchet [EPI], Vincent Miele.

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. http://kissplice.prabi.fr/tools/kissDE/

5.12. LASAGNE

Participants: Pierluigi Crescenzi [Contact, pierluigi.crescenzi@unifi.it, ext. member EPI], Roberto Grossi, Michel Habib, Claudio Imbrenda, Leonardo Lanzi, Andrea Marino.

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.
http://piluc.dsi.unifi.it/lasagne/?page_id=142

5.13. MetExplore

Participants: Michael Barrett, Hubert Charles [EPI], Ludovic Cottret [Contact, Ludovic.Cottret@toulouse.inra.fr], Fabien Jourdan, Marie-France Sagot [EPI], Florence Vinson, David Wildridge.

Web server to link metabolomic experiments and genome-scale metabolic networks.
http://metexplore.toulouse.inra.fr/metexplore/

5.14. Migal

Participants: Julien Allali [Contact, julien.allali@labri.fr], Marie-France Sagot [EPI, Contact, marie-france.sagot@inria.fr].

RNA, tree comparison
Algorithm for comparing RNA structures.
http://www-igm.univ-mlv.fr/~allali/logiciels/index.en.php

5.15. Mirinho

Participants: Cyril Fournier [EPI], Susan Higashi [EPI, Contact, susan.higashi@inria.fr], Christian Gautier [EPI], Christine Gaspin, Marie-France Sagot [EPI].

Predicts, at a genome-wide scale, microRNA candidates.
http://mirinho.gforge.inria.fr/

5.16. MotusWEB

Participants: Ludovic Cottret, Fabien Jourdan, Vincent Lacroix [EPI, Contact, vincent.lacroix@univ-lyon1.fr], Odile Rogier, Marie-France Sagot [EPI].

Algorithm for searching and inferring coloured motifs in metabolic networks (web-based version - offers different functionalities from the downloadable version).
http://pbil.univ-lyon1.fr/software/motus_web/

5.17. Motus

Participants: Ludovic Cottret, Fabien Jourdan, Vincent Lacroix [EPI, Contact, vincent.lacroix@univ-lyon1.fr], Odile Rogier, Marie-France Sagot [EPI].

Algorithm for searching and inferring coloured motifs in undirected graphs (downloadable version - offers different functionalities from the web-based version).
http://pbil.univ-lyon1.fr/software/motus/

5.18. PhEVER

Participants: Christian Gautier [EPI], Vincent Lotteau, Leonor Palmeira [Contact, mlpalmeira@ulg.ac.be], Chantal Rabourdin-Combe, Simon Penel.

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.
http://pbil.univ-lyon1.fr/databases/phever/

5.19. PepLine

Participants: Jérôme Garin, Alain Viari [EPI, Contact, alain.viari@inria.fr].

Pipeline for the high-throughput analysis of proteomic data.

5.20. Pitufo and family

Participants: Vicente Acuña [EPI], Ludovic Cottret [Contact, Ludovic.Cottret@toulouse.inra.fr], Alberto Marchetti-Spaccamela [EPI, ext. member], Paulo Vieira Milreu [EPI, Contact, pvmilreu@gmail.com], Marie-France Sagot [EPI], Leen Stougie [EPI, ext. member], Fabio Viduani-Martinez.

Algorithms to enumerate all minimal sets of precursors of target compounds in a metabolic network.
http://sites.google.com/site/pitufosoftware/

5.21. RepSeek

Participants: Guillaume Achaz [Contact, achaz@abi.snv.jussieu.fr], Eric Coissac, Alain Viari [EPI].

Finding approximate repeats in large DNA sequences.
http://wwwabi.snv.jussieu.fr/public/RepSeek/

5.22. Smile

Participants: Laurent Marsan, Marie-France Sagot [EPI, Contact, marie-france.sagot@inria.fr].

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

5.23. UniPathway

Participants: Eric Coissac, Anne Morgat [EPI, Contact, anne.morgat@inria.fr], Alain Viari [EPI].

Database of manually curated pathways developed with the Swiss-Prot group.
http://www.unipathway.org

6. New Results

6.1. Evolution of the genomes of endosymbionts in insects: the case of Hamiltonella defensa interacting with its various partners

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 synthesize 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 favors 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. 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. We highlighted the specialization 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 phenotype and genomic evolution.

The publications related to this area of research are either submitted or in preparation (to be submitted in the first months of the year).

6.2. Cardinium cBtQ1 providing insights into genome reduction, symbiont motility, and its settlement in Bemisia tabaci

Many insects harbor inherited bacterial endosymbionts. Although some of them are not strictly essential and are considered facultative, they can be a key to host survival under specific environmental conditions, such as parasitoid attacks, climate changes, or insecticide pressures. The whitefly *Bemisia tabaci* is at the top of the list of organisms inflicting agricultural damage and outbreaks, and changes in its distribution may be associated to global warming. In partnership with the group of Andrès Moya at the ICBiBE (Institut Cavanilles de Biodiversitat i Biologia Evolutiva), the genome of *Cardinium* cBtQ1, a facultative bacterial endosymbiont of *B. tabaci*, was sequenced and analysed [23].

6.3. Mitochondrial respiration and genomic analysis provide insight into the influence of the symbiotic bacterium on host trypanosomatid oxygen consumption

Certain trypanosomatids, such as *Angomonas deanei*, co-evolve with an endosymbiotic bacterium in a mutualistic relationship that is characterised by intense metabolic exchanges. We were able to show that the symbionts were able to respire for up to 4 h after isolation from the host. Moreover, our work suggests that the symbiont influences the mitochondrial respiration of the host protozoan [5].

6.4. Telling metabolic stories to explore metabolomics data

The increasing availability of metabolomics data enables to better understand the metabolic processes involved in the immediate response of an organism to environmental changes and stress. The data usually come in the form of a list of metabolites whose concentrations significantly changed under some conditions, and are thus not easy to interpret without being able to precisely visualize how such metabolites are interconnected.

We presented a method that enables to organize the data from any metabolomics experiment into metabolic stories [18]. Each story corresponds to a possible scenario explaining the flow of matter between the metabolites of interest. These scenarios may then be ranked in different ways depending on which interpretation one wishes to emphasize for the causal link between two affected metabolites: enzyme activation, enzyme inhibition or domino effect on the concentration changes of substrates and products. Equally probable stories under any selected ranking scheme can be further grouped into a single anthology that summarizes, in a unique subnetwork, all equivalently plausible alternative stories. An anthology is simply a union of such stories. We detailed an application of the method to the response of yeast to cadmium exposure. We used this system as a proof of concept for our method, and we showed that we are able to find a story that reproduces very well the current knowledge about the yeast response to cadmium. We further showed that this response is mostly based on enzyme activation. We also provided a framework for exploring the alternative pathways or side effects this local response is expected to have in the rest of the network. We discussed several interpretations for the changes we see, and we suggested hypotheses that could in principle be experimentally tested. Noticeably, our method requires simple input data and could be used in a wide variety of applications.

6.5. MiRNA and co: Methodologically exploring the world of small RNAs

We developed a reliable, robust, and much faster method for the prediction of pre-miRNAs. 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 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.

We then moved on to the second issue, that involved problems related to the prediction and analysis of non-coding RNAs (ncRNAs) in the bacterium *Mycoplasma hyopneumoniae*. A method, called ALVINHO, was thus developed for the prediction of targets in this bacterium, together with a pipeline for the segmentation of a numerical sequence and detection of conservation among ncRNA sequences using a k-partite graph.

We finally addressed a problem related to motifs, that is to patterns, that may be composed of one or more parts, that appear conserved in a set of sequences and may correspond to functional elements. This had already been addressed in a robust method called Smile. However, depending on the input parameters, the output may be too large to be tractable, as was realized in other works of the team. We then presented some clustering solutions to group the motifs that may correspond to a same biological element, and thus to better distinguish the biologically significant ones from noise that may be present in what often are large outputs from many motif extraction algorithms.

The publications related to this area of research are either submitted or in preparation (to be submitted in the first months of the year).

6.6. Efficient Algorithms for analysing RNA-seq Data

In the last years, we had addressed the problem of identifying and quantifying variants (alternative splicing and genomic polymorphism) in RNA-seq data when no reference genome is available, without assembling the full transcripts. Based on the fundamental idea that each variant corresponds to a recognizable pattern, a bubble, in a de Bruijn graph constructed from the RNA-seq reads, we propose a general model for all variants in such graphs. We then introduced an exact algorithm, called KISSPLICE, to extract alternative splicing events. We had showed that it enables to identify more correct events than general purpose transcriptome assemblers.

The main time bottleneck in the KISSPLICE algorithm is the bubble enumeration step. Thus, in an effort to make our method as scalable as possible, we had modified Johnson's cycle listing algorithm (Johnson (1975)) to enumerate bubbles in general directed graphs, while maintaining the same time complexity. We now proposed, using a different enumeration technique, an algorithm to list bubbles with path length constraints in weighted directed graphs [29]. For a graph with n vertices and m edges, the method we propose lists all bubbles with a given source in $O(n(m + nlogn))$ delay. Moreover, we experimentally showed that this algorithm is several orders of magnitude faster than the listing algorithm of KISSPLICE to identify bubbles corresponding to alternative splicing events.

Additionally, we showed that the same techniques used to list bubbles can be applied to one classical enumeration problem: K-shortest paths problems [29]. We considered a different parameterisation of the K-shortest paths problem: instead of bounding the number of st-paths, we bound the weight of the st-paths. We present a general scheme to list bounded length st-paths in weighted graphs that takes $O(nt(n, m))$ time per path, where $t(n, m)$ is the time for a single source shortest path computation. This algorithm uses memory linear in the size of the graphs, independent of the number of paths output. For undirected non-negatively weighted graphs, we also show an improved algorithm that lists all st-paths with bounded length in $O((m + t(n, m)))$ time per path.

The main memory bottleneck in KISSPLICE is the construction and representation of the de Bruijn graph. Thus, again with the goal to make our method as scalable as possible, we propose a new compact way to build and represent a de Bruijn graph improving over the state of the art [22]. We show both theoretically and experimentally that our approach uses 30% to 40% less memory than such state of the art, with an insignificant impact on the construction time. Our de Bruijn graph representation is general, in other words it is not restricted to the variation finding or RNA-seq context, and can be used as part of any algorithm that represents NGS data with de Bruijn graphs.

A major issue when analysing transcriptomes using short sequencing reads is to be able to deal with repeats that are longer than the reads. We proposed a first explicit model for large families of inexact repeats in the de Bruijn Graphs generated from RNA-seq data [21]. Taking advantage of this modelling, we also proposed an efficient algorithm which enumerates alternative splicing events without traversing repeat-induced subgraphs, therefore offering a first answer to one the main question left open at the end of Gustavo Sacomoto's PhD [4].

Motivated by previous work on the classical problem of listing cycles, we also studied from a more purely theoretical point of view how to list chordless cycles [28]. We thus developed an amortized $\tilde{O}(|V|)$-delay algorithm for listing chordless cycles in undirected graphs. Chordless cycles are very natural structures in undirected graphs, with an important history and distinguished role in graph theory. The best known solution to list all the C chordless cycles contained in an undirected graph $G = (V, E)$ takes $O(|E|2 + |E|{\cdot}C)$ time. In this paper we provide an algorithm taking $\tilde{\mathrm{O}}(|E| + |V|{\cdot}C)$ time. We also show how to obtain the same complexity for listing all the P chordless st-paths in G (where C is replaced by P).

6.7. Reference-free detection of isolated SNPs

Detecting single nucleotide polymorphisms (SNPs) between genomes is becoming a routine task with next-generation sequencing. Generally, SNP detec- tion methods use a reference genome. As non-model organisms are increasingly investigated, the need for reference-free methods has been amplified. Most of the existing reference-free methods have fundamental limitations: they can only call SNPs between exactly two datasets, and / or they require a prohibitive amount of computational resources. V. Lacroix participated in the developement of a method, called DISCOSNP to detect both heterozygous and homozygous isolated SNPs from any number of read datasets, without a reference genome, and with very low memory and time footprints (billions of reads can be analyzed with a standard desktop computer) [25]. To facilitate downstream genotyping analyses, DISCOSNP ranks predictions and outputs quality and coverage per allele. Compared to finding isolated SNPs using a state-of-the-art assembly and mapping approach, DISCOSNP requires significantly less computational resources, shows similar precision / recall values, and highly ranked predictions are less likely to be false positives. An experimental validation was conducted on an arthropod species (the tick *Ixodes ricinus*) on which de novo sequencing was performed. Among the predicted SNPs that were tested, 96% were successfully genotyped and truly exhibited polymorphism.

6.8. Endothelial, epithelial, and fibroblast cells exhibit specific splicing programs independently of their tissue of origin

Alternative splicing is the main mechanism of increasing the proteome diversity coded by a limited number of genes. It is well established that different tissues or organs express different splicing variants. However, organs are composed of common major cell types, including fibroblasts, epithelial, and endothelial cells. By analysing large-scale data sets generated by The ENCODE Project Consortium and after extensive RT-PCR validation, we demonstrated that each of the three major cell types expresses a specific splicing program independently of its organ origin [17]. Furthermore, by analysing splicing factor expression across samples, publicly available splicing factor binding site data sets (CLIP-seq), and exon array data sets after splicing factor depletion, we identified several splicing factors that contribute to establishing these cell type-specific splicing programs.

6.9. Length and symmetry on the sorting by weighted inversions problem

Large-scale mutational events that occur when stretches of DNA sequence move throughout genomes are called genome rearrangement events. In bacteria, inversions are one of the most frequently observed rearrangements. In some bacterial families, inversions are biased in favour of symmetry as shown by recent research. In addition, several results suggest that short segment inversions are more frequent in the evolution of microbial genomes. Despite the fact that symmetry and length of the reversed segments seem very important, they have not been considered together in any problem in the genome rearrangement field. We defined the problem of sorting genomes (or permutations) using inversions whose costs are assigned based on their lengths and asymmetries [27]. We presented five procedures and assessed their performance on small sized permutations. The ideas presented in the paper provide insights to solve the problem and set the stage for a proper theoretical analysis.

6.10. Efficient tree reconciliation enumerator plus cophylogeny reconstruction algorithm via an Approximate Bayesian Computation

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. Any further biological interpretation of them must therefore take this into account, making the capacity to enumerate all optimal solutions a crucial point. Only two algorithms currently exist that attempt such enumeration; in one case not all possible solutions are produced while in the other not all cost vectors are currently handled. Our objective in addressing this problem was two-fold. The first was to fill this gap, and the second was to test whether the number of solutions generally observed can be an issue in terms of interpretation.

We presented a polynomial-delay algorithm called EUCALYPT for enumerating all optimal reconciliations [12]. 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 were 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 allows 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.

Despite an increasingly vast literature on cophylogenetic reconstructions for studying host-parasite associations, understanding the common evolutionary history of such systems remains a problem that is far from being solved. Most algorithms for host-parasite reconciliation use an event-based model, where the events include in general (a subset of) cospeciation, 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. This is what we did ourselves in EUCALYPT. 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 [8]. 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.

6.11. Others

Other works, often experimental were also developed during 2014 and published in a number of papers [6], [7], [9], [10], [11], [13], [16], [19], [20], [24], [26].

7. Partnerships and Cooperations

7.1. National Initiatives

7.1.1. ABS4NGS

- Title: Solutions Algorithmiques, Bioinformatiques et Logicielles pour le Séquençage Haut Débit
- Coordinator: E. Barillot
- BAMBOO participant(s): V. Lacroix

- Type: ANR (2012-2015)
- Web page: Not available

7.1.2. Colib'read

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

7.1.3. Exomic

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

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

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

7.1.5. ExHyb

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

7.1.6. IMetSym

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

7.1.7. ImmunSymbArt

- Title: Immunity and Symbiosis in Arthropods
- Coordinator: D. Bouchon
- BAMBOO participant(s): F. Vavre
- Type: ANR Blanc (2010-2014)
- Web page: Not available

7.1.8. Metagenomics of Bemisia tabaci

- Title: Metagenomics of *Bemisia tabaci* symbiotic communities
- Coordinator: L. Mouton (LBBE, UCBL)
- BAMBOO participant(s): F. Vavre, M.-F. Sagot
- Type: Genoscope Project
- Web page: Not available

7.1.9. SpeciAphid

- Title: Evolutionary genetics and mechanisms of plant adaptation in aphids
- Coordinator: Jean-Christophe Simon (IGEPP, INRA, Rennes)
- BAMBOO participant(s): H. Charles, S. Colella, Y. Rahbé
- Type: ANR (2012-2014)
- Web page: Not available

7.2. European Initiatives

7.2.1. FP7 & H2020 Projects

7.2.1.1. BacHBerry

Title: BACterial Hosts for production of Bioactive phenolics from bERRY fruits

Coordinator: Jochen Förster, DTU Danemark

BAMBOO participant(s): R. Andrade, L. Bulteau, A. Julien-Laferriére, V. Lacroix, D. Parrot, M.-F. Sagot, A. Viari, M. Wannagat

Type: FP7 - KBBE (2013-2016)

Web page: http://www.bachberry.eu/

7.2.1.2. DroParCon

- Title: Drosophila parasitoid consortium
- Coordinator: Jochen Förster (Novo Nordisk Foundation Center for Biosustainability (CFB), Copenhagen, Danemark)
- BAMBOO participant(s): F. Vavre
- Type: PHC (2012-2014)
- Web page: http://www.droparcon.org

7.2.1.3. Microme

- Title: The Microme Project: A Knowledge-Based Bioinformatics Framework for Microbial Pathway Genomics
- Coordinator: P. Kersey (EBI)
- European partners: Amabiotics (France), CEA (France), CERTH (Greece), CSIC (Spain), CNIO (Spain), DSMZ (Germany), EBI (UK), HZI (Germany), Isthmus (France), Molecular Nertwork (Germany), SIB (Switzerland), Tel Aviv Univ. (Israel), Université Libre de Bruxelles (Belgium), WTSI (UK), Wageningen Univ. (The Netherlands)
- BAMBOO participant(s): Anne Morgat
- Type: Collaborative Project. Grant Agreement Number 222886-2
- Web page: http://www.microme.eu

7.2.1.4. SISYPHE

- Title: Species Identity and SYmbiosis Formally and Experimentally explored

- Coordinator: M.-F. Sagot
- BAMBOO participant(s): Whole BAMBOO team
- Type: ERC Advanced Grant (2010-2015)
- Web page: http://pbil.univ-lyon1.fr/members/sagot/htdocs/team/projects/sisyphe/sisyphe.html

7.2.1.5. SWIPE

- Title: Predicting whitefly population outbreaks in changing environments
- Coordinator: E. Zchori-Fein
- BAMBOO participant(s): F. Vavre
- Type: European ERA-NET program ARIMNET (2012-2015)
- Web page: Not available

7.2.1.6. Symbiox

- Title: Role of the oxidative environment in the stability of symbiotic associations
- Coordinator: F. Vavre
- BAMBOO participant(s): F. Vavre
- Type: Marie Curie IOF for Natacha Kremer (2011-2014)
- Web page: Not available

7.3. International Initiatives

7.3.1. Inria International Labs

BAMBOO 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.

7.3.2. Inria International Partners

Bamboo has an Inria International Partnership, called AMICI (see http://team.inria.fr/bamboo/amici/), with three partners in Italy (Universities of Rome “La Sapienza", Florence, and Pisa) and one in the Netherlands (Free University of Amsterdam / CWI). There are two unifying interests to all the projects of AMICI: algorithmics, and biology. At the present time, mostly because the current work of BAMBOO is centered on the ERC project SISYPHE (“Species Identity and SYmbiosis Formally and Experimentally explored"), the biology is very oriented to the general study, at the molecular level, of the symbiotic relation (genomics and other associated “omics", evolution, biochemical and interaction networks). This should evolve in future to extend the symbiotic study to either the ecological or a more health-oriented level, or to address new biology-related problems using mathematical modelling and techniques, and algorithmics.

7.3.3. Participation In other International Programs

BAMBOO 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 preliminary web page for the LIA LIRIO is available at this address: http://team.inria.fr/bamboo/en/cnrs-lia-laboratoire-international-associe-lirio/.

BAMBOO 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, renewable once. 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.

7.4. International Research Visitors

7.4.1. Visits of International Scientists

During 2014, the team had 4 international scientists visiting our group for at least one week. These included:

- Franciele Maboni, Federal University of Rio Grande do Sul, Porto Alegre, Brazil, two visits of, respectively, 15 days and 1 week;
- Maria Cristina Motta, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, two visits of approximately 15 days;
- Susana Vinga, INESC-ID, IST Lisbon, Portugal, visit of 1 week;
- Arnaldo Zaha, Federal University of Rio Grande do Sul, Porto Alegre, Brazil, visit of 10 days.

The above does not count the frequent visits of our external collaborators, members of the Inria International Partnership AMICI or of the LIA LIRIO.

8. Dissemination

8.1. Promoting Scientific Activities

8.1.1. Scientific events organisation

8.1.1.1. General chair, scientific chair

Cristina Vieira is director of the GDRE "Comparative genomics" since the GDRE was renewed in 2010. Marie-France Sagot co-organised the Colloquium *EMOTIONS 2014* (http://emotions2014.sciencesconf.org/) as well as the International School *ADDICTION 2014* (http://team.inria.fr/bamboo/en/events/school-addiction/). She is also since 2010 member and since 2014 Chair of the Steering Committee of the International Conference *LATIN* (http://www.latintcs.org/). She is since member of the Steering Committee of the *European Conference on Computational Biology* (*ECCB*).

8.1.2. Scientific events selection

8.1.2.1. Member of the conference program committee

Marie-France Sagot was a member of the program committee for the following international conferences in 2014: 22nd Annual International Conference on Intelligent Systems in Molecular Biology (ISMB), Prague Stringology Conference 2014 (PSC), 18th Annual International Conference on Research in Computational Molecular Biology (RECOMB), 12th RECOMB Comparative Genomics 2014 (RECOMB-CG), RECOMB Satellite Workshop On Massively Parallel Sequencing (RECOMB-SEQ), 14th Workshop on Algorithms in Bioinformatics (WABI).

8.1.2.2. Reviewer

Besides the above, various other members of BAMBOO have been reviewer for conferences. These include Christian Baudet, Laurent Bulteau, Vincent Lacroix, Arnaud Mary, Gustavo Sacomoto, and Blerina Sinaimeri.

8.1.3. Journal

8.1.3.1. Member of the editorial board

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*.

Cristina Vieira is Executive Editor of *Gene*, and since 2014 member of the Editorial Board of *Mobile DNA*.

8.1.3.2. Reviewer for Journals

Susan Higashi was reviewer for *Algorithms for Molecular Biology*, and *Evolutionary Bioinformatics*. Vincent Lacroix was reviewer for *BMC Genomics*, *PLoS Computational Biology*, and *Methods*. Gustavo Sacomoto was reviewer for *Bioinformatics*, and *BMC Bioinformatics*. Fabrice Vavre was reviewer for *PLoS Biology*, *Genome Biology and Evolution*, *Applied and Environmental Microbiology*, *Journal of Applied Entomology*, *Symbiosis*, *Microbiology*, *Bulletin of Entomological Research*.

8.2. Teaching - Supervision - Juries

8.2.1. Teaching

The members of BAMBOO teach both at the Department of Biology of the University of Lyon (in particular within the MIV (Mathematics and Computer Science for the Life Sciences) specialty) and at the department of Bioinformatics of the Insa (National Institute of Applied Sciences). 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 École Normale Supérieure. 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. 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, except in 2013-2014 where he taught 120 hours as he had a partial sabbatical funded by Inria (through the ERC AdG Sisyphe grant). He teaches bioinformatics and statistics. Arnaud Mary who was recruited in October as an Associate Professor taught (and will teach) 150 hours in 2014-2015 (L1: mathematics; L2: bioinformatics; M1: data analysis; M1: computer science) as he has a one-year partial sabbatical as a recently recruited personnel of the University of Lyon. Fabrice Vavre taught approximately 20 hours in different Master 2 courses in Lyon, Poitiers and Amiens.

Alice Julien-Laferrière taught 114 hours of Applied Mathematics and Bioinformatics at the Department of Biology (undergraduate students), Hélène Lopez-Maestre and Laura Urbini taught each 32 hours of Mathematics and Statistics at the Department of Biology (undergraduate students). Two other PhD students taught more occasionally in 2012-2013: Gustavo Sacomoto taught 16 hours of graph algorithms at the Department of Biology (M1: Bioinformatics), and Mariana Ferrarini taught 23 hours at the Department of Biology (L3: Bioinformatics). The postdocs are also involved in teaching. Cecilia Klein taught 7 hours in 2014 at the IUT / University of Lyon 1, Blerina Sinaimeri taught 17 hours of graph algorithms at the University of Lyon 1 and INSA (M2), and Christian Baudet taught 24 hours of computer science at the University of Lyon 1 (L3) and the INSA (L1).

All members of the BAMBOO team are affiliated to the doctoral school E2M2 (Ecology-Evolution-Microbiology-Modelling).

8.2.2. Supervision

The following are the PhDs defended in BAMBOO in 2014.

Gustavo Sacomoto, University of Lyon 1, March 6, supervisors: P. Crescenzi (Univ. of Florence, Italy), V. Lacroix, and M.-F. Sagot.

Beatrice Donati, University of Lyon 1 and of Florence (Italy), November 12, supervisors: P. Crescenzi (Univ. of Florence, Italy) and M.-F. Sagot.

Pierre-Antoine Rollat-Farnier, University of Lyon 1, November 22, supervisors: L. Mouton, M.-F. Sagot, and F. Vavre.

Susan Higashi, University of Lyon 1, November 24, supervisors: S. Colella, C. Gautier, and M.-F. Sagot.

8.2.3. Juries

- H. Charles: Reviewer of the PhDs of Diana Stefan, University Joseph Fourier, France, and member of the jury for Nicolas Parisot (University Blaise Pascal, Clermont-Ferrand, France) and Sylvain Prigent (Inria/Irisa Rennes).
- M.-F. Sagot: Reviewer of the PhDs of Kimon Froussios, King's College, London, UK, and of Caroline Baroukh, University of Perpignan, Narbonne, France.
- F. Vavre: Member of the jury for the PhDs of: Myriam Badaoui, Université de Poitiers, France; Wen-Juan Ma, University of Groningen, The Netherlands. He was reviewer for the HDR of Olivier Duron, Université de Montpellier 2, France.
- C. Vieira: Member of the jury for the PhDs of: Maialen Sistiaga, Université du Pays Basque, Bilbao; Pierre-Antoine Rollat-Farnier, Université Lyon 1; Laetitia Delabaere, Université Lyon 1; Guillaume Minard, Université Lyon1; Yann Lesecque, Université Lyon1; Antoine Bridier-Nahmias, Paris VI; Abdelhakim Negoua, Université de Marrakech, Marrakech, Maroc. Reviewer of the PhDs of Frank Touret, EPHE, Lyon, and of Domitile Chaloppin, ENS Lyon. Member of the HDR jury of Abderrahman Khila, ENS Lyon.

9. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] B. Donati. *Graph models and algorithms in (co-)evolutionary contexts*, Universite Claude Bernard Lyon 1 ; Università degli studi di Firenze, November 2014, https://hal.inria.fr/tel-01095298

[2] S. Higashi. *MiRNA and co: Methodologically exploring the world of small RNAs*, Universite Claude Bernard Lyon 1, November 2014, https://hal.archives-ouvertes.fr/tel-01096833

[3] P.-A. Rollat-Farnier. *Evolution of the genomes of the endosymbionts in phloem-sap feeding insects: the case of Hamiltonella defensa interacting with its various partners*, Université Claude Bernard Lyon 1, November 2014, https://hal.inria.fr/tel-01102943

[4] G. Sacomoto. *Efficient Algorithms for de novo Assembly of Alternative Splicing Events from RNA-seq Data*, Universite Claude Bernard Lyon 1, March 2014, https://hal.inria.fr/tel-01095280

Articles in International Peer-Reviewed Journals

[5] 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", August 2014, pp. 1-11 [*DOI :* 10.1017/S0031182014001139], https://hal.inria.fr/hal-01073754

[6] M. BADAWI, I. GIRAUD, F. VAVRE, P. GRÈVE, R. CORDAUX. *Signs of Neutralization in a Redundant Gene Involved in Homologous Recombination in Wolbachia Endosymbionts*, in "Genome Biology and Evolution", 2014, vol. 10, n^o^ 6, pp. 2654-2664 [*DOI :* 10.1093/GBE/EVU207], https://hal.inria.fr/hal-01091858

[7] D. BASQUIN, A. SPIERER, F. BEGEOT, D. E. KORYAKOV, A.-L. TODESCHINI, S. RONSSERAY, C. VIEIRA, P. SPIERER, M. DELATTRE. *The Drosophila Su(var)3-7 gene is required for oogenesis and female fertility, genetically interacts with piwi and aubergine, but impacts only weakly transposon silencing*, in "PLoS ONE", 2014, vol. 9, n^o^ 5, e96802, https://hal.inria.fr/hal-01092623

[8] 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", December 2014, 49 p. [*DOI :* 10.1093/SYSBIO/SYU129], https://hal.inria.fr/hal-01092972

[9] T. BOIVIN, H. HENRI, F. VAVRE, C. GIDOIN, P. VEBER, J.-N. CANDAU, E. MAGNOUX, A. ROQUES, M.-A. AUGER-ROZENBERG. *Epidemiology of asexuality induced by the endosymbiotic Wolbachia across phytophagous wasp species: host plant specialization matters*, in "Molecular Ecology Notes", April 2014, pp. 2362-2375, https://hal.inria.fr/hal-01092614

[10] C. M. A. CARARETO, E. H. HERNANDEZ, C. VIEIRA. *Genomic regions harboring insecticide resistance-associated Cyp genes are enriched by transposable element fragments carrying putative transcription factor binding sites in two sibling Drosophila species*, in "Gene", 2014, vol. 537, n^o^ 1, pp. 93-99 [*DOI :* 10.1016/J.GENE.2013.11.080], https://hal.inria.fr/hal-00922703

[11] E. A. G. CARNELOSSI, E. LERAT, H. HENRI, S. MARTINEZ, C. M. A. CARARETO, C. VIEIRA. *Specific activation of an I-like element in Drosophila interspecific hybrids*, in "Genome Biology and Evolution", June 2014, pp. 1806-17, https://hal.inria.fr/hal-01079996

[12] 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

[13] M. FABLET, A. AKKOUCHE, V. BRAMAN, C. VIEIRA. *Variable expression levels detected in the Drosophila effectors of piRNA biogenesis*, in "Gene", 2014, vol. 537, n^o^ 1, pp. 149-153 [*DOI :* 10.1016/J.GENE.2013.11.095], https://hal.inria.fr/hal-00922704

[14] M. FEDERICO, P. PETERLONGO, N. PISANTI, M.-F. SAGOT. *Rime: Repeat identification*, in "Discrete Applied Mathematics", 2014, vol. 163, n^o^ 3, pp. 275–286 [*DOI :* 10.1016/J.DAM.2013.02.016], https://hal.inria.fr/hal-00802023

[15] N. KREMER, F. VAVRE. *Microbial impacts on insect evolutionary diversification: from patterns to mechanisms*, in "Current Opinion in Insect Science", October 2014, vol. 4, pp. 29–34 [*DOI :* 10.1016/J.COIS.2014.08.003], https://hal.inria.fr/hal-01104106

[16] W.-J. MA, F. VAVRE, L. W. BEUKEBOOM. *Manipulation of arthropod sex determination by endosymbionts: diversity and molecular mechanisms*, in "Sex. Dev.", 2014, pp. 59-73, https://hal.inria.fr/hal-01092616

[17] P. MALLINJOUD, J.-P. VILLEMIN, H. MORTADA, M. POLAY ESPINOZA, F.-O. DESMET, S. SAMAAN, E. CHAUTARD, L.-C. TRANCHEVENT, D. AUBOEUF. *Endothelial, epithelial, and fibroblast cells exhibit*

specific splicing programs independently of their tissue of origin, in "Genome Research", 2014, vol. 24, n^o 3, pp. 511-521 [*DOI :* 10.1101/GR.162933.113], https://hal.archives-ouvertes.fr/hal-01091290

[18] P. V. MILREU, C. C. KLEIN, L. COTTRET, V. ACUÑA, E. BIRMELÉ, M. BORASSI, C. JUNOT, A. MARCHETTI-SPACCAMELA, A. MARINO, L. STOUGIE, F. JOURDAN, P. CRESCENZI, V. LACROIX, M.-F. SAGOT. *Telling metabolic stories to explore metabolomics data: a case study on the yeast response to cadmium exposure*, in "Bioinformatics", January 2014, vol. 30, n^o 1, pp. 61-70 [*DOI :* 10.1093/BIOINFORMATICS/BTT597], https://hal.inria.fr/hal-00922567

[19] 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", May 2014, pp. 2654-2664, https://hal.inria.fr/hal-01092608

[20] D. ROQUIS, J. M. J. LEPESANT, E. VILLAFAN, J. BOISSIER, C. VIEIRA, C. COSSEAU, C. GRUNAU. *Exposure to hycanthone alters chromatin structure around specific gene functions and specific repeats in Schistosoma mansoni*, in "Frontiers in Genetics", 2014, vol. 5, 207 p. , https://hal.inria.fr/hal-01092618

[21] G. SACOMOTO, B. SINAIMERI, C. MARCHET, V. MIELE, M.-F. SAGOT, V. LACROIX. *Navigating in a Sea of Repeats in RNA-seq without Drowning*, in "Lecture Notes in Bioinformatics", 2014, vol. 8701, pp. 82-96 [*DOI :* 10.1007/978-3-662-44753-6_7], https://hal.inria.fr/hal-01079947

[22] K. SALIKHOV, G. SACOMOTO, G. KUCHEROV. *Using cascading Bloom filters to improve the memory usage for de Brujin graphs*, in "Algorithms for Molecular Biology", 2014, vol. 9, n^o 1, 2 p. [*DOI :* 10.1186/1748-7188-9-2], https://hal.inria.fr/hal-00971576

[23] D. SANTOS-GARCIA, P.-A. ROLLAT-FARNIER, F. BEITIA, E. ZCHORI-FEIN, F. VAVRE, L. MOUTON, A. MOYA, A. LATORRE, F. J. SILVA. *The genome of Cardinium cBtQ1 provides insights into genome reduction, symbiont motility, and its settlement in Bemisia tabaci*, in "Genome Biology and Evolution", March 2014, pp. 1013-1030, https://hal.inria.fr/hal-01092610

[24] P. SAPOUNTZIS, G. DUPORT, S. BALMAND, K. GAGET, S. JAUBERT-POSSAMAI, G. FEBVAY, H. CHARLES, Y. RAHBÉ, S. COLELLA, F. CALEVRO. *New insight into the RNA interference response against cathepsin-L gene in the pea aphid, Acyrthosiphon pisum: Molting or gut phenotypes specifically induced by injection or feeding treatments*, in "Insect Biochemistry and Molecular Biology", May 2014, epub ahead of print, forthcoming [*DOI :* 10.1016/J.IBMB.2014.05.005], https://hal.archives-ouvertes.fr/hal-01002574

[25] R. URICARU, G. RIZK, V. LACROIX, E. QUILLERY, O. PLANTARD, R. CHIKHI, C. LEMAITRE, P. PETERLONGO. *Reference-free detection of isolated SNPs*, in "Nucleic Acids Research", November 2014, pp. 1 - 12 [*DOI :* 10.1093/NAR/GKU1187], https://hal.inria.fr/hal-01083715

[26] D. VELA, A. FONTDEVILA, C. VIEIRA, M. P. GARCÍA GUERREIRO. *A genome-wide survey of genetic instability by transposition in Drosophila hybrids*, in "PLoS ONE", 2014, vol. 9, n^o 2, e88992, https://hal.inria.fr/hal-01092626

International Conferences with Proceedings

[27] C. BAUDET, U. DIAS, Z. DIAS. *Length and Symmetry on the Sorting by Weighted Inversions Problem*, in "BSB - 9th Brazilian Symposium on Bioinformatics", Belo Horizonte, Brazil, S. CAMPOS (editor), Advances

in Bioinformatics and Computational Biology, October 2014, vol. 8826, pp. 99 - 106 [*DOI :* 10.1007/978-3-319-12418-6_13], https://hal.inria.fr/hal-01092607

[28] R. Ferreira, R. Grossi, R. Rizzi, G. Sacomoto, M.-F. Sagot. *Amortized Õ(|V|) -Delay Algorithm for Listing Chordless Cycles in Undirected Graphs*, in "22th Annual European Symposium on Algorithms", Wroclaw, Poland, Algorithms ESA 2014, Lecture Notes in Computer Science, September 2014, vol. 8737, pp. 418-429 [*DOI :* 10.1007/978-3-662-44777-2_35], https://hal.inria.fr/hal-01081031

[29] R. Rizzi, G. Sacomoto, M.-F. Sagot. *Efficiently listing bounded length st-paths*, in "Twenty-Fifth International Workshop on Combinatorial Algorithms (IWOCA 2014)", Duluth, Minnesota, France, September 2014, forthcoming, https://hal.inria.fr/hal-01091924