IN PARTNERSHIP WITH:
CNRS

Université Rennes 1

Activity Report 2011

Project-Team SYMBIOSE

Biological systems and models, bioinformatics and sequences

IN COLLABORATION WITH: Institut de recherche en informatique et systèmes aléatoires (IRISA)

RESEARCH CENTER
Rennes - Bretagne-Atlantique

THEME
Computational Biology and Bioinformatics

Table of contents

1. Members **1**
2. Overall Objectives **2**
3. Scientific Foundations **2**
3.1. Sequence and Structure Modeling 2
3.2. System Biology 3
3.3. High Performance Computing 3
4. Application Domains **3**
4.1. Data and knowledge management 3
4.2. Comparative genomics 4
4.3. From structural analysis to systems biology 4
5. Software **4**
5.1. Main softwares 4
5.1.1. Biomaj : Data synchronization and processing workflow 4
5.1.2. GASSST: Short reader mapper for large genomic dataset 4
5.1.3. Protomata learner: fine characterization of protein families 5
5.2. Bioinformatics community tools 5
5.3. Parallel softwares 5
5.4. Softwares for Next Generation Sequencing data 6
5.5. Genome structure 6
5.6. Protein sequence and structure 6
5.7. Systems biology 6
6. New Results **7**
6.1. Advanced tools for data management 7
6.2. Sequences assembly, alignement and comparison 7
6.3. Genome Structure 8
6.4. Protein Sequences and Structures 9
6.5. Systems Biology 9
7. Contracts and Grants with Industry **10**
7.1. Contracts with Industry 10
7.2. Grants with Industry 10
8. Partnerships and Cooperations **10**
8.1. Regional Initiatives 10
8.1.1. Genopole initiatives 10
8.1.2. Partnership with INRA 11
8.2. National Initiatives 11
8.2.1. ANR contracts 11
8.2.1.1. BIOWIC 11
8.2.1.2. LEPIDOLF 11
8.2.1.3. MAPPI 11
8.2.1.4. PELICAN 11
8.2.1.5. ECS 12
8.2.1.6. BIOTEMPO 12
8.2.2. Programs from research institutions 12
8.2.3. Transfert and service ressources - GenOuest ressource center 12
8.3. European Initiatives 13
8.4. International Initiatives 13
8.4.1. INRIA Associate Teams 13
8.4.2. INRIA International Partners 14
8.4.3. Visits of International Scientists 14

8.4.4. Participation In International Programs 14
8.4.4.1. Chile. Inria-Conycit 2011-12 14
8.4.4.2. Argentina - MinCYT-Inria 2011-12 14
8.4.4.3. Germany. Egide Procope Program 2011-12 14
9. Dissemination **15**
9.1. Animation of the scientific community 15
9.1.1. Administrative functions: scientific committees, journal boards, jury 15
9.1.2. Conference program committees 15
9.1.3. Meeting organization and scientific animation 16
9.1.4. Jury of PhD Theses 16
9.2. Teaching 16
10. Bibliography **17**

Project-Team SYMBIOSE

Keywords: Computational Biology, Genetic Networks, Next Generation Sequencing, Genomics, Protein Structure

1. Members

Research Scientists

Dominique Lavenier [Team leader, Senior Researcher, Cnrs, HdR]
François Coste [Junior researcher, Inria]
Claire Lemaitre [Junior researcher, Inria]
Jacques Nicolas [Senior researcher, Inria, HdR]
Pierre Peterlongo [Junior researcher, Inria]
Anne Siegel [Senior researcher, Cnrs, HdR]

Faculty Members

Rumen Andonov [Professor, Univ. Rennes 1, HdR]
Catherine Belleannée [Associate Professor, Univ. Rennes 1]
Michel Le Borgne [Associate Professor, Univ. Rennes 1]
Raoul Vorc'h [Associate Professor, Univ. Rennes 1]
Antonio Mucherino [Associate Professor, Univ. Rennes 1, since sept. 2011]

External Collaborators

Nathalie Theret [Research director, INSERM, Rennes, HdR]
Jérémie Bourdon [Associate Professor, Univ. Nantes]

Technical Staff

Olivier Collin [GENOUEST, permanent senior Research engineer, Cnrs]
Charles Deltel [GENOUEST, permanent senior Research engineer, Inria, 50% time]
Olivier Sallou [GENOUEST, permanent senior Research engineer, Univ. Rennes 1]
Fabrice Legeai [permanent Engineer, INRA, 20% time dedicated to the symbiose project]
François Moreews [permanent Engineer, INRA, 20% time dedicated to the symbiose project]
Anthony Bretaudeau [GENOUEST, non permanent junior engineer, ANR PELICAN]
Delphine Naquin [GENOUEST, non permanent junior engineer, Région Bretagne grant]
Jonathan Piat [GENOUEST, non permanent junior engineer, ANR grant (BioWIC), until Aug. 2011]
Aurélien Roult [GENOUEST, non permanent junior engineer, Région Bretagne grant]
Romaric Sabas [GENOUEST, non permanent junior engineer, Inria grant (ADT BioMAJ), until sept. 2011]
Andres Burgos [non permanent junior engineer, ADT Inria, until nov. 2011]
Pavel Senin [non permanent junior engineer, ANR Lepidolf, until sept. 2011]
Claudia Hériveau [non permanent junior engineer, ADT Inria, since nov. 2011]
Erwan Drezen [non permanent engineer, KoriPlast, since June 2011]

PhD Students

Mathilde Le Boudic-Jamin [MENRT, since oct. 2011]
Gaëlle Garet [Région, since oct. 2011]
Sylvain Prigent [MENRT, since oct. 2011]
Santiago Videla [CNRS/ANR BIOTEMPO, since sept. 2011]
Oumarou Abdou-Arbi [MENRT]
Geoffroy Andrieux [MENRT]
Guillaume Chapuis [Région Bretagne/ENS]
Guillaume Rizk [MENRT, until apr. 2011]
Valentin Wucher [Région Bretagne/INRA, since nov. 2011]
Andres Aravena [Chilean fundings]

Rayan Chikhi [MENRT/ENS]
Matthias Gallé [Inria/CORDI, until feb. 2011]
Nicolas Maillet [ANR Mappi]

Post-Doctoral Fellows

Pavlos Antoniou [ARC Alcovna, until oct. 2011]
Pierre Blavy [INRA, ASC]
Raluca Uricaru [INRA project PEAPOL, since sept. 2011]
Thomas Derrien [INRA project Myzus, since nov. 2011]
Brivael Trelhu [ATER, Univ. Rennes 2, since oct. 2011]

Administrative Assistant

Marie-Noëlle Georgeault [Assistant, Inria]

2. Overall Objectives

2.1. A Bioinformatics Center

Symbiose is a bioinformatics research project. It focuses on methodological research at the interface between computer science and molecular biology. The project addresses both the pragmatic needs of high throughput resource management and the longer-term needs of the development of original algorithms and applications through dedicated researches.

The Symbiose team gathers two entities: a research group and a technical platform, called GenOuest.

- The *research group* focuses on high performance computing for large-scale genomic data and modeling of large-scale biological systems. Research activities cover sequence comparison, Next Generation Sequence processing, comparative genomic, identification of genome structures, structural biology, dynamic systems, and gene regulation network.
- The *GenOuest platform* belongs to Biogenouest, the French west life science network. Since 2009, it also belongs to the IBiSA[1] network and is certified ISO 9001:2008. The platform coordinates the activities of the RENABI-GO[2] regional center , one of the six French bioinformatics resource centers, and offers different bioinformatics services: computing power, storage, databanks, development, training, etc.

Both entities tightly collaborate to offer a full technological, research and training support to the biological community. Research and technological development projects are conducted in collaboration with INRA, Inserm and CNRS biological teams.

This environment offers the opportunity to locally mix computer scientists with strong expertise in high performance computing and dynamical modeling, together with genomic research labs. In the competitive field of molecular biology, we are concerned with the storage, analysis and interpretation of large-scale and multi-timescales datasets produced by other platforms and research teams, including –although not exclusively – the analysis of Next Generation Sequencing (NGS) data.

3. Scientific Foundations

3.1. Sequence and Structure Modeling

This track concerns the search for relevant (e.g. functional) spatial or logical structures in macromolecules, either with intent to model specific spatial structures (secondary and tertiary structures, disulfide bounds, ...) or

[1] GIS IBiSA: Infrastructures en Biologie Sante et Agronomie
[2] RENABI: Réseau national des plates-formes de bioinformatique, GO: Grand Ouest

general biological mechanisms (transposition, ...). In the framework of **language theory and combinatorial optimization**, we address various types of problems: design of grammatical models on biological sequences and machine learning of grammatical models from sequences; efficient filtering and model matching in data banks; protein structure prediction.

Corresponding disciplinary fields are language theory, algorithmic on words, machine learning, data analysis and combinatorial optimization.

3.2. System Biology

We address the question of constructing accurate models of biological systems with respect to available data and knowledge. The availability of high-throughput methods in molecular biology has led to a tremendous increase of measurable data along with resulting knowledge repositories, gathered on the web (e.g. KEGG,MetaCyc, RegulonDB). However, both measurements as well as biological networks are prone to incompleteness, heterogeneity, and mutual inconsistency, making it highly non-trivial to draw biologically meaningful conclusions in an automated way. Based on this statement, we develop methods for the analysis of large-scale biological networks which formalize various reasoning modes in order to highlight incomplete regions in a regulatory model and to point at network products that need to be activated or inactivated to globally explain the experimental data. We also consider small-scale biological systems for a fine understanding of conclusions that can be drawn on active pathways from available data, working on deducible properties rather than simulation.

Corresponding disciplinary fields are model checking, constraint-based analysis and dynamical systems.

3.3. High Performance Computing

HPC for bioinformatics aims to bring efficient computing solutions in the two following challenging areas: processing of high throughput genomic data and processing of computational intensive algorithms. These two areas have in common to be highly time-consuming, but for different reasons. The first has to handle very huge amounts of data while the second has to solve very large optimization problems. More precisely, the first area required to organize, to structure, to index, and more generally, to manipulate very large data structures, making memory management issue a real challenge. The second area involves high complexity algorithms, but on a much more reduced data set.

In both cases, space and time limitations can be pushed away by the use of parallel machines as they can provide large aggregated memory space and/or high computational power. We believe that the design of parallel and optimized parallel algorithms is a key issue to face the avalanche of genomic data, and that all forms of parallelism must be exploited, from cloud computing to hardware accelerators such as GPGPU.

Corresponding disciplinary fields are optimization, parallelism, processing mass of data, hardware accelerator and advanced indexing structures.

4. Application Domains

4.1. Data and knowledge management

Multiple technologies are producing raw data that have to be cleared and assembled into meaningful observations. It is the realm of statistical studies, with sophisticated normalization procedures, most of them being included in routine treatments. Information is produced in a highly distributed way, in each laboratory. Standardization, structuring of data banks, detection of redundancies and inconsistencies, integration of several sources of data and knowledge, extraction of knowledge from texts, all these are very crucial tasks for bioinformatics [41]. High throughput techniques are also a source of algorithmic issues (assembling of fragments, design of probes).

4.2. Comparative genomics

Referring to a set of already known sequences is the most important method for studying new sequences, in the search for homologies. The basic issue is the alignment of a set of sequences, where one is looking for a global correspondence between positions of each sequence. A more complex issue consists in aligning structures. More macroscopic studies are also possible, involving more complex operations on genomes such as permutations. Genotyping studies consider Single Nucleotide Polymorphism data, which correspond to mutations observed at given positions in a sequence with respect to a population. Analyzing this type of data and relating them to phenotypic data leads to new research issues. Once sequences have been compared, phylogenies, that is, trees tracing back the evolution of genes, may be built from a set of induced distances.

4.3. From structural analysis to systems biology

This large domain aims at extracting biological knowledge from Xome studies, where X varies from genes to metabolites. Biological sequences and networks of components in the cell must verify a number of important constraints with respect to stable and accessible conformations, functional mechanisms and dynamics. These constraints result in the conservation during evolution of "patterns" and types of interactions to be deciphered. Many advanced researches consider now the study of life as a system, abstracted in a network of components governed by interaction laws, mostly qualitative or quantitative for reduced systems.

5. Software

5.1. Main softwares

Participants: Olivier Collin [correspondant], Dominique Lavenier, François Coste, Olivier Sallou, Romaric Sabas, Guillaume Rizk, Andres Burgos.

We highlight here 3 softwares of the team which received considerable care this year, in particular to improve their ergonomy and diffusion. In the following sections, all softwares of the team will be described, classified according to their applicative domain.

5.1.1. Biomaj : Data synchronization and processing workflow

BioMAJ (BIOlogie Mise A Jour) is a workflow engine dedicated to data synchronization and processing. The Software automates the update cycle and the supervision of the locally mirrored databank repository. Thanks to the funding of INRIA's ADT, the BioMAJ software has been ergonomically improved and is diffusion enhanced. It is now part of a Linux distribution (Debian-med). The tool is now used on many bioinformatics core facilities in France and Europe. It is used as an infrastructure tool but also as a key component of new resources. For example the AnnotQTL tool relies heavily on BioMAJ. Another example is popgenie, an integrative explorer of the Populus genome in Sweden has been built on top of BioMAJ.

[Web site: http://biomaj.genouest.org]

5.1.2. GASSST: Short reader mapper for large genomic dataset

GASSST is a short read mapper allowing very large genomic dataset to be processed. It takes as input raw data (reads) coming from next generation sequencing machines and map them over full genomes. In 2011, the GASSST software has been tuned to meet industrial requirements and transfered to the GenomeQuest Company. A specific license agreement has been set up between INRIA and GenomeQuest for integrating GASSST into the GenomeQuest NGS tool suite.

web site: http://www.irisa.fr/symbiose/projects/gassst/

5.1.3. Protomata learner: fine characterization of protein families

Protomata-Learner V2.0 is a tool to infer weighted automata for the characterization of (structural or functional) families of proteins from a sample of (unaligned) sequences belonging to the family. Protomata-Learner has been completely rewritten thanks to the ADT "Suite logicielle pour la modélisation de familles protéiques par automates": based on a better formalisation and thanks to the implementation of efficient weighting techniques, this new version is significantly faster and gives better results. Special care has been given to the integration of the different programs to propose an easy-to-use suite.

Protomata-Learner has been tested and improved on real use-case thanks to collaborations established in Lepidolf and Pelican ANR projects. New scanning algorithms (Forward scores) and procedures for choosing automatically the best set of parameters have been developed. New signatures for the studied families of proteins have been established and are used for the predictions of candidates by our partners.

[Web site: http://tools.genouest.org/tools/protomata/]

5.2. Bioinformatics community tools

Participants: Olivier Collin [**contact**], Olivier Sallou, Charles Deltel, François Moreews, Anthony Bretaudeau, Delphine Naquin, Aurélien Roult, Romaric Sabas, Claudia Hériveau.

- **BioMAJ** See first section above.
- **GRISBI** The GRISBI project is aiming to set up a grid infrastructure devoted to the Bioinformatics community. This infrastructure is built upon the resources available on different bioinformatics facilities through gLite middleware. [Web site: http://www.grisbio.fr]
- **Mobylenet** In partnership with other bioinformatics platforms, GenOuest is setting up a distributed network of bioinformatics resources built upon web portals based on the Mobyle platform. [Web site: http://mobylenet.rpbs.univ-paris-diderot.fr:8080/]
- **MetaData platform** Seqcrawler is an indexing platform for biological meta data and sequences, providing a google like web interface. It can scale from single computers to the cloud. [Web site: http://seqcrawler.sourceforge.net/]
- **DrMotifs** DrMotifs is a new software resource aiming at the integration of different software commonly used in pattern search and discovery. This resource will also integrate new software elaborated by the Symbiose team. [Web site: http://www.drmotifs.org] [Blog site: http://drmotifs.genouest.org]

5.3. Parallel softwares

Participants: Dominique Lavenier [**contact**], Charles Deltel, Erwan Drezen, Guillaume Chapuis, Guillaume Rizk.

- **PLAST: intensive bank sequence comparison**. PLAST is a parallel version of BLAST-like software targetting multiple parallel hardware such as FPGA accelerator or GPU boards.
  web site: http://www.irisa.fr/symbiose/projects/plast/
- **SLICEE** (Service Layer for Intensive Computation Execution Environment) is part of the BioWIC project. This software proposes (1) to abstract the calls to the cluster scheduler by handling command submission; (2) to take care of exploiting the data parallelism with data specific methods; (3) to manage data using a cache references mecanism and route data between tasks. [Web site: http://vapor.gforge.inria.fr/]
- **QTL-map** is a GPU parallel version of the QTLMap Software developed in cooperation with INRA
  web site: http://www.inra.fr/qtlmap

5.4. Softwares for Next Generation Sequencing data

Participants: Dominique Lavenier [**contact**], Pierre Peterlongo, Guillaume Rizk, Rayan Chikhi.

- **GASSST: short reads mapper**. See first section above.
- **kisSnp and kisSplice : variant identification without the use of a reference genome**. kisSnp is a tool to find single nucleotide polymorphisms (SNP) by comparing two sets of raw NGS reads. web site: http://alcovna.genouest.org/kissnp/ KisSplice finds alternative splicings but also short insertions, deletions and duplications, SNPs and sequencing errors in one or two RNA-seq sets, without assembly nor mapping on a reference genome. web site: http://alcovna.genouest.org/kissplice/
- **Blastree:** is a tool for computing intensive approximate pattern matching in a string graph. web site: http://alcovna.genouest.org/blastree/
- **Mapsembler: targeted assembly software.** Mapsembler takes as input a set of NGS raw reads and a set of input sequences (starters). It first determines if each starter is read-coherent, e.g. whether the reads confirm the presence of each starter in the original sequence. Then for each read-coherent starter, Mapsembler outputs its sequence neighborhood as a linear sequence or as a graph, depending on the user choice. web site: http://alcovna.genouest.org/mapsembler/

5.5. Genome structure

Participants: Jacques Nicolas [**contact**], Catherine Belleannée, Pierre Peterlongo, Raoul Vorc'h, Anthony Bretaudeau, Olivier Sallou.

- **CRISPI: CRISPR identification**. CRISPI is a user-friendly web interface with many graphical tools and facilities allowing extracting CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats), finding out CRISPR in personal sequences or calculating sequence similarity with spacers. web site: http://crispi.genouest.org
- **Logol** is a language and a tool to define biological patterns to look for in one or more sequences (dna/rna/proteins). Patterns can be complex: the tool allows the use of variables to look for repetitions for example, the use of gaps and morphisms (reverse word complement for example), etc. web site: http://www.genouest.org/spip.php?article758

5.6. Protein sequence and structure

Participants: Rumen Andonov [**contact**], François Coste, Andres Burgos, Pavel Senin.

- **A_purva: Scoring similarities between proteins**. A_purva is a Contact Map Overlap maximization (CMO) solver. Given two protein structures represented by two contact maps, A_purva computes the amino-acid alignment which maximizes the number of common contacts. web site: http://apurva.genouest.org
- **Protomata learner: fine characterization of protein families** See first section above.

5.7. Systems biology

Participants: Anne Siegel [**contact**], Michel Le Borgne, François Moreews, Anthony Bretaudeau.

- **Bioquali: confront knowledge-based regulatory models with data**. Bioquali tests the consistency between an interaction graph and transcriptomic data. It outputs nodes in the network whose variation cannot be globally explained by the other available observations. web site:http://bioquali.genouest.org Cytoscape java web start

6. New Results

6.1. Advanced tools for data management

Participants: Olivier Collin [**contact**], Dominique Lavenier, François Moreews, Olivier Sallou, Anthony Bretaudeau, Jonathan Piat.

- **Annotation and databases**: The AnnotQTL server is a tool designed to gather the functional annotation of genes from several institutional databases for a specific chromosomal region. [14] [Online publication: http://dx.doi.org/10.1093/NAR/GKR361]. SigReannot-mart is a query environment populated with regularly updated annotations for different oligo sets. It stores the results of the SigReannot pipeline that has mainly been used on farm and aquaculture species [17] [Online publication: http://database.oxfordjournals.org/content/2011/bar025]. BioMart Central Portal is a first of its kind, community-driven effort to provide unified access to dozens of biological databases spanning genomics, proteomics, model organisms, cancer data, ontology information and more. [11] [Online publication: http://database.oxfordjournals.org/content/2011/bar041].
- **Bioinformatics Workflow for Intensive Computation** SLICEE proposes to abstract the calls to the cluster scheduler by handling command submission and takes care of exploiting the data parallelism. It enables an easy implementation maintaining and sharing for bioinformatics workflows using intensive computation ressources [32]. OBIWEE is a virtual cluster deployment tool associated with SLICEE. It can be deployed either on private cloud or a public cloud architecture. It helps at facing the increasing demand for bioinformatics intensive treatments, in a context of large dissemination of sequencing technologies usages. [29] http://vapor.gforge.inria.fr/. We also developed a library of bioinformatics softwares implemented on manycore structures such as GPU [27].

6.2. Sequences assembly, alignement and comparison

Participants: Dominique Lavenier [**contact**], Claire Lemaitre, Pierre Peterlongo, Fabrice Legeai, Guillaume Chapuis, Rayan Chikhi, Nicolas Maillet, Delphine Naquin, Raluca Uricaru, Pavlos Antoniou, Thomas Derrien.

- **Hardware accelerator**: Designing FPGA-based accelerators is a difficult and time-consuming task that can be eased by High Level Synthesis Tools. A C-to-hardware methodology has been used to develop an efficient systolic array for the genomic sequence alignment problem. [42], [25][Online publication 1: http://www.eetimes.com/design/programmable-logic/4217568/How-to-accelerate-genomic-sequence-alignment-4X-using-half-an-FPGA?Ecosystem=programmable-logic] [Online publication 2: http://www.springerlink.com/content/37l00567qm18h146/]
- **De novo assembly of NGS data**: A novel framework has been introduced for de novo assembly of next-generation sequencing data. The new paired string graphs and localized assembly models are implemented in the Monument assembler [24]. [Online publication: http://www.springerlink.com/content/f5g305j5k73x3k14/]
- **International competition of de novo genome assembly**: The Symbiose team (IRISA/CNRS/ENS Cachan Brittany) participated to this competition. [3]. [Online publication: http://genome.cshlp.org/content/early/2011/09/16/gr.126599.111.abstract]
- **Indexation of NGS data**: A novel data structure is described for indexing NGS data. The structure is coupled with filtering algorithms that enable memory-efficient and parallel indexing. [23]
- **Breakpoints in genomes**: We analysed the correlation between 3D chromatin interaction data and breakpoint regions resulting from evolutionary rearrangements in the human genome. We found that two loci distant in the human genome but adjacent in the mouse genome are significantly more often observed in close proximity in the human nucleus than expected. [21]. [Online publication: http://www.biomedcentral.com/1471-2164/12/303]

- **Repeat detection**: A tool has been presented for detecting long similar fragments that occur two or more times in a set of biological sequences. This is achieved by using a filter as a preprocessing step, and by using the information that the filter has gathered also in the successive inference phase. [26]. [Online publication: http://www.stringology.org/event/2011/p08.html]
- **Targeted assembly of NGS data**: Mapsembler is an iterative targeted assembler which processes large datasets of reads on commodity hardware. Mapsembler checks for the presence of given regions of interest in the reads and reconstructs their neighborhood, either as a plain sequence (consensus) or as a graph (full sequence structure). [39]
- **Transcriptome assembly and annotation**: We established and analyzed two catalogues of transcripts by assembling EST sequences, and performed their functional annotations using the gene ontology for the following 2 species : spodoptera littoralis [15] and cabomba [20].
- **Substitution matrices**: A general and simple methodology has been proposed to build new matrices fitted to specific compositional bias of proteins. It was then applied to the large scale comparison of Mollicute AT-rich genomes [16]. [Online publication: http://www.biomedcentral.com/1471-2105/12/457/]

6.3. Genome Structure

Participants: Jacques Nicolas [**contact**], Dominique Lavenier, François Coste, Catherine Belleannée, Olivier Sallou, Fabrice Legeai, Guillaume Rizk, Guillaume Chapuis, Matthias Gallé, Anthony Bretaudeau.

- **GPU accelerated RNA folding algorithm** The main kernel of the widely used RNA folding package Unafold has been accelerated using GPU boards by reordering computations to enable tiled computations and good data reuse [37], [2].
- **GPU accelerated QTL algorithm** Our GPU/multicore implementation performs up to 20 times faster than the previous multicore implementation and allows extensive QTL analysis to be conducted in a reasonable time, while maintaining the same level of precision [35].
- **Hierarchical structure of genomes.** In [8], we proposed to split the classical smallest grammar problem into two tasks: (1) choosing the constituents of the grammar and (2) finding the *smallest grammar parsing* given these constituents. This defines properly the search space for this problem and, as we have shown how to solve in polynomial time the second task, this opens doors for new algorithms finding smaller grammars as shown on a generic compression benchmark (up to 10%). In [7], we have worked on the scalability to propose a new algorithm able to handle whole genomes: on this kind of sequences, the size reduction is still about 10% for a comparable execution time with respect to state-of-the art algorithms.
- **Data compression** By using grammars with rigid patterns as words, we were able to achieve a compression rate up to 25% better compared to the previous best DNA grammar-based coder, and just below state-of-the-art dedicated DNA compressors [1].
- **CRISPR Modeling and identification:** CRISPR (Clustered regularly interspaced short palindromic repeats) are small repeats present in a number of bacterial and archaeal species. We proposed the most complete database on these elements (http://crispi.genouest.org), elaborating for the first time a complete study of the palindromic nature of these repeats. The analysis has made an extensive use of our Logol Parser to decipher stem-loop structures [40].
- **Aphid genetics** We participated in a genetic study aiming at comparing the rates of evolution of genes enclosed in aphid sexual chromosome (X) to autosomal genes. In order to do so, we provided particular microsatellites for the selection of genomic sequences, as well as tools for studying their genomic environment. [13] [Online publication: http://mbe.oxfordjournals.org/content/early/2011/10/12/molbev.msr252]

6.4. Protein Sequences and Structures

Participants: Rumen Andonov [**contact**], Antonio Mucherino, François Coste, Jacques Nicolas, Andres Burgos, Gaëlle Garet, Pavel Senin, Mathilde Le Boudic-Jamin.

- **Branch & Prune Algorithm**: We proposed an extension of the Branch & Prune (BP) algorithm for the Discretizable Molecular Distance Geometry Problem (DMDGP) which is able to exploit all symmetries of the research domain of the corresponding combinatorial optimization problem [30]
- **Modeling protein sequences with long distance correlations** To initiate this new line of research, we have set up a framework to learn context-free grammars on protein sequences based on the identification of conservation blocks and substitutability of non-terminals. A first implementation of the learning algorithms showed the interest of this approach [38]
- **Maximum Contact Map Overlap (CMO)** is a popular measure for quantifying the similarity between protein structures. A new integer programming model was presented for CMO and an exact branch-and-bound algorithm was designed with bounds obtained by a novel Lagrangian relaxation. The efficiency of the approach was demonstrated on known benchmarks on which sets our approach significantly outperforms the best existing exact algorithms [4].
- **Alignment of protein structures** First successes were obtained on provably optimal pairwise alignment of protein inter-residue distance matrices, using the popular Dali scoring function. We proposed the first mathematical model for computing optimal structural alignments based on dense inter-residue distance matrices and present algorithm engineering techniques to handle the huge integer linear programs [22]. In a second paper, a strategy was proposed for sparsifying distance matrices in which only the distances needed for uniquely reconstructing the conformations of the proteins are kept. [31].
- **Protein Family Identification** Identification of protein families is a computational biology challenge that needs efficient and reliable methods. First, we used the comparison tool A_purva, which is based on Contact Map Overlap (CMO), to classify protein structure coming from the CATH database. The obtained results showed that A_purva was able to correctly classify 92% of the structures, and that introducing the notion of dominance drastically reduced the computational time needed for classifying the protein structures [33]. Then, we introduced this concept of dominance in a novel combined approach based on Distance Alignment Search Tool (DAST), which contains an exact algorithm with bounds. Our experiments showed that this method successfully finds the most similar proteins in a set without solving all instances [28].
- **Local Protein Threading, sequence-structure alignment**: A novel approach to PTP has been investigated. It aligns a part of a protein structure onto a protein sequence in order to detect local similarities [9]. [Online publication: http://www.sciencedirect.com/science/article/B6TYW-50G78H4-1/2/947312da7a7bbf175cab7b3288ba4f03]

6.5. Systems Biology

Participants: Anne Siegel [**contact**], Jérémie Bourdon, Michel Le Borgne, Nathalie Theret, Geoffroy Andrieux, Oumarou Abdou-Arbi, Sylvain Prigent, Pierre Blavy, Andres Aravena, Santiago Videla, Valentin Wucher, Brivael Trelhu.

- **Average-case analysis for quantitative data integration** We proposed a probabilistic modeling framework that integrates heterogeneous data. Average case analysis methods were used in combination with Markov chains to link qualitative information about transcriptional regulations to quantitative information about protein concentrations. The approach was illustrated by modeling the carbon starvation response in Escherichia coli. It accurately predicted the quantitative time-series evolution of several protein concentrations using only knowledge of discrete gene interactions and a small number of quantitative observations on a single protein concentration [6]. [Online publication: http://dx.plos.org/10.1371/journal.pcbi.1002157]

- **Combining genetic and metabolic regulations**: We mixed Gale-Nikaido reduction steps and differential inequalities to understand how genetic regulation modifies the behavior of a very abstracted model of lipid metabolism [18] [Online publication: http://www.springerlink.com/content/n437048670560782/]
- **Extract relevant information with respect to a cancer phenotype**: We designed dedicated logical rules to model the static response of biomolecular interactions implied in the cancer network. This allowed us to trace back genes implied in the cancer phenotype [12]. [Online publication: http://www.computer.org/portal/web/csdl/doi/10.1109/TCBB.2010.71]
- **Integrative biology for brown algal** We proposed a protocol focusing on integrating heterogeneous knowledge gained on brown algal metabolism. The resulting abstraction of the system helps understanding how brown algae cope with changes in abiotic parameters within their unique habitat [19].
- **Search for key regulators** A method was proposed to model the effects of all transcriptional and metabolic regulations contained in transpath in a single influence network. The network was analyzed to find a set of candidates that explain the variations of a set of targets [34].
- **Identification of co-regulation patterns**. We introduced a new approach based on the compilation of Simple Shared Motifs (SSM), groups of sequences defined by their length and similarity and present in conserved sequences of gene promoters. We proved that Simple Shared Motifs analysis provides a clearer definition of expression networks [10]. [Online publication: http://www.biomedcentral.com/1471-2105/12/365]
- **Probabilistic models for systems biology** We reviewed, in a book chapter, some classical concepts concerning probabilistic models and their applications in systems biology. Probabilistic boolean networks were presented in deep with a focus on the effect of synchronization of genes and on stochastic simulation of such networks [36].

7. Contracts and Grants with Industry

7.1. Contracts with Industry

GASSST-GQ is an industrial contract with the GenomeQuest Company for tuning the GASSST software with industrial requirements. It is coordinated by D. Lavenier (EPI Symbiose) and JJ. Codani (GenomeQuest). In 2011, the GASSST software has been transfered to GenomeQuest. [Genome Quest web announcement] [Inria web announcement]

7.2. Grants with Industry

KoriPLAST is a project with the Korilog company aiming to transfer the PLAST software developed in the Symbiose Team. It is funded by the Britany region. [Inria web announcement]

The Peapol project is funded by Sofiproteol company whose mission is to develop the French vegetable oil and protein industry, open up new markets, and ensure an equal distribution of value among its members. The Peapol project counts two collaborators, Biogemma, and INRA, the latter working in collaboration with the Symbiose team, in charge of algorithmic research in the context of the project. This collaboration enabled to hire in the Symbiose team Raluca Uricaru for 18 months on an INRA post doctoral position.

8. Partnerships and Cooperations

8.1. Regional Initiatives

8.1.1. Genopole initiatives

Participants: Olivier Collin, Delphine Naquin, Aurélien Roult.

We benefit from the strong implication of the GenOuest Ressource center in the regional Genopole to have long-term research and development relationships with most of laboratories in Brittany involved in molecular biology.

As a technological platform belonging to Biogenouest, the Life Science network of the West of France, the bioinformatics platform is funded by the Brittany Region. This funding allowed the creation of two short term contract positions, shared with other technological platforms:

- A sequencing facility: creation of a bioinformatics environment for the management and analysis of Next Generation Sequencing data,
- a proteomic facility: system and network administration, managing a small cluster and providing expertise for an upgrade of the computational and network infrastructure.

This alliance with other facilities proved to be a very good way to reinforce bonds and cooperation, giving birth to new research subjects both within the Symbiose team and biological team.

8.1.2. Partnership with INRA

Participants: Fabrice Legeai, François Moreews, Pierre Blavy, Raluca Uricaru, Thomas Derrien, Valentin Wucher.

We have a strong and long term collaboration with biologists of INRA in Rennes : Bio3P, APBV and SENAH units. This partnership concerns both service and research activities and is acted by the hosting of two ingineers (F. Legeai, F. Moreews) and by the co-supervision of three post-doctorants and one PhD student. In particular, the collaboration with the APBV team, including the co-supervision of a post-doc, are built upon an INRA project PEAPOL including an industrial partner, Biogemma.

8.2. National Initiatives

8.2.1. ANR contracts

8.2.1.1. BIOWIC

Participants: Dominique Lavenier, Olivier Collin, Rumen Andonov, François Moreews, Jonathan Piat, Guillaume Rizk.

The BioWIC project aims to speed up both the design and the execution of bioinformatics workflows. It is funded by ANR call ARPEGE and coordinated by D. Lavenier from Jan. 2009 to June 2012. http://biowic.inria.fr/

8.2.1.2. LEPIDOLF

Participants: François Coste, Fabrice Legeai, Jacques Nicolas, Andres Burgos, Pavel Senin.

The LEPIDOLF project aims at better understanding olfactory mechanisms in insects. The goal is to establish the antennal transcriptome of the cotton leafworm Spodoptera littoralis, a noctuid representative of crop pest insects. It is funded by ANR call Blanc and coordinated by E. Jacquin-Joly from UMR PISC (INRA) from 2009 to 2012.

8.2.1.3. MAPPI

Participants: Dominique Lavenier, Pierre Peterlongo, Guillaume Chapuis, Rayan Chikhi, Nicolas Maillet.

The MAPPI project aims to develop new algorithms and Bioinformatics methods for processing high trougthput genomic data. It is funded by ANR call COSINUS and coordinated by M. Raffinot (LIAFA, Paris VII) from Oct 2010 to Dec. 2013. http://mappi.arthy.org/

8.2.1.4. PELICAN

Participants: Olivier Collin, François Coste, Anthony Bretaudeau, Andres Burgos.

The PELICAN project addresses competition for light in the ocean: An integrative genomic approach of the ecology, diversity and evolution of cyanobacterial pigment types in the marine environment. It is coordinated by F. Partensky (CRNS Roscoff) from 2010 to 2013. http://www.sb-roscoff.fr/anr-pelican/

8.2.1.5. ECS

Participant: Olivier Collin.

The ECS project explores cooperation in plant symbioses. It provides new insights on how plant-microbe interactions shape the ecological processes and evolutionary trajectories of natural and agricultural ecosystems. This project seeks also to identify new symbiotic partners for plants. It is funded by ANR (Systerra) and coordinated by P. Vandenkoornhuyse from UMR 6553 (CAREN) from 2010 to 2013. http://ecs-project.univ-rennes1.fr/news.php

8.2.1.6. BIOTEMPO

Participants: Anne Siegel, Jérémie Bourdon, François Coste, Jacques Nicolas, Michel Le Borgne, Geoffroy Andrieux, Sylvain Prigent, Santiago Videla, Andres Aravena.

The BioTempo project aims at developing some original methods for studying biological systems. The goal is to introduce partial quantitative information either on time or on component observations to gain in the analysis and interpretation of biological data. Three biological applications are considered regulation systems used by biomining bacteria, TGF-beta signaling and initiation of sea-urchin translation. It is funded by ANR Blanc (SIMI2) and coordinated by A. Siegel from 2011 to 2014. http://biotempo.genouest.org/wiki.php/Accueil

8.2.2. Programs from research institutions

Participants: Anne Siegel, François Coste, Olivier Collin, Charles Deltel, Dominique Lavenier, Michel Le Borgne, Claire Lemaitre, Pierre Peterlongo, Jérémie Bourdon, Pavlos Antoniou, Andres Burgos, Guillaume Chapuis.

- **Alcovna** The Alcovna project aims to explore possibilities of extracting information among possibly huge sets of reads without reference genome and avoiding to assemble the data. It is funded by INRIA ARC call and coordinated by P. Peterlongo from oct. 2009 to sept. 2011. http://alcovna.genouest.org
- **BioManyCores** The BioManyCores project aims to develop a library of bioinformatics softwares implemented on manycore structures such as GPU. It is funded by INRIA ADT call and supervised by J.S. Varré in Sequoia Team in Lille. http://www.biomanycores.org/
- **ParaQtlMap** The ParaQtlMap project is a join initiative from EPI Symbiose and Genetique Animale. to design high performance software for detecting quantitative trait locus. It is funded by INRIA/INRA call and coordinated by D. Lavenier (EPI Symbiose) and P. Leroy (GA INRA) from oct. 2010 to sept. 2012. https://qgp.jouy.inra.fr/index.php?option=com_content&task=view&id=17&Itemid=28
- **Protomata** The protomata project aims at developping a software for modeling proteins families with automatas. It is funded by INRIA ADT call (2010-2011) and supervised by F. Coste.
- **QuantOursin** The QuantOursin project aims at developping modeling tools based on probabilistic framework and average analysis, and apply then to the initiation of urchin translation. It is funded by a PEPS program at CNRS and coordinated by A. Siegel from april. 2010 to december. 2012. http://quantoursin.genouest.org/wiki.php/Accueil

8.2.3. Transfert and service ressources - GenOuest ressource center

Participants: Olivier Collin, Olivier Sallou, Charles Deltel, Anthony Bretaudeau, Delphine Naquin, Aurélien Roult, Romaric Sabas, Claudia Hériveau.

- **GRISBI** The project intends at developping a production grid dedicated to bioinformatics, by gathering computational ressources of six french ressource centers. It is funded by IBISA and coordinated by C. Blanchet (IPCP Lyon) from 2009 to 2011. http://www.grisbio.fr
- **DrMotifs** is a project dedicated to develop tools for pattern discovery and research. The resource integrates the tools in a workflow plugged on different databases in order to provide a user friendly tool geared toward motif discovery. It is funded by IBISA and coordinated by O. Collin (Symbiose) from 2010 to 2011. http://drmotifs.genouest.org

- **BioMaj** The project aims at developing a workflow engine dedicated to data synchronization and processing.The Software automates the update cycle and the supervision of the locally mirrored databank repository. It is funded by Inria ADT program from 2009 to 2011 and coordinated by O. Collin. http://biomaj.genouest.org.
- **Inria Biosciences Resources** This new project (recruitement of an engineer in november) is focused on the different ways to improve the visibility of Inria's bioinformatics team software production. To achieve this goal, a new web resource will be built. This resource will allow end-users to test and evaluate the new bioinformatics tools created by Inria. It is funded by INRIA ADT program from 2011 to 2013, it involves 8 research teams and is coordinated by Symbiose (Jacques Nicolas and Olivier Collin).

8.3. European Initiatives

8.4. International Initiatives

8.4.1. INRIA Associate Teams

8.4.1.1. INTEGRATIVEBIOCHILE

Title: Bioinformatics and mathematical methods for heterogeneous omics data

INRIA principal investigator: Anne Siegel

International Partner:

Institution: University of Chile (Chile)

Laboratory: University of Chile, CMM

Duration: 2011 - 2013

See also: http://www.irisa.fr/symbiose/people/asiegel/EA/

IntegrativeBioChile is an Associate Team between INRIA project-team "Symbiose" and the "Laboratory of Bioinformatics and Mathematics of the Genome" hosted at CMM at University of Chile. The Associated team is funded from 2011 to 2013. The project aims at developping bioinformatics and mathematical methods for heterogeneous omics data. Within this program, we funded long-stay visitings in France to initiate long-term research lines, in complement to short visit funded by and inria-conycit program.

- **Reconstruction of regulatory networks.** This research line was developed within the visits of A. Aravena and A. Maass in March 2011 (funded by conycit-inria and by a mobility grant from UEB). It was pushed further during the visit of J. Bourdon in October 2011 and A. Siegel in November 2011 in Chile.
- **Reconstruction and study of metabolic network with reduction methods.** This research line was initiated during the one-month visit of Marko Budinich (engineer) in Rennes in April 2011. It was pushed further with visits from french researchers in Santiago funded by Inria-Conycit program (D. Eveillard and S. Prigent).
- **Detection of structural variations in genomes** This research line was studied during the one-month visit of Alex Di Genova (engineer) in Rennes in May 2011.

8.4.2. INRIA International Partners

- **CWI, The Netherlands** : In 2011 we have been collaborating very actively with the Algorithmic computational biology from CWI Life Sciences, The Netherlands in the domain of protein structure comparison. Inken Wohler, a PhD student from CWI visited Symbiose for 6 months partially supported by Inria internship grant. Two papers have been published in the framework of this cooperation [31], [22].

8.4.3. Visits of International Scientists

8.4.3.1. Internship

- Andres Aravena, from CMM (Chile) received a 3 months Mobility Grant from "University Européenne de Bretagne" to visit the Symbiose team between april and july.
- Inken Wohlers, from CWI (Amsterdam The Netherlands) received a 6 months Inria internship and visited Symbiose from December 2010 to May 2011.

8.4.4. Participation In International Programs

8.4.4.1. Chile. Inria-Conycit 2011-12

Partner: University of Chile, *Laboratory of Bioinformatics and Mathematics of the Genome*, Chile.

Title: IntegrativeBiomining

Financial support: Conicyt-Inria program 2011-12

INRIA principal investigator: Anne Siegel

The project wishes at developing methods combining combinatorial and constraint-based approaches with probabilistic/optimization approaches to integrate and explore heterogeneous, multi-scale and large-scale data produced in molecular biology.

Within the project, the following visits were funded (A) 2 monthes visit in Rennes (A. Aravena, phD student, February-March 2011) (B) 3 weeks visit in Rennes (A. Maass, professor, March 2011) (C) 15 days visit in Santiago de Chile (A. Siegel, assistant professor, july 2011). (D) 15 days visit in Santiago de Chile (F. Coste, junior researcher, october 2011) (E) 1 month visit in Santiago de Chile (S. Prigent, phD student, November 2011)

8.4.4.2. Argentina - MinCYT-Inria 2011-12

Partner: Universidad Nacional de Córdoba, *Grupo de Procesamiento de Lenguaje Natural (PLN)*, Argentina.

Title: Modélisation linguistique de séquences génomiques par apprentissage de grammaires

Financial support: MinCYT-Inria program 2011-12

The project aims at developing new grammatical inference methods to learn automatically linguistic models of genomic sequences.

Within the project, the following visits were funded (A) 15 days visit in Cordobà (F. Coste, junior researcher, June 2011) (B) 1 month visit in Rennes (R. Carrascosa, ph-D student, August 2011) (C) 15 days visit in Rennes (G. Infante-Lopez, professor, November 2011).

8.4.4.3. Germany. Egide Procope Program 2011-12

Partner: Postdam university, Institut fu¨r Informatik Wissensverarbeitung und Informationssysteme, Germany

Title: Querying Biological Systems with Answer Set Programming.

Financial support: Egide Procope Program 2011-12

The project aims at developing new methods for constructing and querying biological networks with a new constraint-based programming (answer set programming) mastered in Postdam university.

Within the project, the following visits were funded (A) 1 week visit in Rennes (S. Thiele, phD student, May 2011) (B) 15 days visit in Postdam and Heidelberg (P. Blavy, phD student, June 2011) (C) 1 week visit in Rennes (T. Schaub, professor, October 2011) (D) 1 week visit in Rennes (M. Gebser, Post-doc, October 2011) (E) 1 week visit in Postdam (J. Nicolas, senior researcher, December 2011) (F) 1 week visit in Postdam (S. Videla, phD student, December 2011) (G) 1 week visit in Postdam (V. Wurcher, phD student, December 2011)

9. Dissemination

9.1. Animation of the scientific community

9.1.1. Administrative functions: scientific committees, journal boards, jury

- Scientific Advisory Board of ITMO Genetics Genomics and Bioinformatics [J. Nicolas].
- Scientific Advisory Board of GDR BIM " Molecular Bioinformatics"[J. Nicolas].
- Member of the Evaluation Committee of Inria [A. Siegel]
- Member of the IRISA laboratory council [F. Coste]
- Member of the administrativ council of ISTIC [R. Andonov]
- ANR committees [D. Lavenier / Modèles numériques]
- Member of ReNaBi steering committee and coordinator for ReNaBi-GO (Grand Ouest). This regional centre includes the platforms of Nantes, Rennes and Roscoff [O. Collin]
- Scientific Advisory Board of Biogenouest [J. Bourdon, O. Collin, J. Nicolas].
- Steering committee of the International Inference community (ICGI) [F. Coste]
- Recruitment committees: engineer, junior research, assistant professor, scientific study officer [O. Collin, A. Siegel, P. Peterlongo, D. Lavenier, R. Andonov]
- Member of the "new paradigm computation" group (OMNT, Observatoire des micro et nanotechnologies) [D. Lavenier]
- Member of the Editorial Board of The Scientific World JOURNAL, bioinformatics domain [D. Lavenier]
- Member of the International Expert Committee (MEI, Research Ministry) [D. Lavenier]
- Member of SCAS (Service Commun d'Action Sociale) of Univ. Rennes 1 [C. Belleannée]

9.1.2. Conference program committees

- International Symposium on String Processing and Information Retrieval (SPIRE 2011)
- International Conference on Field Programmable Logic and Applications (FPL)
- International Conference on Engineering of Reconfigurable Systems and Algorithms (ERSA)
- International Conference on Contemporary Computing (ICCC)
- Southern Programmable Logic Conference (SPL)
- ACM International Conference on Computing Frontiers (UCHPC Workshop) (CF)
- International Conference on ReConFigurable Computing and FPGAs (ReConFig)
- ParCo: Parallel Computing with FPGAs (ParaFPGA)
- ICCS: Workshop on Emerging Parallel Architectures (WEPA)
- International Conference on DNA Computing and Molecular Programming (DNA 17)
- Conference d'apprentissage (CAP'11).

9.1.3. Meeting organization and scientific animation

- **Seminar** A weekly seminar of bioinformatics is organized within the laboratory. Attendees are member of the symbiose team, biologists from Brittany and computer scientists from the laboratory. A thematic day meeting on modeling issues was also organized by the team [web site: http://www.irisa.fr/symbiose/seminaires/].
- **GenOuest annual meeting** GenOuest annual meeting The 9th annual meeting of GenOuest computing center took place october 2011, the 18th. This session was dedicated to ontologies in life science. There were 65 attendees [web site: http://www.genouest.org/spip.php?article850].
- **SeqBI** A workshop entitled "Algorithmique, combinatoire du texte et applications en bioinformatique" took place at Inria Rennes the 10th and 11th of January 2011. It was funded by GDR BIM. 50 people coming from all France attended to this workshop.[web site: http://www.irisa.fr/symbiose/people/ppeterlongo/seqbi/].

9.1.4. Jury of PhD Theses

- *Referee of Habilitation thesis jury*. F. Clermidy, Université de Grenoble [D. Lavenier]
- *Member of Ph-D thesis jury*. T. Riaz, Université Joseph Fourrier Grenoble [P. Peterlongo]. N. Philippe, Université de Montpellier [D. Lavenier]. S. Laroum, Université d'Angers [D. Lavenier], A. Ben Hassena, ENSSAT [F. Coste].
- *Referee of Ph-D thesis*. C. Rezvoy, ENS Lyon [D. Lavenier]. C. Teodorov, Université Rennes 1 [D. Lavenier], N. Terrapon, Université de Montpellier [J. Nicolas], F. Cliquet, Université de Nantes [J. Nicolas], O. Gaci, Université du Havre [R. Andonov].

9.2. Teaching

Licence : Imperative Programming and scientific computation, 36h, L2, Univ. Rennes 1, France.

Licence : Programming, 20h, L2 Univ. Rennes 1, France.

Licence : Algorithmic methods, 96h, L3, Istic, Rennes France

Licence : Graph algorithms, 20h, L3, Istic, Rennes France

Licence : Programming, 78h, L3, Univ. Rennes 1, France.

Licence : Web programming, 26h, L3, Univ. Rennes 1, France.

Licence : Network provisioning, 20h, L3, Univ. Rennes 1, France.

Licence : Operations research, 30h, L3, Univ. Rennes 1, France.

Licence : Architectures and Systems, 64h, L3, ENS Cachan, Rennes, France

Licence : Database, 21h, L3, Istic Univ. Rennes 1 France

Licence : Architectures, 50h, L3, Istic Univ. Rennes 1 France

Master : Compilation, 32h, M1, Univ. Rennes 1, France.

Master : Operations research, 100h, M1, Univ. Rennes 1, France.

Master : Logic Programming, 32h, M1, Istic Univ. Rennes 1 France

Master : Dynamical systems for biological networks, 16h, M2, Univ. Rennes 1, France

Master : Programmation Objet, 15h, M2, Univ. Rennes 1, France

Master : Sequence algorithms, 29h, M2, Univ. Rennes 1, France

Master : Bioinformatics, 12h, M2, ESEO Angers, France

Master : Symbolic sequential data, 10h, Univ. Rennes 1, France

Master : Bioinformatics, 3h, M2, Istic Univ. Rennes 1 France

Master : Numerical and combinatorial optimization, 12h, M2, Univ. Rennes 1, France.

Master : Modelling of protein structures, 15h, M2, Univ. Rennes 1, France.

PhD defense : Matthias Gallé, *Searching for Compact Hierarchical Structures in DNA by means of the Smallest Grammar Problem*[1], supervised by J. Nicolas and G. Infante-Lopez (Argentine), defended on Februrary 15th 2011 [online manuscript: http://tel.archives-ouvertes.fr/tel-00595494_v1/]

PhD defense : Guillaume Rizk, *Parallélisation sur matériel graphique : contributions au repliement d'ARN et à l'alignement de séquences*[2], supervised by D. lavenier, defended on April 12th 2011 [online manuscript: http://tel.archives-ouvertes.fr/tel-00634901/]

PhD in progress : Oumarou Abdou-Arbi *Analyse Automatisée et générique des réseaux métaboliques en nutrition*, started in October 2010, supervised by A. Siegel and T. Tabsoba (Burkina-Faso).

PhD in progress : Geoffroy Andrieux, *Discrete approach modeling of biological signaling pathway*, started in October 2010, supervised by N. Théret (Inserm) and M. Le Borgne

PhD in progress : Andres Aravena, *Introduire des approches combinatoires dans des modèles probabilistes pour la découverte d'évènements de régulation d'un système biologique à partir de données hétérogènes*, started in July 2011, supervised by A. Maass (CMM, University of Chile) and A. Siegel.

PhD in progress : Guillaume Chapuis, *Bioinformatique parallèle*, started in October 2010, supervised by D. Lavenier

PhD in progress: Rayan Chikhi, *Computational theory for de novo assembly of short sequencing reads*, started in October 2008, supervised by D. Lavenier

PhD in progress : Gaëlle Garet, *Discovery of enzymatic functions in the framework of formal languages*, started in October 2011, supervised by J. Nicolas and F. Coste.

PhD in progress : Mathilde Le Boudic-Jamin, *Through Flexible Protein-Protein Docking*, started in October 2011, supervised by R. Andonov

PhD in progress : Nicolas Maillet, *Algorithme pour l'assemblage de données NGS de métagénomique*, started in November 2010, supervised by D. Lavenier and P. Peterlongo

PhD in progress : Sylvain Prigent, *Modélisation par contraintes pour le contrôle génomique et physiologique de l'adaptation des algues brunes à la salinité de l'eau*, started in October 2011, supervised by A. Siegel and T. Tonon (UMR 7150, station biologique de Roscoff)

PhD in progress : Santiago Videla, *Applying logic programming to the construction of robust predictive and multi- -scale models of bioleaching bacteria*, started in November 2011, supervised by A. Siegel

PhD in progress : Valentin Wucher, *Modélisation d'un réseau de régulation d'ARN pour prédire des fonctions de gènes impliqués dans le mode de reproduction du puceron du pois*, started in November 2011, supervised by J. Nicolas and D. Tagu (INRA)

10. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] M. GALLÉ. *Searching for Compact Hierarchical Structures in DNA by means of the Smallest Grammar Problem*, Université Rennes 1, February 2011, http://hal.inria.fr/tel-00595494/en.

[2] G. RIZK. *Parallélisation sur matériel graphique : contributions au repliement d'ARN et à l'alignement de séquences*, Université Rennes 1, January 2011, http://hal.inria.fr/tel-00634901/en.

Articles in International Peer-Reviewed Journal

[3] D. A. EARL, K. BRADNAM, ..., R. CHIKHI, D. LAVENIER, G. CHAPUIS, D. NAQUIN, N. MAILLET, ..., I. KORF, B. PATEN. *Assemblathon 1: A competitive assessment of de novo short read assembly methods*, in "Genome Research", September 2011, International competition of de novo genome assembly. The Symbiose team (IRISA/CNRS/ENS Cachan Brittany) participated to this competition. [*DOI :* 10.1101/GR.126599.111], http://hal.inria.fr/inria-00637571/en.

[4] R. ANDONOV, N. MALOD-DOGNIN, N. YANEV. *Maximum Contact Map Overlap Revisited*, in "Journal of Computational Biology", January 2011, vol. 18, n° 1, p. 1-15 [*DOI :* 10.1089/CMB.2009.0196], http://hal.inria.fr/inria-00536624/en.

[5] V. BERTHÉ, A. SIEGEL, W. STEINER, P. SURER, J. THUSWALDNER. *Fractal tiles associated with shift radix systems*, in "Advances in Mathematics", January 2011, vol. 226, n° 1, p. 139-175 [*DOI :* 10.1016/J.AIM.2010.06.010], http://hal.inria.fr/hal-00407999/en.

[6] J. BOURDON, D. EVEILLARD, A. SIEGEL. *Integrating quantitative knowledge into a qualitative gene regulatory network.*, in "PLoS Computational Biology", September 2011, vol. 7, n° 9, e1002157 [*DOI :* 10.1371/JOURNAL.PCBI.1002157], http://hal.inria.fr/hal-00626708/en.

[7] R. CARRASCOSA, F. COSTE, M. GALLÉ, G. INFANTE-LOPEZ. *Searching for Smallest Grammars on Large Sequences and Application to DNA*, in "Journal of Discrete Algorithms", 2011 [*DOI :* 10.1016/J.JDA.2011.04.006], http://hal.inria.fr/inria-00536633/en.

[8] R. CARRASCOSA, F. COSTE, M. GALLÉ, G. INFANTE-LOPEZ. *The Smallest Grammar Problem as Constituents Choice and Minimal Grammar Parsing*, in "Algorithms", October 2011, n° 4, p. 262-284 [*DOI :* 10.3390/A4040262], http://hal.inria.fr/inria-00638445/en.

[9] G. COLLET, R. ANDONOV, J.-F. GIBRAT, N. YANEV. *Local protein threading by Mixed Integer Programming*, in "Discrete Applied Mathematics", 2011, vol. 159, n° 16, p. 1707-1716 [*DOI :* 10.1016/J.DAM.2010.05.024], http://hal.inria.fr/inria-00536537/en.

[10] J. GRUEL, M. LE BORGNE, N. LEMEUR, N. THÉRET. *Simple Shared Motifs (SSM) in conserved region of promoters: a new approach to identify co-regulation patterns*, in "BMC Bioinformatics", September 2011, 12:365, http://hal.inria.fr/hal-00641785/en.

[11] J. M. GUBERMAN, J. AI, ..., F. MOREEWS, ..., J. ZHANG, A. KASPRZYK. *BioMart Central Portal: an open database network for the biological community*, in "Database: The Journal of Biological Databases and Curation", September 2011 [*DOI :* 10.1093/DATABASE/BAR041], http://hal.inria.fr/inria-00638749/en.

[12] C. GUZIOLOWSKI, S. BLACHON, T. BAUMURATOVA, G. STOLL, O. RADULESCU, A. SIEGEL. *Designing Logical Rules to Model the Response of Biomolecular Networks with Complex Interactions: An Application to Cancer Modeling.*, in "IEEE/ACM Trans Comput Biol Bioinform", 2011, vol. 8, n° 5, p. 1223-1234 [*DOI :* 10.1109/TCBB.2010.71], http://hal.inria.fr/inria-00538134/en.

[13] J. JAQUIÉRY, S. STOECKEL, C. RISPE, L. MIEUZET, F. LEGEAI, J.-C. SIMON. *Accelerated evolution of sex chromosomes in aphids, an X0 system.*, in "Molecular Biology and Evolution", October 2011, to appear [*DOI :* 10.1093/MOLBEV/MSR252], http://hal.inria.fr/hal-00639983/en.

[14] F. LECERC, A. BRETAUDEAU, O. SALLOU, C. DÉSERT, Y. BLUM, S. LAGARRIGUE, O. DEMEURE. *AnnotQTL: a new tool to gather functional and comparative information on a genomic region*, in "Nucleic Acids Research", May 2011, to appear [*DOI :* 10.1093/NAR/GKR361], http://hal.inria.fr/hal-00597398/en.

[15] F. LEGEAI, S. MALPEL, N. MONTAGNÉ, C. MONSEMPES, F. COUSSERANS, C. MERLIN, M.-C. FRANÇOIS, M. MAÏBÈCHE-COISNÉ, F. GAVORY, J. POULAIN, E. JACQUIN-JOLY. *An Expressed Sequence Tag collection from the male antennae of the Noctuid moth Spodoptera littoralis: a resource for olfactory and pheromone detection research.*, in "BMC Genomics", 2011, vol. 12, 86 [*DOI :* 10.1186/1471-2164-12-86], http://hal.inria.fr/hal-00639985/en.

[16] C. LEMAITRE, A. BARRÉ, C. CITTI, F. TARDY, F. THIAUCOURT, P. SIRAND-PUGNET, P. THEBAULT. *A novel substitution matrix fitted to the compositional bias in Mollicutes improves the prediction of homologous relationships*, in "BMC Bioinformatics", 2011, to appear, http://hal.inria.fr/hal-00643361/en.

[17] F. MOREEWS, G. RAUFFET, P. DEHAIS, C. KLOPP. *SigReannot-mart: a query environment for expression microarray probe re-annotations.*, in "Database (Oxford)", 2011, vol. 2011, bar025 [*DOI :* 10.1093/DATABASE/BAR025], http://hal.inria.fr/inria-00638684/en.

[18] O. RADULESCU, A. SIEGEL, E. PECOU, C. CHATELAIN, S. LAGARRIGUE. *Genetically regulated metabolic networks: Gale-Nikaido modules and differential inequalities*, in "Transactions on Computational Systems Biology", 2011, n^{o} Lecture Notes in Computer Science 6575, p. 110-130 [*DOI :* 10.1007/978-3-642-19748-2_6], http://hal.inria.fr/inria-00538136/en.

[19] T. TONON, D. EVEILLARD, S. PRIGENT, J. BOURDON, P. POTIN, C. BOYEN, A. SIEGEL. *Toward systems biology in brown algae to explore acclimation and adaptation to the shore environment*, in "OMICS", 2011, http://hal.inria.fr/inria-00636965/en.

[20] A. C. M. VIALETTE-GUIRAUD, M. ALAUX, F. LEGEAI, C. FINET, P. CHAMBRIER, S. C. BROWN, A. CHAUVET, C. MAGDALENA, P. J. RUDALL, C. P. SCUTT. *Cabomba as a model for studies of early angiosperm evolution.*, in "Annals of Botany", September 2011, vol. 108, n^{o} 4, p. 589-98 [*DOI :* 10.1093/AOB/MCR088], http://hal.inria.fr/hal-00639984/en.

[21] 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^{o} 1, 303 [*DOI :* 10.1186/1471-2164-12-303], http://hal.inria.fr/inserm-00617206/en.

[22] I. WOHLERS, R. ANDONOV, G. W. KLAU. *Algorithm engineering for optimal alignment of protein structure distance matrices*, in "Optimization Letters", April 2011 [*DOI :* 10.1007/s11590-011-0313-3], http://hal.inria.fr/inria-00586067/en.

International Peer-Reviewed Conference/Proceedings

[23] R. CHIKHI, G. CHAPUIS, D. LAVENIER. *Parallel and memory-efficient reads indexing for genome assembly*, in "Parallel Bio-Computing 2011", torun, Poland, October 2011, http://hal.inria.fr/inria-00637536/en.

[24] R. CHIKHI, D. LAVENIER. *Localized genome assembly from reads to scaffolds: practical traversal of the paired string graph*, in "11th Workshop on Algorithms in Bioinformatics (WABI 2011)", T. PRZYTYCKA, M.-F. SAGOT (editors), Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2011, vol. 6833, p. 39-48 [*DOI :* 10.1007/978-3-642-23038-7_4], http://hal.inria.fr/inria-00637535/en.

[25] A. CORNU, S. DERRIEN, D. LAVENIER. *HLS Tools for FPGA : faster development with better performances*, in "Proceeding of the 7th International Symposium on Applied Reconfigurable Computing", Belfast, United Kingdom, March 2011, vol. 6578, p. 67-78, http://hal.inria.fr/hal-00637830/en.

[26] 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, August 2011, http://hal.inria.fr/inria-00608208/en.

[27] M. GIRAUD, S. JANOT, J.-F. BERTHELOT, C. DELTEL, L. JOURDAN, D. LAVENIER, H. TOUZET, J.-S. VARRÉ. *Biomanycores, open-source parallel code for many-core bioinformatics*, in "Bioinformatics Open Source Conference (BOSC 2011)", Vienne, Austria, 2011, http://hal.inria.fr/inria-00623390/en.

[28] N. MALOD-DOGNIN, M. LE BOUDIC-JAMIN, P. KAMATH, R. ANDONOV. *Using Dominances for Solving the Protein Family Identification Problem*, in "11th Workshop on Algorithms in Bioinformatics (WABI 2011)", T. PRZYTYCKA, M.-F. SAGOT (editors), Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2011, vol. 6833, p. 201-212 [*DOI :* 10.1007/978-3-642-23038-7_18], http://hal.inria.fr/inria-00609432/en.

[29] F. MOREEWS, J. PIAT, O. SALLOU. *OBIWEE : an open source bioinformatics cloud environment*, in "BOSC 2011 - 12th Annual Bioinformatics Open Source Conference", Vienne, Austria, July 2011, http://hal.inria.fr/inria-00638715/en.

[30] A. MUCHERINO, C. LAVOR, L. LIBERTI. *A symmetry-driven BP algorithm for the Discretizable Molecular Distance Geometry Problem*, in "IEEE International Conference on Bioinformatics and Biomedicine", Atlanta, United States, IEEE, 2011, http://hal.inria.fr/inria-00637771/en.

[31] A. MUCHERINO, I. WOHLERS, G. W. KLAU, R. ANDONOV. *Sparsifying Distance Matrices for Protein-Protein Structure Alignments*, in "CTW2011", Rome, Italy, June 2011, http://hal.inria.fr/hal-00642794/en.

[32] J. PIAT, F. MOREEWS, O. COLLIN, D. LAVENIER, A. CORNU. *SLICEE: A Service oriented middleware for intensive scientific computation*, in "SERVICES 2011", WASHINGTON DC, United States, IEEE (editor), July 2011, http://hal.inria.fr/inria-00638694/en.

National Peer-Reviewed Conference/Proceedings

[33] M. LE BOUDIC-JAMIN, N. MALOD-DOGNIN, A. CORNU, J. NICOLAS, R. ANDONOV. *Identification rapide de familles protéiques par dominance*, in "12th Annual Congress of the French National Society of Operations Research and Decision Science (ROADEF)", Saint-Étienne, France, École Nationale Supérieure des Mines de Saint-Étienne, March 2011, vol. 2, p. 791-792, Publié dans le douzième congrès de la Société Française de Recherche Opérationnelle et d'Aide à la Décision (ROADEF 2011)., http://hal.inria.fr/inria-00611457/en.

Workshops without Proceedings

[34] P. BLAVY, F. GONDRET, T. SVEN, C. GUZIOLOWSKI, S. LAGARRIGUE, J. VAN MILGEN, A. SIEGEL. *A new method for modeling reactions and regulations to analyze high-throughout data*, in "4th International

Symposium on Animal Functional Genomics", Dublin, Ireland, Sorcha De Gras, October 2011, http://hal.inria.fr/hal-00640720/en.

[35] G. CHAPUIS, O. FILANGI, P. LEROY, J.-M. ELSEN, D. LAVENIER. *GPU Accelerted QTLMap*, in "The 15th QTL-MAS workshop", France, May 2011, http://hal.inria.fr/hal-00637840/en.

Scientific Books (or Scientific Book chapters)

[36] J. BOURDON, D. EVEILLARD. *Probabilistic Approaches for Investigating Biological Networks*, in "Algorithms in Computational Molecular Biology: Techniques, Approaches and Applications", M. ELLOUMI, A. Y. ZOMAYA (editors), Wiley, January 2011, 1066, http://hal.inria.fr/hal-00531568/en.

[37] D. LAVENIER, G. RIZK, S. RAJOPADHYE. *GPU accelerated RNA folding algorithm*, in "GPU Computing Gems", W. MEI W. HWU (editor), ELSEVIER, February 2011, 560, pages 199-210 [*DOI :* 10.0123849888], http://hal.inria.fr/hal-00637827/en.

Research Reports

[38] G. GARET. *Apprentissage de grammaires algébriques par alignement multiple de séquences protéiques*, Université de Rennes 1, 2011, http://hal.inria.fr/hal-00639334/en.

[39] P. PETERLONGO, R. CHIKHI. *Mapsembler, targeted assembly of larges genomes on a desktop computer*, INRIA, March 2011, n^{o} RR-7565, http://hal.inria.fr/inria-00577218/en.

[40] C. VROLAND, C. BELLEANNÉE, J. NICOLAS. *Recherche et annotation des structures de CRISPR dans l'ensemble des génomes eucaryotes*, INRIA, 2011, http://hal.inria.fr/hal-00643408/en.

Scientific Popularization

[41] C. BLANCHET, O. COLLIN. *Un déluge de données*, in "Biofutur -Paris-", August 2011, http://hal.inria.fr/hal-00639962/en.

Other Publications

[42] A. CORNU, S. DERRIEN, D. LAVENIER. *How to accelerate genomic sequence alignment 4X using half an FPGA*, July 2012, to appear, http://hal.inria.fr/hal-00637833/en.