Activity Report 2011

Team BAMBOO

An algorithmic view on genomes, cells, and environments

RESEARCH CENTER
Grenoble - Rhône-Alpes

THEME
Computational Biology and Bioinformatics

Table of contents

Team BAMBOO

Keywords: Computational Biology

1. Members

Research Scientists

Yvan Rahbé [Senior Researcher Inra (DR), HdR]
Marie-France Sagot [Team leader, Senior Researcher Inria (DR), HdR]
Fabrice Vavre [Senior Researcher (DR) Cnrs, HdR]
Alain Viari [Senior Researcher Inria (DR)]

Faculty Members

Etienne Birmelé [Associate Professor, Univ. Évry, délégation INRIA for one year since Sept. 1st, 2011]
Hubert Charles [Associate Professor, Insa-Lyon, HdR]
Christian Gautier [Professor, University Claude Bernard, HdR]
Vincent Lacroix [Associate Professor, University Claude Bernard]
Cristina Vieira [Professor, IUF, University Claude Bernard, HdR]

Technical Staff

Lauranne Duquenne [CDD ERC, until April 30, 2011]

PhD Students

Beatrice Donati [scholarship ERC, supervisors: Marie-France Sagot and Pierluigi Crescenzi, started Nov. 1st, 2011]
Pierre-Antoine Farnier [scholarship ERC, supervisors: Marie-France Sagot, Laurence Mouton and Fabrice Vavre]
Alexandre Freire [scholarship sandwich-PhD CNPq, Brazil, supervisor: Carlos Eduardo Ferreira, until Nov. 30, 2011]
Susan Higashi [scholarship ERC, supervisors: Marie-France Sagot, Christian Gautier and Stefano Colella, started Jul. 1st, 2011]
Cecília Coimbra Klein [scholarship ERC, supervisors: Marie-France Sagot and Ana Tereza Vasconcelos]
Nuno Mendes [scholarship FCT, Ministry of Science and Technology, Portugal, surpervisors: Ana Teresa Freitas and Marie-France Sagot, defended in June 6, 2011]
Paulo Vieira Milreu [scholarship ANR, supervisors: Vincent Lacroix and Marie-France Sagot]
Alexandra Popa [scholarship Ministère de la Recherche, supervisors: Christian Gautier and Dominique Mouchiroud, defended May 24, 2011]
Gustavo Sacomoto [scholarship ERC, supervisors: Marie-France Sagot, Vincent Lacroix and Pierluigi Crescenzi, started Feb 15, 2011]
Patrícia Simões [scholarship FCT, Ministry of Science and Technology, Portugal, supervisors: Marie-France Sagot and Sylvain Charlat]

Post-Doctoral Fellows

Vicente Acuña [CDD ANR]
Christian Baudet [CDD ANR, started Jan. 1st, 2011]
Matteo Brilli [CDD ERC]
Janice Kielbassa [CDD ERC]
Igor Nor [scholarship ANR, until Feb 28, 2011]

Visiting Scientists

Pablo Ivan [Master student, Lab Nacional de Computação Científica, Petrópolis, Brazil, co-supervision: Marisa Fabiana Nicolás, Ana Tereza Vasconcelos and Marie-France Sagot, from Feb. 1st until Jul. 31st, 2011]
Allan Martíns [PhD student, Universidade Federal do Rio de Janeiro (UFRJ), Brazil, supervision: Cristina Motta, from April 11 until August 10, 2011]

Administrative Assistant

Florence Bouheddi [Secretary (SAR) Inria]

Others

Pierluigi Crescenzi [Professor, University of Florence, Italy, external collaborator]
Manolo Gouy [Senior Researcher (DR) Cnrs, external collaborator, HdR]
Roberto Grossi [Professor, University of Pisa, Italy, external collaborator]
Alberto Marchetti-Spaccamela [Professor, University La Sapienza, Rome, Italy, external collaborator]
Anne Morgat [Scientist, Swiss Institute of Bioinformatics, external collaborator]
Nadia Pisanti [Associate Professor, University of Pisa, Italy, external collaborator]
Leen Stougie [Free University Amsterdam and CWI, Amsterdam, the Netherlands, external collaborator]

2. Overall Objectives

2.1. Highlights

There are two highlights we wish to stress for 2011. One is scientific and relates to contributions we made this year to the problems of intra- and inter-chromosomal interactions. The other highlight is both scientific and organisational. It concerns the setting up of an INRIA International Partnership with our close collaborators in Italy and the Netherlands.

2.1.1. Intra- and inter-chromosomal interactions

This year, we were able to make two contributions which both rely on the use of recently published data of 3D co-localisation of DNA fragments in human cells. In both cases, our findings are novel and broaden our view of what is a gene and what drives its (change of) location on the genome.

On the one hand, from the joint study of the network of spatial proximities of human genomic loci and a dataset of evolutionary breakpoints between human and mouse, we were able to provide evidence that evolutionary breakpoints tend to cluster spatially in human cells, which leads us to propose the new notion of *spatial synteny*, which generalises the widely used concept of genomic synteny.

On the other hand, in the framework of the extension of the ENCODE project to chromosome 21 and 22, we had the opportunity to identify a new category of transcripts, which we call *chimeric transcripts*. These transcripts contain exons from different genes, which can themselves be located far apart on the same chromosome, or on different chromosomes. We further found that the network formed by these connected genes is enriched in cliques of sizes 3 and 4, which seems to indicate that transcription and/or splicing of these sets of genes co-occur in time and space, as is confirmed by the confrontation of our expression dataset to a dataset indicating the co-localisation of DNA fragments in 3D.

2.1.2. INRIA International Partner: AMICI

The INRIA International Partner project AMICI is the continuation and extension of the INRIA Associated Team SIMBIOSI that started in January 2009 and ended in December 2011 (see the web page for SIMBIOSI at http://pbil.univ-lyon1.fr/members/sagot/htdocs/team/projects/simbiosi/simbiosi.html). It includes, beside the EPI, four partners: University of Rome La Sapienza, group headed by Alberto Marchetti-Spaccamela; Free University of Amsterdam and CWI, group headed by Leen Stougie; University of Florence, group headed by Pierluigi Crescenzi; University of Pisa, group headed by Nadia Pisanti. More information on AMICI may be found at http://piluc.dsi.unifi.it/amici.

3. Scientific Foundations

3.1. Formal methods

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.

3.2. 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. Experimentaly, 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 pratically, (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.

3.3. 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. Biology with a focus on symbiosis

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

5. Software

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. BaobabLuna

Participants: Marília Braga [Contact, mdvbraga@gmail.com], Marie-France Sagot [EPI], Eric Tannier.

Manipulation of signed permutations in the context of genomic evolution.
http://pbil.univ-lyon1.fr/software/luna/

5.3. Cassis

Participants: Christian Baudet [EPI, Contact, christian.baudet@univ-lyon1.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. Cravela

Participants: Ana Teresa Freitas, Nuno Mendes [EPI, Contact, ndm@kdbio.inesc-id.pt], Marie-France Sagot [EPI, Contact, marie-france.sagot@inria.fr].

Framework for the identification and evaluation of miRNA precursors (finished), targets (in development) and regulatory modules(in development).
http://www.cravela.org/

5.5. C3P

Participants: Frédéric Boyer, Anne Morgat [EPI, ext. member], Alain Viari [EPI, Contact, alain.viari@inria.fr].

Merging two or more graphs representing biological data (e.g. pathways, ...).
http://www.inrialpes.fr/helix/people/viari/cccpart

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. Ed'Nimbus

Participants: Pierre Peterlongo [Contact, pierre.peterlongo@inria.fr], Marie-France Sagot [EPI].

Algorithm for detecting and filtering repeats in sequences prior to multiple alignments.

5.8. GeM

Participants: Gisèle Bronner, Christian Gautier [EPI, Contact, christian.gautier@univ-lyon1.fr], Bruno Spataro.

Database for comparative genomic analysis of complete vertebrate genomes.
http://pbil.univ-lyon1.fr/gem/gem_home.php

5.9. Gobbolino

Participants: Vicente Acuña [EPI], Etienne Birmelé [EPI, délégation], 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], Leen Stougie [EPI, ext. member].

Algorithm to enumerate all metabolic stories in a metabolic network given a set of metabolites of interest.
Code available on request.

5.10. 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.
http://alcovna.genouest.org/kissnp/

5.11. kisSplice

Participants: Rayan Chikhi, Janice Kielbassa [EPI], Vincent Lacroix [Contact, EPI], Pierre Peterlongo [Contact, pierre.peterlongo@inria.fr], Gustavo Sacomoto [EPI], Marie-France Sagot [EPI], Raluca Uricaru.

Algorithm for de-novo calling alternative splicing events from RNA-seq data.
http://alcovna.genouest.org/kissplice/

5.12. 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/

5.13. Migal

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

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

5.14. 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/

5.15. 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 (downloable version - offers different functionalities from the web-based version).
http://pbil.univ-lyon1.fr/software/motus/

5.16. 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/

5.17. PepLine

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

Pipeline for the high-throughput analysis of proteomic data.

http://www.grenoble.prabi.fr/protehome/software/pepline

5.18. Pitufo

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.

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

5.19. PSbR

Participants: Yoan Diekmann, Marie-France Sagot [EPI, Contact, marie-france.sagot@inria.fr], Eric Tannier.

Algorithm for testing the evolution and conservation of common clusters of genes.
http://pbil.univ-lyon1.fr/members/sagot/htdocs/team/software/PSbR/

5.20. 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/

5.21. 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.

5.22. Tuiuiu

Participants: Alair Pereira do Lago, Pierre Peterlongo [Contact, pierre.peterlongo@inria.fr], Nadia Pisanti, Gustavo Sacomoto [EPI], Marie-France Sagot [EPI].

Multiple repeat search filter with edit distance.
http://mobyle.genouest.org/cgi-bin/Mobyle/portal.py?form=tuiuiu

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. Sex-specific impact of meiotic recombination on nucleotide composition

Meiotic recombination is an important evolutionary force shaping the nucleotide landscape of genomes. For most vertebrates, the frequency of recombination varies slightly to considerably between the sexes (heterochiasmy). We extended the examination of the evolutionary impact of heterochiasmy beyond primates to include four additional eutherian mammals (mouse, dog, pig, and sheep), a metatherian mammal (opossum), and a bird (chicken). We compared sex-specific recombination rates with nucleotide substitution patterns evaluated on transposable elements. Our results, based on a comparative approach, revealed a great diversity of the relationship between heterochiasmy and nucleotide composition. We found that the stronger male impact on this relationship is a conserved feature of human, mouse, dog, and sheep. In contrast, variation in genomic GC content in pig and opossum is more strongly correlated with female, rather than male, recombination rate. We also showed that the sex-differential impact of recombination is mainly driven by the chromosomal localisation of recombination events, not the overall average recombination rate. We proposed a new explanation for the evolutionary impact of heterochiasmy on nucleotide composition. This work has been submitted for publication. This work was done in collaboration with D. Mouchiroud.

6.2. Modelling the influence of karyotype on the distribution of meiotic recombination

Given the important evolutionary role of recombination, recent work has focused on understanding the dynamics of this molecular process. We analysed the variation of recombination rate among species in relation to their karyotype. Specifically, we developed a non-linear model between the total genetic and physical lengths of chromosomes. Our model incorporates important biological knowledge of the recombination process. It further allows the estimation of two main parameters of recombination: the additional recombination rate per Mb and the per-species average strength of interference. Since the model is defined on data from genetic maps, at the global level of the karyotype, it can be applied even on low-resolution data and, hence, can result in the exploration of multiple species. By analysing the variability of our models recombination parameters among species, we showed that the recombination rate and the interference strength are regulated at the Mb scale of the genomes. We found that the genome size is a strong predictor of the recombination rate, while the average physical length of chromosomes is positively correlated with the interference parameter of our model. These relations represent valuable tools for the estimation of recombination parameters even for species lacking genetic maps. This work is in submission. This work was done in collaboration with D. Mouchiroud.

6.3. Finding long and multiple repeats with edit distance

We developed an algorithm, FILMRED, for detecting long similar fragments occurring at least twice in a set of biological sequences (a conference paper [25] has already appeared, a journal version is in preparation). The problem becomes computationally challenging when a non negligible number of insertions, deletions and substitutions are allowed. The algorithm is exact and manages instances whose size and combination of parameters cannot be handled by other currently existing method. This is achieved by using a filter as a preprocessing step, and then the information that this filter has gathered in the following inference phase. FILMRED can deal with very long repeats (up to a few thousands) occurring possibly several times, with a difference rate (substitutions and indels) of 10% or more. This work was done in collaboration with N. Pisanti and P. Peterlongo. The software will be made available in a near future.

6.4. Genomics of symbiosis

Insect symbioses are model systems for studying the evolution of bacterial genomes. Importantly, evolution is directly related to the type of interaction and may also be influenced by the presence of other symbionts [11]. We are currently studying the genomes of different symbionts. The first one is the genome of the *Wolbachia* strain that has recently become obligatory for the reproduction of *Asobara tabida*. First results indicate that this genome may have been recently invaded by mobile elements. The second project concerns the sequencing of the different symbionts that co-exist in the insect *Bemisia tabaci*. The initial sequences are promising and we hope to close the genome of four different symbionts in this system, which will allow the study of the complementation between symbionts and lateral gene transfers among symbionts sharing the same intracellular arena. This work is done in collaboration among others with L. Mouton.

6.5. Bacterial syntenies

The automatic identification of syntenies across multiple species is a key step in comparative genomics that helps biologists shed light both on evolutionary and functional problems. We developed a versatile algorithm to extract all syntenies from multiple bacterial species based on a clear-cut and very flexible definition of the synteny blocks that allows for gene quorum, partial gene correspondence, gaps, and a partial or total conservation of the gene order [8]. We then applied this algorithm to two different kinds of studies. The first one is a search for functional gene associations. In this context, we compared our algorithm to a widely used heuristic - I-ADHORE - and showed that at least up to ten genomes, the problem remains tractable with our exact definition and algorithm. The second application was linked to evolutionary studies: we verified in a multiple alignment setting that pairs of orthologs in synteny are more conserved than pairs outside, thus extending a previous pairwise study. We then showed that this observation is in fact a function of the size

of the synteny: the larger the block of synteny is, the more conserved the genes are. This work was done in collaboration with F. Boyer.

6.6. Spatial synteny in Eukaryotes

Folding and intermingling of chromosomes has the potential of bringing close to each other loci that are very distant genomically or even on different chromosomes. On the other hand, genomic rearrangements also play a major role in the reorganisation of loci proximities. Whether the same loci are involved in both mechanisms has been studied in the case of somatic rearrangements, but never from an evolutionary standpoint. From the joint study of the network of spatial proximities of human genomic loci and a dataset of evolutionary breakpoints between human and mouse, we were able to provide evidence that evolutionary breakpoints tended to cluster spatially in human cells, which led us to propose the new notion of spatial synteny, which generalises the concept of genomic synteny. This work was submitted in 2010 and was accepted in 2011 [24]. It was done in collaboration with C. Lemaitre.

6.7. Chimeric transcripts in Eukaryotes

In the framework of the extension of the ENCODE project to chromosomes 21 and 22, we had the opportunity to identify a new category of transcripts, which we call chimeric transcripts, since they contain exons from different genes, which can themselves be located far apart on the same chromosome, or on different chromosomes [9]. We further found that the network formed by these connected genes is enriched in cliques of sizes 3 and 4, which seems to indicate that transcription and/or splicing of these sets of genes co-occur in time and space, as is confirmed by the confrontation of our expression dataset to a dataset indicating the co-localisation of DNA fragments in 3D. This work was done in collaboration with, among others, R. Guigó.

6.8. KisSplice: de-novo calling alternative splicing events from RNA-seq data

We addressed the problem of identifying polymorphisms in RNA-seq data when no reference genome is available, and avoiding an assembly. Based on the fundamental idea that each polymorphism will correspond to a recognisable pattern in a De Bruijn graph constructed from the RNA-seq reads, we propose a general model for all polymorphisms in such graphs. We then introduce an exact algorithm to extract alternative splicing events and show that it enables to identify more correct events than current transcriptome assemblers. Additionally, when we applied our method on a 50M reads dataset from human, we were able to identify 3884 events, out of which 57% are not present in the annotations, which confirms recent estimates showing that the complexity of alternative splicing has been largely underestimated so far. This work has been submitted to publication. This work was done in collaboration with P. Peterlongo.

6.9. Transcriptomics of symbiosis in the Asobara tabida-Wolbachia association

Wolbachia has evolved a very peculiar phenotype in the host *Asobara tabida* where it is obligatory for oogenesis. Through transcriptomic approaches (Sanger sequencing of mRNA, in vitro substraction of transcriptomes), we have established a first reference transcriptome of this insect. The analyses done demonstrate that *Wolbachia* interferes with different host pathways, and notably regulation of oxidative stress, apoptosis and autophagy, which are known to be involved in host-pathogen interactions. RNAseq has now been performed on this system and analyses are underway, which will allow a finer investigation of the interaction using the algorithm KISSPLICE (see above) developed in the EPI for the analysis of NGS data without reference genome.

6.10. Navigating through unexplored pre-miRNA candidates

The computational search for novel miRNA precursors often involves some sort of structural analysis with the aim of identifying which type of structures are prone to being recognised and processed by the cellular miRNA-maturation machinery. A natural way to tackle this problem is to perform clustering over the candidate structures along with known miRNA precursor structures. Mixed clusters allows then the identification of

candidates that are similar to known precursors. Given the large number of candidate pre-miRNAs that can be identified in single-genome approaches, even after applying several filters for robustness and stability, a conventional structural clustering approach is unfeasible. We presented a method, MINDIST, to represent candidate structures in a feature space which summarises key sequence/structure characteristics of each candidate. We demonstrated that proximity in this feature space is related to sequence/structure similarity, and we selected candidates which have a high similarity to known precursors. Additional filtering steps were then applied to further reduce the number of candidates to those with greater transcriptional potential. Our method was compared to another single-genome method (TRIPLETSVM) in two datasets, showing better performance in one and comparable performance in the other. Additionally, we showed that our approach allows for a better interpretation of the results. The MinDist method is available upon request and will be made available online. This work has been submitted to publication. This work was done in collaboration with A. T. Freitas and R. Backofen.

6.11. Exploration of the genetic network of Buchnera aphidicola

Aphids are important agricultural pests which can grow and reproduce thanks to their intimate symbiosis with the γ-proteobacterium *Buchnera aphidicola* that furnishes them with essential amino acids lacking in their phloem sap diet. We investigated how *B. aphidicola*, with its reduced genome containing very few transcriptional regulators, responds to variations in the metabolic requirements of its host by concentrating attention on the leucine metabolic pathway [23]. We showed that leucine is a limiting factor for aphid growth and displays a stimulatory feeding effect. Our metabolic analyses demonstrated that symbiotic aphids are able to respond to leucine starvation or excess by modulating the neosynthesis of this amino acid. Taken together, our data showed that the response of *B. aphidicola* to the leucine demand of its host is multimodal and dynamically regulated, providing new insights concerning the genetic regulation capabilities of this bacterium in relation to its symbiotic functions.

6.12. Annotation database system to ease the development and update of BioCyc databases

In recent years, genomes from an increasing number of organisms have been sequenced, but their annotation remains a time-consuming process. The BIOCYC databases offer a framework for the integrated analysis of metabolic networks. The PATHWAY TOOL SOFTWARE SUITE allows the automated construction of a database starting from an annotated genome, but it requires prior integration of all annotations into a specific summary file or into a GenBank file. To allow the easy creation and update of a BIOCYC database starting from the multiple genome annotation resources available over time, we developed an ad hoc data management system that we called Cyc Annotation Database System (CYCADS) [22]. The CYCADS pipeline for annotation management was used to build the ACYPICYC database for the pea aphid *Acyrthosiphon pisum*, TRICACYC for *Tribolium castaneum* and DROMECYC for *Drosophila melanogaster*. This work will be extended to create a database for other arthropods. This work was done in collaboration among others with S. Collela.

6.13. Representation and curation of metabolic pathways: UniPathway

UNIPATHWAY (http://www.unipathway.org) is a manually curated database for the representation and annotation of metabolic pathways developed in collaboration with the Swiss Institute of Bioinformatics (Swiss-Prot group). UNIPATHWAY provides explicit chemical representations of enzyme-catalysed and spontaneous chemical reactions, as well as a hierarchical representation of metabolic pathways. This hierarchy uses linear subpathways as the basic building block for the assembly of larger and more complex pathways, including species-specific pathway variants. All of the pathway data in UNIPATHWAY has been extensively cross-linked to existing pathway resources such as KEGG and METACYC, as well as sequence resources such as the UNIPROT KNOWLEDGEBASE (UNIPROTKB). UNIPATHWAY has been used to provide a controlled vocabulary for pathway annotation within UNIPROTKB records since UNIPROT release 14.7 (January 2009). In release 2011_08 of UNIPROT, UNIPATHWAY provides annotation for 118,390 distinct Swiss-Prot protein

records and 783,299 TrEMBL protein records. On the UNIPROTKB web site, each of these records is linked to the appropriate pathway description in the UNIPATHWAY web site. In 2011, the complete description of the UNIPATHWAY database has been published in *Nucleic Acids Research* (Jan. 2012 Database Issue) and has been chosen by the editors of *Nucleic Acids Research* to appear on the Featured Articles page (top 5% of *NAR* papers : http://www.oxfordjournals.org/our_journals/nar/featured_articles.html) [16].

6.14. Representation and curation of biochemical reactions: Rhea

RHEA (http://www.ebi.ac.uk/rhea) is a project developed in collaboration with the Swiss Institute of Bioinformatics (SIB) and the European Institute for Bioinformatics (EBI). It aims at providing a comprehensive resource of expert-curated biochemical reactions. RHEA provides a non-redundant set of chemical transformations for use in a broad spectrum of applications, including metabolic network reconstruction and pathway inference. RHEA includes enzyme-catalysed reactions (covering the IUBMB Enzyme Nomenclature list), transport reactions and spontaneously occurring reactions. RHEA reactions are described using chemical species from the Chemical Entities of Biological Interest ontology (ChEBI) and are stoichiometrically balanced for mass and charge. They are extensively manually curated with links to source literature and other public resources on metabolism including enzyme and pathway databases. This cross-referencing facilitates the mapping and reconciliation of common reactions and compounds between distinct resources, which is a common first step in the reconstruction of genome scale metabolic networks and models. The complete description of the database will appear in the Jan. 2012 *NAR* Database issue [5].

6.15. Metabolic reconstruction of Klebsiella pneumoniae str. Kp13

Klebsiella pneumoniae str. Kp13 is a multidrug resistant pathogen involved in nosocomial outbreaks in Brazil. The objectives of this study still underway are two-fold: (1) to perform a graph-based metabolic reconstruction of the small-molecules network of strain Kp13 and (2) from the reconstructed network evaluate what makes this pathogen so successful in colonising its human host. Manual annotation of the network was performed and a choke-point analysis was carried out, which yielded interesting targets such as L-rhamnose biosynthesis enzymes that may be related to the virulence of this bacterium. The MetAnnot platform within MetExplore was used for the manual curation of the network and graph export/import. A paper is in preparation. This work is being done in collaboration with A. T. Vasconcelos and M. Nicolás.

6.16. Clustering of elementary modes and metabolic modules identification

While it is commonly admitted that metabolism is modular, the identification of metabolic modules remains an open topic. In fact, what remains open comes even upstream of any identification problem, and refers instead to the question of defining a good model for modules in metabolic networks. One would hope that such a model might enable, for instance, to automatically derive the metabolic pathways that have been discovered and painstakingly established by biochemists over the years. Elementary modes, that are informally defined as metabolic subnetworks that can function at steady state, meaning that all internal metabolites are produced and consumed in equal rates (that is, nothing accumulates internally), represent one starting point for a definition of modules that has been considered. There are two difficulties related to this however. One is that enumerating elementary modes has itself been proven (by members of the EPI) to be a hard problem, while the second is that even small networks (around 100 reaction nodes) can have millions of elementary modes. Clustering them based on, for instance, the amount of overlap, that is of shared reactions, is one idea that has been used. We attempted another definition of modules using elementary modes that is related to a form of node-covering. This has been submitted to publication. The corresponding software is available on request. This work was done in collaboration with C. E. Ferreira, E. Moreno, and P. Crescenzi.

6.17. Enumeration of metabolic stories

In many cases, we are interested in understanding how metabolism reacts when an organism is submitted to some environmental stress, that is, to establish which metabolic processes are involved in an organism's adaptation to such stress. In order to do it, elements of the metabolism such as metabolites are monitored to determine which are over- or under-produced during the stress response as compared to the organism's state in normal conditions. Such quantitative and qualitative measurements are called metabolomics. The affected part may represent only a small portion of the network, that is, involve a small subset of metabolites. The aim then is to identify subnetworks that enable to link together all elements in this subset. More formally, given such subset and a metabolic network represented as a digraph where nodes are metabolites and edges link two metabolites when one is the input of a reaction that produces the other, we are interested in identifying all maximal directed acyclic graphs that cover all the metabolites in the subset of interest, and have no sources or targets that are not one of these metabolites. Such maximal DAGs are called metabolic stories. We established already that finding one metabolic story is easy (paper submitted) and used the algorithm developed, GOBBOLINO (see the Softwares Section), for practical purposes (second publication in preparation). This work was done in collaboration with P. Crescenzi, A. Marchetti-Spaccamela, L. Stougie, A. Marino, F. Jourdan, and L. Cottret.

6.18. Identification of all the minimal precursor sets for a given set of targets

Once the metabolic network of an organism has been defined, the question of how are produced the essential metabolites for the organism arises. In particular, it is important to know which are the metabolites that the organism needs to obtain from its environment to produce those essential metabolites. In the case of symbiosis, this environment could be the host, and determining such metabolites one way of exploring the dialog that is established between different organisms that entertain a close and often long term relation. We call such metabolites that must be obtained from the environment precursors. We had already in 2008 established the complexity of the problem of, given a network and a set of targets of interest, enumerating all minimal sets of precursors enabling to produce the targets, and given one first algorithm. The algorithm has since been much improved (journal paper submitted). The algorithm has been developed into a software called PITUFO (see the Softwares Section). This work was done in collaboration with A. Marchetti-Spaccamela, L. Stougie, and L. Cottret.

6.19. Metabolic network comparison

All living organisms have very similar metabolic needs as they all uptake nutrients from their environment, degrade them into basic building blocks such as amino acids and nucleotides, which in turn are essential input for protein and DNA synthesis. A natural expectation is therefore that they share an even reduced core of metabolic functions necessary to carry out this basic small molecule metabolism. Comparing the small molecule metabolism of 58 bacteria carefully selected and representing a range of lifestyles, we found not a single enzymatic reaction common to all of them. This absence of a metabolic core is essentially due to intracellular symbionts. These results were in preparation in 2010 and are now submitted. The work was done in a collaboration with Ludovic Cottret (INSA Toulouse) and Ana Tereza Vasconcelos.

6.20. Wolbachia detection

Wolbachia is a large monophyletic genus of intracellular bacteria, traditionally detected using PCR assays. Its considerable phylogenetic diversity and impact on arthropods and nematodes make it urgent to assess the efficiency of these screening protocols. The sensitivity and range of commonly used PCR primers and of a new set of 16S primers were evaluated on a wide range of hosts and *Wolbachia* strains [20]. We showed that certain primer sets are significantly more efficient than others but that no single protocol can ensure the specific detection of all known *Wolbachia* infections. This work was done in collaboration among others with S. Charlat.

6.21. Genetic architecture of parasite infection

The problem here is to understand the genetic architecture of a parasitic invasion by investigating the different phenotypes such invasion produces in the host. One such phenotype is called "cytoplasmic incompatibility". Briefly, when a parasite invades a male host, it induces developmental arrest, ultimately, death of the host's offspring unless the fertilised embryo carries the same symbiont inherited from its mother, that is, unless the female is also infected. This has been tentatively explained by a toxin/antitoxin model that involves a toxin deposited by the parasites in the male's sperm that induces the death of the zygote unless neutralised by an antidote produced by the parasites present in the egg. One toxin/antitoxin pair is linked to one gene. Given a set of observed cytoplasmic incompatibilities, the question is how many genes are required to explain it. Formally, and skipping many intermediate modelling steps, this translates into, given a 0/1 matrix M for pairs of male/female (a 0 indicating that either the male is not infected or, if it is, then so is the female meaning that there is no incompatibility, and a 1 indicating that the male is infected while the female is not), what is the minimum number of "rectangles" that enable to cover all the 1s in M? A "rectangle" in this case is a subset of columns and rows such that, if permuted, they can be arranged in the form of a rectangle with only 1s (publication Nor et al., 2010 by the BAMBOO Team and collaborators). One rectangle corresponds to a gene. The model can then be made more complex by considering that genes may have different alleles (different forms), and are expressed in variable quantities. The quantitative version of the problem in particular translates into having to find a minimum number of "triangles" that cover all 1s. All the above problems translate also into different versions of edge covers of a bipartite graph that are for the most part algorithmically original, and always not fully resolved (meaning, there remains open questions, notably regarding complexity). Work on these problems within the Associate Team SIMBIOSI and the results already obtained should lead to a publication in 2012. This work was done in collaboration among others with S. Charlat.

6.22. Population structure and dynamics

We recently started a collaboration with the Pasteur Institute in Cambodia (Dr. P. Buchy) and the CIRAD at Montpellier (Dr. R. Frutos) on viral population structure and dynamics. In this context, we developed an exploratory statistical approach to characterise mutational patterns in viral populations. The basic idea is to use Multiple Correspondence Analysis (MCoA) on a multiple alignment of nucleic sequences. To this purpose, the alignment is encoded as a boolean table where rows correspond to the sequences and columns to the presence/absence of characters. This can be done simply by considering each of the four possible bases as a different character, or by considering only two possible states: one for the major base and one for all other (minor) bases. This technique turned out to be very effective in representing co-mutation patterns within a population of sequences. As for simple CoA, the plot is quite easy to interpret by biologists, both in terms of proximities of sequences and characters (*i.e.* mutations). Moreover, it has some strong relationships with parsimony phylogeny that need to be clarified. In [10], we applied this technique to study the structure and time evolution of the Dengue virus (serotype I) population in Cambodia, using heterochronous sequence samples. Beside its methodological aspect, this work also introduced new topics related to population genetics in BAMBOO (*e.g.* the use of coalescent theory to reconstruct population dynamics). Another approach, widely used in ecology, to measure biodiversity and to determine the species composition of environmental samples is the "DNA-barcoding" technique. In collaboration with the Laboratoire d'Ecologie Alpine (Univ. Joseph Fourier Grenoble), we developed a new software for identifying new barcode markers and their associated PCR primers [19].

7. Partnerships and Cooperations

7.1. National Initiatives

7.1.1. Adapthantroph

- Title: Adaptation des insectes aux anthroposystèmes

- Coordinator: M. Harry
- BAMBOO participant(s): C. Vieira
- Type: ANR Génoplante (2009-2012)
- Web page: Not available

7.1.2. Alcovna

- Title: ALgorithms for COmparing and Visualzing Non Assembled data
- Coordinator: Pierre Peterlongo
- BAMBOO participant(s): J. Kielbassa, V. Lacroix, G. Sacomoto, M.-F. Sagot
- Type: ARC INRIA (2010-2011)
- Web page: http://alcovna.genouest.org/

7.1.3. AphiCible

- Title: Impact de la recombinaison et de la conversion génique biaisée sur l'évolution de génomes
- Coordinator: Y. Rahbé
- BAMBOO participant(s): Y. Rahbé and H. Charles
- Type: ANR Génoplante (2008-2011)
- Web page: Not available

7.1.4. Cogebi

- Title: Symbiosis, digestion and reproduction as aphid physiological processes to identify new targets for insecticides
- Coordinator: L. Duret (LBBE)
- BAMBOO participant(s): C. Gautier, E. Tannier
- Type: ANR Génomique Animale (2008-2011)
- Web page: Not available

7.1.5. ImmunSymbArt

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

7.1.6. 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: none

7.1.7. NeMo

- Title: Network Motifs

- Coordinator: S. Robin (AgroParisTech, Paris)
- BAMBOO participant(s): V. Lacroix, M.-F. Sagot
- Type: ANR Blanc (2008-2011)
- Web page: http://nemo.ssbgroup.fr/

7.1.8. MIRI

- Title: Mathematical Investigation of "Relations Intimes"
- Coordinator: M.-F. Sagot
- BAMBOO participant(s): V. Acuña, C. Baudet, C. Gautier, V. Lacroix, P. Milreu, C. Klein, I. Nor, M.-F. Sagot, P. Simões
- Type: ANR Blanc (2009-2012)
- Web page: http://pbil.univ-lyon1.fr/members/sagot/htdocs/team/projects/miri/miri.html

7.2. European Initiatives

7.2.1. FP7 Project

7.2.1.1. SISYPHE

- 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

- 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

7.2.2. Collaborations in European Programs, except FP7

7.2.2.1. METNET4SysBio

- Title: System level analysis of animal metabolism by multicompartment graph- and constraint-based modelling
- Coordinator: H. Charles (INSA Lyon, France)
- BAMBOO participant(s): V. Acuña, H. Charles, C. Gautier, V. Lacroix, Y. Rahbé, M.-F. Sagot
- European Partner: Angela Douglas, York University, UK
- Type: ANR-BBSRC BioSys (2007-2011)

7.2.2.2. SIMBIOSI

- Title: Mathematical and algorithmic investigation of symbiosis
- Coordinators: M.-F. Sagot (France), A. Marchetti-Spaccamela (Italy), L. Stougie (the Netherlands)
- BAMBOO participant(s): Whole BAMBOO Team
- Type: Associated Team INRIA (2009-2011)
- Web page: http://pbil.univ-lyon1.fr/members/sagot/htdocs/team/projects/simbiosi/simbiosi.html

7.2.3. Major European Organizations with which you have followed Collaborations

Partner 1: Pierluigi Crescenzi, Univ. Florence, Italy
Algorithmic (graphs, trees, sequences), complexity

Partner 2: Ana Teresa Freitas, INESC-ID, IST Lisbon, Portugal
NGS, metabolism, small RNAs, motifs

Partner 3: Alberto Marchetti-Spaccamela, Univ. Rome La Sapienza, Italy
Algorithmic (graphs, trees), complexity

Partner 4: Nadia Pisanti and Roberto Grossi, Univ. Pisa, Italy
Algorithmic (graphs, trees, sequences)

Partner 5: Leen Stougie, Free Univ. Amsterdam and CWI, the Netherlands
Algorithmic (graphs, trees), complexity

7.3. International Initiatives

7.3.1. INRIA International Partners: AMICI

- Title: Algorithms and Mathematics for Investigating Communication and Interactions intra- and inter-organisms
- Coordinators: M.-F. Sagot (France), A. Marchetti-Spaccamela (Univ. Rome, Italy), L. Stougie (Free Univ. Amsterdam and CWI, the Netherlands), P. Crescenzi, Univ. Florence, Italy), N. Pisanti (Univ. Pise, Italy)
- BAMBOO participant(s): Whole BAMBOO Team
- Type: INRIA International Partner
- Web page: http://piluc.dsi.unifi.it/amici/

7.3.2. INRIA-Faperj (Brazil) project: RAMPA

- Title: Bioinformatics for the Reconstruction and Analysis of the Metabolism of PArasites
- Coordinators: M.-F. Sagot (France), A. T. Vasconcelos (LNCC, Brazil)
- BAMBOO participant(s): Whole BAMBOO Team
- Type: Faperj-INRIA
- Web page: not yet available

7.4. Exterior research visitors

Etienne Birmelé, Associate Professor, University of Évry, France, various visits of 1 week until délégation in Sept. 2011
Pierluigi Crescenzi, Professor, University of Florence, Italy, various visits of 1-2 weeks

Roberto Grossi, Professor, University of Pisa, Italy, various visits of 1 week
Alberto Marchetti-Spaccamela, Professor, University La Sapienza, Rome, Italy, visit of 1 week
Andrea Marino, PhD student (Supervisor: Pierluigi Crescenzi), University of Florence, Italy, various visits of 1-2 weeks
Eduardo Moreno, Associate Professor, University Adolfo Ibañez, Chile, visit of 1 week
Nadia Pisanti, Associate Professor, University of Pisa, Italy, various visits of 1 week
Gianluca Rossi, Associate Professor, University of Rome Tor Vergata, Italy, visit of 1 week
Leen Stougie, Free University Amsterdam and CWI, Amsterdam, the Netherlands, visit of 1 week
Ana Tereza Vasconcelos, CNPq Grant, Lab Nacional de Computação Científica, Petrópolis, Brazil, visit of 1 year from Sept. 1st, 2010 until Aug. 31st, 2011
Susana Vinga, Professor, INESC-ID, IST Lisbon, Portugal, visit of 1 week
Maria Emilia Walter Telles, University of Brasília, Brazil, visit of 3 months

8. Dissemination

8.1. Animation of the scientific community

Hubert Charles is director of studies of the "Bioinformatique et Modélisation (BIM)" track at the INSA-Lyon. He is co-director of the Biosciences Department of the INSA-Lyon, and co-director of the Doctoral School E2M2. He was member of the PhD committee of Léo Coutellec (INSA-Lyon).

Christian Gautier is Director of the PRABI. He is member of the "Conseil d'administration" of the University of Lyon 1, and of the selecting committee for biology at the École Polytechnique, Paris. He is deputy director of the IFR 41 "Bio-environnement et santé" of the University of Lyon 1.

Vincent Lacroix was a member of the PhD committee of Florian Sikora (University of Paris-Est).

Yvan Rahbé is director of the INSA-Lyon BF2I laboratory. He was reviewer of the PhDs of Julien Chuche (University of Bordeaux), Sandrine Nsango (University of Strasbourg) and Benjamin Fartek (University of the Réunion).

Marie-France Sagot is a member of the Scientific Advisory Board ("Conseil Scientifique (COS") for the INRIA Grenoble Rhône-Alpes Research Center. She was co-chair of the 11th Workshop on Algorithms and BInformatics (WABI), at the Max-Planck Institute, Saarbrücken, Germany, September 5-7 (http://pbil.univ-lyon1.fr/members/sagot/htdocs/wabi2011/wabi2011.html) and is co-chair for one area track of the 20th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) in 2012. She is Editor-in-Chief of *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, and Associate Editor of *BMC Bioinformatics*, *Algorithms for Molecular Biology*, *Journal of Discrete Algorithms*, and *Lecture Notes in BioInformatics*. She is member of the Steering Committee for the European Conference on Computational Biology (ECCB), for the International Symposium on Bioinformatics Research and Applications (ISBRA), and for the Latin American Theoretical Informatics Symposium (LATIN). She was member of the Program Committee for CiE, PSC, RECOMB, SPIRE. She was member of the HDR committee of Etienne Birmelé (University of Évry) and of the PhD committee of David Parsons (INSA-Lyon). She was reviewer of the HDR of Pierre Brézellec (University of Évry), and of the PhDs of Anne-Laure Gaillard (University of Bordeaux) and Gilles Vieira (University of Évry).

Fabrice Vavre is director of the GDR 2153(CNRS) "Interactions multipartenaires dans les populations et les communautés d'insectes". He is also member of the management committee and responsible of a working group in the COST Action FA0701 "Arthropod Symbiosis: from fundamental studies to pest and disease management". He was reviewer of the HDR of Matthieu Sicard (University of Poitiers), and of the PhDs of Magali Thierry (University of the Réunion) and Emilie Dion (University of Rennes).

Alain Viari was, until May 2011, the scientific deputy of the INRIA Grenoble Rhône-Alpes Research Center. He is member of the scientific advisory board of the MIA (Mathematics and Applied Mathematics) at the INRA and of the IMMI (Institut de Microbiologie et Maladies Infectieuses / Aviesan). He is also a member of the Steering Committee of the IRT (Institut de Recherche Technologique) Lyon-Biotech. Since May 2011, he is a member of the Scientific Advisory Board ("Conseil Scientifique (COS") for the INRIA Grenoble Rhône-Alpes Research Center. He was reviewer of the PhD of Giovanni Battaglia (supervisor: Roberto Grossi, Univ. of Pisa, Italy).

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

8.2. Teaching

Four members of the BAMBOO project are professors or associate professors at the University Claude Bernard in Lyon and at the INSA Lyon: Hubert Charles, Christian Gautier (émérite since Oct. 2011), Vincent Lacroix, and Cristina Vieira. They therefore have a full teaching service (at least 192 hours) except for Cristina Vieira who became since 2010 a Junior Member of the Institut Universitaire de France.

Various members of the EPI have developed over the years courses in biometry, bioinformatics and evolutionary biology at all levels of the University as well as at the "École Normale Supérieure" (ENS) of Lyon and the INSA ("Institut National de Sciences Appliquées"). Two members of the EPI have also in the past participated in, or sometimes organised courses or teaching modules at the international level: creation and support of a Master's course in Ho-Chi-Minh, Vietnam, and creation and direction of a PhD Program in Computational Biology in Lisbon, Portugal (http://bc.igc.gulbenkian.pt/pdbc/).

The following are the HDR and PhDs defended in BAMBOO in 2011.

HdR: Etienne Birmelé, Étude structurelle des réseaux: modèles aléatoires, motifs et cycles, Université d'Évry, November 3

PhD: Alexandra Carvalho, Motif representation and discovery, IST, July 13, supervisors A. Oliveira and M.-F. Sagot

PhD: Nuno Mendes, Efficient algorithms for the identification of miRNA motifs in DNA sequences, IST, June 6, supervisors A. T. Freitas and M.-F. Sagot

PhD: Alexandra Popa, The evolution of recombination and genomic structures: a modeling approach, Université Lyon 1, May 24, supervisors C. Gautier and D. Mouchiroud

9. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] E. BIRMELÉ. *Etude structurelle des réseaux: modèles aléatoires, motifs et cycles*, Université d'Évry Val d'Essonne, 2011, Habilitation, Ph. D. Thesis.

[2] A. CARVALHO. *Motif representation and discovery*, Instituto Superior Técnico Lisbon, 2011, Supervisors: A. Oliveira and M.-F. Sagot.

[3] N. MENDES. *Efficient algorithms for the identification of miRNA motifs in DNA sequences*, Université Lyon 1 and Instituto Superior Técnico Lisbon, 2011, Supervisors: M.-F. Sagot and A. T. Freitas.

[4] A. POPA. *The evolution of recombination and genomic structures: a modeling approach*, Université Lyon 1, 2011, Supervisors: C. Gautier and D. Mouchiroud.

Articles in International Peer-Reviewed Journal

[5] R. ALCÁNTARA, K. B. AXELSEN, A. MORGAT, E. BELDA, E. COUDERT, A. BRIDGE, H. CAO, P. DE MATOS, M. ENNIS, S. TURNER, G. OWEN, L. BOUGUELERET, I. XENARIOS, C. STEINBECK. *Rhea–a manually curated resource of biochemical reactions*, in "Nucleic Acids Res", Dec 2011.

[6] S. BERTHOUMIEUX, M. BRILLI, H. DE JONG, D. KAHN, E. CINQUEMANI. *Identification of metabolic network models from incomplete high-throughput datasets*, in "Bioinformatics", 2011, vol. 27, n^{o} 13, p. I186-I195, 19th Annual International Conference on Intelligent Systems for Molecular Biology/10th European Conference on Computational Biology [*DOI :* 10.1093/BIOINFORMATICS/BTR225].

[7] A. BRANCA, B. P. LE RU, F. VAVRE, J.-F. SILVAIN, S. DUPAS. *Intraspecific specialization of the generalist parasitoid Cotesia sesamiae revealed by polyDNAvirus polymorphism and associated with different Wolbachia infection*, in "Molecular Ecology", 2011, vol. 20, n^{o} 5, p. 959-971 [*DOI :* 10.1111/J.1365-294X.2010.04977.X].

[8] Y.-P. DENIELOU, M.-F. SAGOT, F. BOYER, A. VIARI. *Bacterial syntenies: an exact approach with gene quorum*, in "BMC Bioinformatics", 2011, vol. 12 [*DOI :* 10.1186/1471-2105-12-193].

[9] S. DJEBALI, J. LAGARDE, P. KAPRANOV, V. LACROIX, J. MUDGE, C. UCLA, S. FOISSAC, C. HOWALD, J. CHRAST, P. RIBECA, D. MARTIN, R. MURRAY, X. YANG, L. GHAMSARI, C. LIN, I. BELL, E. DUMAIS, J. DRENKOW, N. VAN BERKUM, B. LAJOIE, M. VIDAL, J. STAMATOYANNOPOULOS, J. HARROW, T. HUBBARD, J. DEKKER, A. FRANKISH, K. SALEHI-ASHTIANI, A. REYMOND, S. ANTONARAKIS, R. GUIGÓ, T. GINGERAS. *Evidence for Transcript Networks composed of chimeric RNAs in human cells*, in "PLoS One", 2011, to appear.

[10] V. DUONG, C. SIMMONS, L. GAVOTTE, A. VIARI, S. ONG, N. CHANTHA, N. J. LENNON, B. W. BIRREN, S. VONG, J. J. FARRAR, M. R. HENN, V. DEUBEL, R. FRUTOS, P. BUCHY. *Genetic diversity and lineage dynamic of dengue virus serotype 1 (DENV-1) in Cambodia*, in "Infect Genet Evol", Jul 2011.

[11] J. FERRARI, F. VAVRE. *Bacterial symbionts in insects or the story of communities affecting communities*, in "Philosophical Transactions of the Royal Society B-Biological Sciences", 2011, vol. 366, n^{o} 1569, p. 1389-1400 [*DOI :* 10.1098/RSTB.2010.0226].

[12] M. GALARDINI, A. MENGONI, M. BRILLI, F. PINI, A. FIORAVANTI, S. LUCAS, A. LAPIDUS, J.-F. CHENG, L. GOODWIN, S. PITLUCK, M. LAND, L. HAUSER, T. WOIKE, N. MIKHAILOVA, N. IVANOVA, H. DALIGAULT, D. BRUCE, C. DETTER, R. TAPIA, C. HAN, H. TESHIMA, S. MOCALI, M. BAZZICALUPO, E. G. BIONDI. *Exploring the symbiotic pangenome of the nitrogen-fixing bacterium Sinorhizobium meliloti*, in "BMC Genomics", 2011, vol. 12 [*DOI :* 10.1186/1471-2164-12-235].

[13] R. A. GUTIÉRREZ, A. VIARI, B. GODELLE, R. FRUTOS, P. BUCHY. *Biased mutational pattern and quasispecies hypothesis in H5N1 virus*, in "Infect Genet Evol", Oct 2011.

[14] K. KRAAIJEVELD, B. M. REUMER, L. MOUTON, N. KREMER, F. VAVRE, J. J. M. VAN ALPHEN. *Does a parthenogenesis-inducing Wolbachia induce vestigial cytoplasmic incompatibility?*, in "Naturwissenschaften", 2011, vol. 98, n^{o} 3, p. 175-180 [*DOI :* 10.1007/S00114-010-0756-X].

[15] E. LERAT, N. BURLET, C. BIEMONT, C. VIEIRA. *Comparative analysis of transposable elements in the melanogaster subgroup sequenced genomes*, in "Gene", 2011, vol. 473, n° 2, p. 100-109 [*DOI :* 10.1016/J.GENE.2010.11.009].

[16] A. MORGAT, E. COISSAC, E. COUDERT, K. B. AXELSEN, G. KELLER, A. BAIROCH, A. BRIDGE, L. BOUGUELERET, I. XENARIOS, A. VIARI. *UniPathway: a resource for the exploration and annotation of metabolic pathways*, in "Nucleic Acids Res", Nov 2011.

[17] A. NECSULEA, A. POPA, D. N. COOPER, P. D. STENSON, D. MOUCHIROUD, C. GAUTIER, L. DURET. *Meiotic Recombination Favors the Spreading of Deleterious Mutations in Human Populations*, in "Human Mutation", 2011, vol. 32, n° 2, p. 198-206 [*DOI :* 10.1002/HUMU.21407].

[18] L. PALMEIRA, S. PENEL, V. LOTTEAU, C. RABOURDIN-COMBE, C. GAUTIER. *PhEVER: a database for the global exploration of virus-host evolutionary relationships*, in "Nucleic Acids Research", 2011, vol. 39, n° Database issue, p. D569-75 [*DOI :* 10.1093/NAR/GKQ1013], http://hal.inria.fr/hal-00562509/en.

[19] T. RIAZ, W. SHEHZAD, A. VIARI, F. POMPANON, P. TABERLET, E. COISSAC. *ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis*, in "Nucleic Acids Res", Nov 2011, vol. 39, n° 21, e145.

[20] P. M. SIMÕES, G. MIALDEA, D. REISS, M.-F. SAGOT, S. CHARLAT. *Wolbachia detection: an assessment of standard PCR Protocols*, in "Molecular Ecology Resources", 2011, vol. 11, n° 3, p. 567-572 [*DOI :* 10.1111/J.1755-0998.2010.02955.X].

[21] UNIPROT CONSORTIUM. *Ongoing and future developments at the Universal Protein Resource*, in "Nucleic Acids Res", Jan 2011, vol. 39, n° Database issue, p. D214-9.

[22] A. F. VELLOZO, A. S. VÉRON, P. BAA-PUYOULET, J. HUERTA-CEPAS, L. COTTRET, G. FEBVAY, F. CALEVRO, Y. RAHBÉ, A. E. DOUGLAS, T. GABALDÓN, M.-F. SAGOT, H. CHARLES, S. COLELLA. *CycADS: an annotation database system to ease the development and update of BioCyc databases*, in "Database (Oxford)", 2011, vol. 2011, bar008 [*DOI :* 10.1093/DATABASE/BAR008], http://hal.inria.fr/hal-00632965/en.

[23] J. VIÑUELAS, G. FEBVAY, G. DUPORT, S. COLELLA, J.-M. FAYARD, H. CHARLES, Y. RAHBÉ, F. CALEVRO. *Multimodal dynamic response of the Buchnera aphidicola pLeu plasmid to variations in leucine demand of its host, the pea aphid Acyrthosiphon pisum*, in "Molecular Microbiology", 2011, vol. 81, n° 5, p. 1271-1285 [*DOI :* 10.1111/J.1365-2958.2011.07760.X].

[24] A. VÉRON, C. LEMAITRE, C. GAUTIER, V. LACROIX, M.-F. SAGOT. *Close 3D proximity of evolutionary breakpoints argues for the notion of spatial synteny*, in "BMC Genomics", 2011, vol. 12, n° 1, 303 [*DOI :* 10.1186/1471-2164-12-303], http://hal.inria.fr/inserm-00617206/en.

International Conferences with Proceedings

[25] M. FEDERICO, P. PETERLONGO, N. PISANTI, M.-F. SAGOT. *Finding Long and Multiple Repeats with Edit Distance*, in "The Prague Stringology Conference 2011", Prague, Czech Republic, 2011, http://hal.inria.fr/inria-00608208/en.

[26] T. SACOMOTO, G. AKIO, A. PEREIRA DO LAGO. *A New Algorithm for Sparse Suffix Trees*, in "Brazilian Symposium on Bioinformatics", O. NORBERTO DE SOUZA, G. TELLES, M. PALAKAL (editors), Lecture Notes in Computer Science, Springer, 2011, vol. 6832, p. 34-41.

Conferences without Proceedings

[27] V. ACUÑA, E. BIRMELÉ, L. COTTRET, P. CRESCENZI, F. JOURDAN, V. LACROIX, A. MARCHETTI-SPACCAMELA, A. MARINO, P. VIEIRA MILREU, M.-F. SAGOT, L. STOUGIE. *Telling Stories*, in "ICALP 2011 satellite workshop: Graph Algorithms and Applications (GA)", 2011, Talk.

[28] P. BAA-PUYOULET, A. F. VELLOZO, J. HUERTA-CEPAS, G. FEBVAY, T. GABALDON, M.-F. SAGOT, H. CHARLES, S. COLELLA. *ArthropodaCyc: a BioCyc database powered by CycADS to study and compare the metabolism of arthropods*, in "5th Annual Arthropod Genomics Symposium “Arthropod Genomics 2011: Exploring Diversity, Relating Similarity”", 2011, Poster.

[29] C. BAUDET, Z. DIAS. *Partial Enumeration of Traces of Solutions for the Problem of Sorting by Signed Reversals*, in "ACM Conference on Bioinformatics, Computational Biology and Biomedicine", 2011, Poster.

[30] M.-C. CARPENTIER, J. CHARIF, V. LACROIX, M.-F. SAGOT, F. VAVRE. *RNA-seq without a reference genome: a comparison of the mapping and the assembly approaches*, in "JOBIM", 2011, Poster.

[31] C. E. FERREIRA, A. S. FREIRE, G. A. PUGLIA. *A dynamic programming algorithm for the tree mapping problem*, in "VI Latin-America Graphs and Algorithms Sysmposium (LAGOS)", 2011, Talk.

[32] P. IVAN, C. KLEIN, L. COTTRET, A. T. VASCONCELOS, A. GALES, M.-F. SAGOT, M. NICOLÁS. *Initial metabolic reconstruction of Klebsiella pneumoniae Kp13, a multidrug resistant pathogen involved in nosocomial outbreaks in Brazil*, in "X-Meeting", 2011, Poster.

[33] C. KLEIN, L. COTTRET, H. CHARLES, C. GAUTIER, A. T. VASCONCELOS, V. LACROIX, M.-F. SAGOT. *Metabolic network comparison of bacteria in the context of symbiosis*, in "Otto Warburg International Summer School and Research Symposium on Evolutionary Genomics", 2011, Poster.

[34] V. LACROIX. *KisSplice : de-novo calling alternative splicing events from RNA-seq data*, in "Groupement De Recherche Européen - Arc Rhône Alpin (GDRE-RA)", 2011, http://gdrera.univ-lyon1.fr/.

[35] V. LACROIX. *NGS without a reference genome, comparison and visualisation of non assembled data*, in "Statistical Methods for Post-Genomic Data (SMPGD)", 2011, http://smpgd2011.curie.fr/.

[36] A. RABATEL, G. FEBVAY, K. GAGET, G. DUPORT, N. BENDRIDI, Y. RAHBÉ, P. BAA-PUYOULET, H. CHARLES, F. CALEVRO, S. COLELLA. *Integration of metabolic pathways and host regulation in symbiosis: tyrosine and cuticular protein biosynthesis during the pea aphid embryonic development*, in "Sixth International Symposium on Molecular Insect Science", 2011, Poster.