INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team amib

Algorithms and Models for Integrative Biology

Saclay - Île-de-France

Theme : Computational Biology and Bioinformatics

Activity Report

2010

Table of contents

1. Team

Research Scientists

Mireille Régnier [Team leader, Research Director (DR) Inria, HdR]
Julie Bernauer [Research Associate (CR) Inria]
Pierre Nicodème [Research Associate (CR) CNRS]
Yann Ponty [Research Associate (CR) CNRS]
Thomas Simonson [Research Director (DR) Ecole Polytechnique, HdR]

Faculty Members

Patrick Amar [Université Paris -Sud XI]
Jérôme Azé [Université Paris -Sud XI]
Sarah Cohen-Boulakia [Université Paris -Sud XI]
Alain Denise [Université Paris -Sud XI, HdR]
Christine Froidevaux [Université Paris -Sud XI, HdR]
Jean-Marc Steyaert [Ecole Polytechnique, HdR]

PhD Students

Zahira Aslaoui [Université Paris -Sud XI, on leave since 01/10/10]
Feng Lou [Université Paris -Sud XI]
Philippe Rinaudo [Université Paris -Sud XI]
Cédric Saule [Université Paris -Sud XI]
Audrey Sedano [Ecole Polytechnique, 50%]
Thuong Van Du Tran [Ecole Polytechnique]

Post-Doctoral Fellows

Thomas Bourquard [Université Paris-Sud XI, until 31/05/10]
Balaji Raman [Ecole Polytechnique]
Saad Sheikh [Ecole Polytechnique]

Administrative Assistant

Evelyne Rayssac [Secretary (SAR) Inria]

2. Overall Objectives

2.1. Overall Objectives

This project in bioinformatics is mainly concerned with the molecular levels of organization in the cell, dealing principally with RNAs and proteins; we currently concentrate our efforts on structure, interactions, evolution and annotation and aim at a contribution to protein and RNA engineering. On the one hand, we study and develop methodological approaches for dealing with macromolecular structures and annotation: the challenge is to develop abstract models that are computationally tractable and biologically relevant. On the other hand, we apply these computational approaches to several particular problems arising in fundamental molecular biology. These problems, described below, raise different computer science issues. To tackle them, the project members rely on a common methodology for which our group has a significant experience. The trade-off between the biological accuracy of the model and the computational tractability or efficiency is to be addressed in a closed partnership with experimental biology groups.

We investigate the relations between nucleotide sequences, 3D structures and, finally, biochemichal function. All protein functions and many RNA functions are intimately related to the three-dimensional molecular structure. Therefore, we view structure prediction and sequence analysis as an integral part of gene annotation that we study simultaneously and that we plan to pursue on a RNAomic and proteomic scale. Our starting point is the sequence either *ab initio* or with some knowledge such as a 3D structural template or ChIP-Chip experiments. We are interested in deciphering information organization in DNA sequences and identifying the role played by gene products: proteins and RNA, including noncoding RNA. A common toolkit of computational methods is developed, that relies notably on combinatorial algorithms, mathematical analysis of algorithms and data mining. One goal is to provide softwares or platform elements to predict either structures or structural and functional annotation. For instance, a by-product of 3D structure prediction for protein and RNA engineering is to allow to propose sequences with admissible structures. Statistical softwares for structural annotation are included in annotation tools developped by partners, notably our associate team MIGEC.

Our work is organized along two main axes. The first one is structure prediction, comparison and design engineering. The relation between nucleotide sequence and 3D macromolecular structure, and the relation between 3D structure and biochemical function are possibly the two foremost problems in molecular biology. There are considerable experimental difficulties in determining 3D structures to a high precision. Therefore, there is a crucial need for efficient computational methods for structure prediction, functional assignment and molecular engineering. A focus is given on both protein and RNA structures.

The second axis is structural and functional annotation, a special attention being paid to regulation. Structural annotation deals with the identification of genomic elements, e.g. genes, coding regions, non coding regions, regulatory motifs. Functional annotation consists in characterizing their function, e.g. attaching biological information to these genomic elements. Namely, it provides biochemical function, biological function, regulation and interactions involved and expression conditions. High-throughput technologies make automated annotation crucial. There is a need for relevant computational annotation methods that take into account as many characteristics of gene products as possible -intrinsic properties, evolutionary changes or relationships- and that can estimate the reliability of their own results.

3. Scientific Foundations

3.1. RNA and protein structures

3.1.1. RNA

Participants: Patrick Amar, Alain Denise, Yann Ponty, Balaji Raman, Mireille Régnier, Philippe Rinaudo, Cédric Saule, Jean-Marc Steyaert.

Common activity with P. Clote (Boston College and Digiteo).

3.1.1.1. Recoding events and riboswitches

Recoding represents several non conventional phenomena for the translation of messenger RNA (mRNA) into proteins, including *frameshift, readthrough, hopping,* where a single mRNA sequence allows the synthesis of (at least) two different polypeptides. Recoding is mandatory for many virus machinery and viability. We develop two complementary computational methods that aim to find genes subject to recoding events in genomes. The first one is based on a model for the recoding site ; the second one is based on a comparative genomics approach at a large scale. In both cases, our predictions are subject to experimental biological validation by our collaborators at IGM (Institut de Génétique et Microbiologie), Paris-Sud University. This work was funded by the ANR (project RNA-RECOD, ANR BLANC 2006-2010) and is currently funded by DIGITEO. .

Additionally, we are currently developing a combinatorial approach, based on random generation, to design small and structured RNAs. An application of such a methodology to the Gag-Pol HIV-1 frameshifting site will be carried out with our collaborators at IGM. We hope that, upon capturing the hybridization energy at the design stage, one will be able to gain control over the rate of frameshift and consequently fine-tune the expression of *Gag/Pol*. Our goal is to build these RNA sequences such that their hybridization with existing mRNAs will be favorable to independent folding, and will therefore affect the stability of some secondary structures involved in recoding events. Moreover it has been observed, mainly on bacteria, that some mRNA sequences may adopt an alternate fold. Such an event is called a riboswitch. A common feature of recoding events or riboswitches is that some structural elements on mRNA initiate unusual action of the ribosome or allow for an alternate fold under some environmental conditions. One challenge is to predict genes that might be subject to riboswitches. Additionnally, we are currently developing a combinatorial approach, based on random generation, to design small and structured RNAs. Our goal is to build these RNAs such that their hybridization with existing mRNAs will be favorable to independent folding, and will therefore affect the stability of some secondary structures involved in recoding events. An application of such a methodology to the *Gag-Pol HIV-1* frameshifting site will be carried out with our collaborators at IGM. We hope that, upon capturing the hybridization energy at the design stage, one will be able to gain control over the rate of frameshift and consequently fine-tune the expression of *Gag/Pol*.

3.1.1.2. Tertiary structures

One of our major challenges is to go beyond secondary structure, that is an intermediate structure level for RNA, between the single sequence and the full structure (tertiary structure). Over the past decade, few attempts have been made to predict the 3D structure of RNA from sequence only. So far, few groups have taken this leap. Despite the promises shown by their preliminary results, these approaches currently suffer to a limiting scale due to either their high algorithmic complexity or their difficult automation. Using our expertise in algorithmics and modeling, we plan to design original methods, notably within the AMIS-ARN project (selected by the ANR) in collaboration with E.Westhof's group at Strasbourg.

1. *Ab initio* modeling: Starting from the predicted RNA secondary structure, we aim to detect *local structural motifs* in it, giving local 3D conformations. It is based on pairing between complementary bases (A-U and C-G). The recent *Leontis-Westhof classification*, distinguishes twelve different kinds of chemical bonds between two nucleotides, according to the way they are linked together within the tertiary structure. This knowledge turns out to be crucial to determine molecular stability. Moreover, some recent works on RNA biochemistry have shown that RNA molecules are structured by *RNA tertiary motifs*. These motifs, that are known from 3D structure, can be seen as "small bricks" that play a very important role in RNA structuration. We develop graph algorithms for extracting tertiary motifs from RNA structures, and for predicting the tertiary structure from the sequence (thesis of M. Djelloul, defended in 2009). We use the resulting partial structure as a flexible scaffold for a multi-scale reconstruction, notably using game theory. We believe the latter paradigm offers a more realistic view of biological processes than global optimization, used by our competitors, and constitutes a real originality of our project.
2. Comparative modeling: we investigate new algorithms for predicting 3D structures by a comparative approach. This involves comparing multiple RNA sequences and structures at a large scale, that is not possible with current algorithms. Successful methods must rely both on new graph algorithms and on biological expertise on sequence-structure relations in RNA molecules.

3.1.1.3. RNA 3D structure evaluation

The biological function of macromolecules such as proteins and nucleic acids relies on their dynamic structural nature and their ability to interact with many different partners. Their function is mainly determined by the structure those molecules adopt as protein and nucleic acids differ from polypeptides and polynucleotides by their spatial organization. This is specially challenging for RNA where structure flexibility is key.

To address those issues, one has to explore the biologically possible spatial configurations of a macromolecule. The two most common techniques currently used in computational structural biology are Molecular Dynamics

(MD) and Monte Carlo techniques (MC). Those techniques require the evaluation of a potential or force-field, which for computational biology are often empirical. They mainly consist of a summation of bonded forces associated with chemical bonds, bond angles, and bond dihedrals, and non-bonded forces associated with van der Waals forces and electrostatic charge. Even if there exists implicit solvent models, they are yet not very well performing and still require a lot of computation time.

Our goal, in collaboration with the Levitt lab at Stanford University (Associate Team GNAPI http://www.lix.polytechnique.fr/~bernauer/EA_GNAPI/) is to develop knowledge-based potentials, based on measurements on known RNA 3D structure. Such potential are quick to evaluate during a simulation and can be used without having to explicitly address the solvent problem. They can be developed at various level of representation: atom, base, nucleotide, domain and could allow the modelling of a wide size range: from an hairpin to the whole ribosome. We also intend to combine these knowledge-based potentials with other potentials (hybrid modelling) and template-based techniques, allowing accurate modelling and dynamics study of very large RNA molecules. Such studies are still a challenge.

3.2. PROTEINS

Participants: Jérôme Azé, Julie Bernauer, Thomas Bourquard, Thomas Simonson, Jean-Marc Steyaert, Thuong Van Du Tran.

3.2.1. Docking and evolutionary algorithms

As mentioned above, the function of many proteins depends on their interaction with one or many partners. Docking is the study of how molecules interact. Despite the improvements due to structural genomics initiatives, the experimental solving of complex structures remains a difficult problem. The prediction of complexes, *docking*, proceeds in two steps: a configuration generation phase or *exploration* and an evaluation phase or *scoring*. As the verification of a predicted conformation is time consuming and very expensive, it is a real challenge to reduce the time dedicated to the analysis of complexes by the biologists. Various algorithms and techniques have been used to perform exploration and scoring [33]. The recent rounds of the CAPRI challenge show that real progress has been made using new techniques [30]. Our group has strong experience in cutting edge geometric modelling and scoring techniques using machine learning strategies for protein-protein complexes. In a collaboration with A. Poupon, INRA-Tours, a method that sorts the various potential conformations by decreasing probability of being real complexes has been developed. It relies on a ranking function that is learnt by an evolutionary algorithm. The learning data are given by a geometric modelling of each conformation obtained by the docking algorithm proposed by the biologists. Objective tests are needed for such predictive approaches. The *Critical Assessment of Predicted Interaction*, CAPRI, a community wide experiment modelled after CASP was set up in 2001 to achieve this goal (http://www.ebi.ac.uk/msd-srv/capri/). First results achieved for CAPRI'02 suggested that it is possible to find good conformations by using geometric information for complexes. This approach has been followed (see section New results). As this new algorithm will produce a huge amount of conformations, an adaptation of the ranking function learning step is needed to handle them. In the near future, we intend to extend our approach to protein-RNA complexes.

3.2.2. Computational Protein Design

A protein amino acid sequence determines its structure and biological function, but no concise and systematic set of rules has been stated up to now to describe the functions associated to a sequence; experimental methods are time (and money) consuming. Massive genome sequencing has revealed the sequences of millions of proteins, whereas roughly 55.000 3D protein structures, only, are known yet. Structure prediction *in silico* attempts to fill up the gap. It consists in finding a tentative spatial (3D) conformation that a given nucleotidic or aminoacid sequence is likely to adopt, using the modelling by homology. A second problem of interest is *inverse protein folding* or *computational protein design* (CPD): the prediction of (the most favorable) amino-acid sequences that adopt a particular target tertiary structure. One main question is to map the millions of protein sequences extracted from the genomes onto the tens of thousand known 3D structures. This problem has many implications such as protein folding and stability, structure prediction (fold recognition), or protein evolution. Moreover, it is a mandatory step towards the design of new, artificial proteins. The engineering of protein-ligand interactions also has great biological and technological value. For example, the recent engineering of aminoacyl-tRNA synthetase (aaRS) enzymes has led to organisms with a modified genetic code, expanded to include nonnatural aminoacids.

Another novel ingredient is the use of *negative design*: the ability to select against sequences that have undesired properties, such as a tendency to fold into alternate, undesired structures. It can be critical for attaining specificity when competing states are close in (stability) structure space. There are also current efforts to enlarge this thermodynamical point of view by a new knowledge on natural proteins with known conformations.

3.2.3. Transmembrane proteins

Our goal is to predict the structure of different classes of *barrel proteins*. Those proteins contain the two large classes of transmembrane proteins, which carry out important functions. Nevertheless, their structure is yet difficult to determine by standard experimental methods such as X-ray cristallography or NMR. Most existing methods only address single-domain protein structures. Therefore, for large proteins, a preprocessing to determine the protein domains is necessary. Then, a suitable model of energy functions needs to be designed for each specific class. We have designed a pseudo-energy minimization method for the prediction of the super-secondary structure of β-barrel or α-helical-barrel proteins with structural knowledge-based enhancement. The method relies on graph based modelling and also deals with various topological constraints such as Greek key or Jelly roll conformations.

3.3. Annotation and Combinatorics

3.3.1. Word counting

Participants: Alain Denise, Pierre Nicodème, Mireille Régnier, Cédric Saule, Saad Sheikh, Jean-Marc Steyaert.

We aim at enumerating or generating sequences or structures that are *admissible* in the sense that they are likely to possess some given biological property. Team members have a common expertise in enumeration and random generation of combinatorial structures. They have developped computational tools for probability distributions on combinatorial objects, using in particular generating functions and analytic combinatorics. Admissibility criteria can be mainly statistic; they can also rely on the optimisation of some biological parameter, such as an energy function.

The ability to distinguish a significant event from statistical noise is a crucial need in bioinformatics. In a first step, one defines a suitable probabilistic model (null model) that takes into account the relevant biological properties on the structures of interest. A second step is to develop accurate criteria for assessing (or not) their exceptionality. An event observed in biological sequences, is considered as exceptional, and therefore biologically significant, if the probability that it occurs is very small in the null model. Our approach to compute such a probability consists in an enumeration of good structures or combinatorial objects. Thirdly, it is necessary to design and implement efficient algorithms to compute these formulae or to generate random data sets. Two typical examples that motivate research on words and motifs counting are *Transcription Factor Binding Sites*, TFBSs, consensus models of recoding events and some RNA secondary structures. The project has a significant contribution in word enumeration area. When relevant motifs do not resort to regular languages, one may still take advantage of combinatorial properties to define functions whose study is amenable to our algebraic tools. One may cite secondary structures and recoding events.

3.3.2. Random generation

Participants: Alain Denise, Yann Ponty, Cédric Saule.

Analytical methods may fail when both sequential and structural constraints of sequences are to be modelled or, more generally, when molecular *structures* such as RNA structures have to be handled. The random generation of combinatorial objects is an alternative, yet natural, framework to assess the significance of observed phenomena. General and efficient techniques have been developed over the last decades to draw objects uniformly at random from an abstract specification. However, in the context of biological sequences and structures, the uniformity assumption fails and one has to consider non-uniform distributions in order to obtain relevant estimates. Typically, context-free grammars can handle certain kinds of long-range interactions such as base pairings in secondary RNA structures. Stochastic context-free grammars (SCFG's) have long been used to model both structural and statistical properties of genomic sequences, particularly for predicting the structure of sequences or for searching for motifs. They can also be used to generate random sequences. However, they do not allow the user to fix the length of these sequences. We developed algorithms for random structures generation that respect a given probability distribution on their components. For this purpose, we first translate the (biological) structures into combinatorial classes, according to the framework developed by Flajolet *et al*. Our approach is based on the concept of *weighted* combinatorial classes, in combination with the so-named *recursive* method for generating combinatorial structures. Putting weights on the atoms allows to bias the probabilities in order to get the desired distribution. The main issue is to develop efficient algorithms for finding the suitable weights. An implementation is given in the `GenRGenS` software http://www.lri.fr/~genrgens/.

Recently a new paradigm appeared is in *ab initio* secondary structure prediction [28]: in place of classical optimization algorithms, the new approach relies on probabilistic algorithms, based on statistical sampling within the space of solutions. Indeed, we have done significant and original progress in this area recently [2], [6], [13], including combinatorial models for structures with pseudoknots. Our aim is to combine this paradigm with a fragment based approach for decomposing structures, such as the cycle decomposition by F. Major's group [32].

Besides, our work on random generation is also applied in a different fields, namely software testing and model-checking, in collaboration with the Fortesse group at LRI [21].

3.3.3. *Knowledge extraction*

Participants: Zahira Aslaoui, Jérôme Azé, Sarah Cohen-Boulakia, Christine Froidevaux, Mireille Régnier.

Our main goal is to design semi-automatic methods for annotation. A possible approach is to focus on the way we could discover relevant motifs in order to make more precise links between function and motifs sequence. Indeed, a commonly accepted hypothesis is that function depends on the order of the motifs present in a genomic sequence. Examples of relevant motifs can be frameshift motifs, RNA structural motifs, TFBS or PFAM domains. General tools must then be developed in order to assess the significance of the motifs found out. Likewise we must be able to evaluate the quality of the annotation obtained. This necessitates giving an estimate of the reliability of the results that includes a rigorous statement of the validity domain of algorithms and knowledge of the results provenance. We are interested in provenance resulting from workflow management systems that are important in scientific applications for managing large-scale experiments and can be useful to calculate functional annotations. A given workflow may be executed many times, generating huge amounts of information about data produced and consumed. Given the growing availability of this information, there is an increasing interest in mining it to understand the difference in results produced by different executions.

3.3.4. *Systems Biology*

Systems Biology involves the systematic study of complex interactions in biological systems using an integrative approach. The goal is to find new emergent properties that may arise from the systemic view in order to understand the wide variety of processes that happen in a biological system. Systems Biology activity can be seen as a cycle composed of theory, computational modelling to propose a hypothesis about a biological process, experimental validation, and use of the experimental results to refine or invalidate the computational model (or even the whole theory).

In this context, the AMIB group is working on two axes.

On the one hand, we work on helping the design and understanding of the biological relationships between proteins involved in signalling pathways. More precisely, we work with the BIOS group from INRA-TOURS (A. Poupon) within the ASAM project on the understanding of signalling pathways involving G protein-coupled receptors (GPCR). Our aim is to design a knowledge base containing expert rules able to interpret various and highly numerous experimental results and semi automatically construct signalling networks (from a statical point of view). In this work, we are particularly interested in storing information about the quality of each piece of information in the knowledge base, which may depend on various criteria (a piece of data obtained by various experiments or by experiments of high quality etc.).

On the other hand, we concentrate on the computational modelling step of the cycle by developing a computer simulation system, HSIM, that mimics the interactions of biomolecules in an environment modelling the membranes and compartments found in real cells. In collaboration with biologists from the AMMIS lab. at Rouen we have used HSIM to show the properties of grouping the enzymes of the phosphotransferase system and the glycolytic pathway into metabolons in *E. coli*. In another collaboration with the SYSDIAG at Montpellier, we participate at the COMPUBIOTIC project. This is a Synthetic Biology project in the field of medical diagnosis: its goal is to design a small vesicle containing specific proteins and membrane receptors. These components are chosen in a way that their interactions can sense and report the presence in the environment of molecules involved in human pathologies. We used HSIM to help the design and to test qualitatively and quantitatively this *"biological computer"* before *in vitro*.

4. Software

4.1. VARNA

Participants: Yann Ponty [correspondant], Alain Denise.

VARNA is a tool for the automated drawing, visualization and annotation of the secondary structure of RNA, designed as a companion software for web servers and databases. VARNA implements four drawing algorithms, supports input/output using the classic formats *dbn, ct, bpseq* and *RNAML* and exports the drawing as five picture formats, either pixel-based (*JPEG, PNG*) or vector-based (*SVG, EPS* and *XFIG*). It also allows manual modification and structural annotation of the resulting drawings using either an interactive point and click approach, within a web server or through command-line arguments. VARNA is a free software distributed under the terms of the GPLv3.0 license and available at http://varna.lri.fr.

VARNA is currently used by RNA scientists (Cited/used by 10 research articles in 2010), webservers such as the BOULDEALE webserver (http://www.microbio.me/boulderale/), the TFOLD webserver (http://tfold.ibisc.univ-evry.fr/TFold/),the CYLOFOLD webserver (http://cylofold.abcc.ncifcrf.gov/), and by databases such as the IRESITE database (http://iresite.org/) and the SRNATARBASE (http://ccb.bmi.ac.cn/srnatarbase/). Remarkably the applet is currently used by the RFAM database (http://rfam.sanger.ac.uk/), the main source of sequence/structure data for RNA scientist, to display secondary structures.

4.2. HSIM

Participant: Patrick Amar [correspondant].

HSIM is a simulation tool for studying the dynamics of biochemical processes in a virtual bacteria. The model is given using a language based on probabilistic rewriting rules that mimics the reactions between biochemical species. HSIM is a stochastic automaton which implements an entity-centered model of objects. This kind of modelling approach isan attractive alternative to differential equations for studying the diffusion and interaction of the many different enzymes and metabolites in cells which may be present in either small or large numbers. This software is freely available at http://www.lri.fr/~pa/Hsim; A compiled version is available for the Windows, Linux and MacOSX operating systems.

5. New Results

5.1. RNA structures

5.1.1. RNA secondary structure alignement

In [5], in collaboration with groups in Bordeaux, Lille and Marne La VallÃ©e, we provided a thorough analysis of the RNA secondary structure alignment, hierarchy, including a new polynomial time algorithm and an NP-completeness proof. The polynomial time algorithm involves biologically relevant evolutionary operations, such as pairing or unpairing nucleotides. In [8] , we proved that the average complexity of the pairwise ordered tree alignment algorithm of Jiang, Wang and Zhang is in O(nm), where n and m stand for the sizes of the two trees, respectively. It is shown that the same result holds for the average complexity of pairwise comparison of RNA secondary structures, using a set of biologically relevant operations.

5.1.2. Design

In a collaboration with a group of molecular biology in Wuhan, we studied the effect of RNA structures on the activity of exonic splicing enhancers on the SMN1 minigene model by engineering known ESEs into different positions of stable hairpins [16]. For that purpose, we developed an original algorithm for designing simple RNA structures, as hairpins, subject to sequence constraints, as the presence or absence of particular motifs. This work is a first step towards design of complex secondary structures subject to sequence constraints, that will be of great use for biologists.

Furthermore the weighted models introduced within the project have already shown useful within an exploration of the mutational landscape of RNA, performed in a collaboration with Y. Ponty and J. Waldispuhl (McGill, Canada) (Accepted for presentation at the RECOMB'11 conference [22]). In this collaboration, weights were used to compensate a bias toward regions of higher GC-content within sampled sequences, thus allowing for the exploration of more relevant portions of the evolutionary landscape.

5.1.3. RNA knowledge-based potentials and 3D studies

Being able to build such potentials require good initial experimental data. We have a manually curated database of biologically interesting structures on which statistical analysis can be performed (reasonably-sized and non-redundant). The database server for the dataset http://csb.stanford.edu/rna is up and running and was used to initiate discussions on template based RNA construction. This discussion involves Ph. Rinaudo and A. Denise.

Following various adjustments on the handling of topology and covalent bonds filtering, the RNA knowledge-based potential now performs reasonably well in its different flavors (depending on how local covalent bonds are treated on a coarse grained level). It can easily be used in three well-known Molecular Dynamics (MD) and modeling software suites ENCAD [34], GROMACS (v3 and 4) [35] and MOSAICS [31]. A large number of new decoys were generated to fully evaluate and compare the potential to the Rosetta RNA scoring function [27] which includes corrections for base pairing and stacking constraints. These decoys were generated using different techniques. Beyond generating nearânative decoys for crystallographically determined RNA structures, we also generated decoys for some structures with coordinates solved by NMR. We found that these decoys are extremely hard to handle as the ânativeâ conformation is illâdefined and likely consists of an ensemble of similar structures that a single model may not adequately represent. We show that our KB potential often performs just as well, or better than Rosetta RNA, which is the current stateâofâtheâart scoring standard for RNA structure. We also noticed that the inclusion of structural terms in the potential such as pairing and stacking is not always increasing the accuracy of the scoring procedure.

This work was done in collaboration with A. Sim, X. Huang an M. Levitt (Stanford University - GNAPI Associate team).

5.1.4. Thermodynamics

A non-Boltzmannian Monte Carlo algorithm was designed by Wang and Landau to estimate the density of states for complex systems, such as the Ising model, that exhibit a phase transition. In [10], we applied the Wang-Landau (WL) method to compute the density of states for secondary structures of a given RNA sequence, and for hybridizations of two RNA sequences. Our method is shown to be much faster than existent software, such as RNAsubopt. From density of states, we compute the partition function over all secondary structures and over all pseudoknot-free hybridizations. The advantage of the WL method is that by adding a function to evaluate the free energy of arbitary pseudoknotted structures and of arbitrary hybridizations, we can estimate thermodynamic parameters for situations known to be NP-complete.

5.2. Proteins structures

5.2.1. Protein-protein interaction

A protein-protein docking procedure traditionally consists in two successive tasks: a search algorithm generates a large number of candidate solutions, and then a scoring function is used to rank them in order to extract a native-like conformation. We have already demonstrated that using Voronoi constructions and a defined set of parameters, we could optimize an accurate scoring function. However, the precision of such a function is still not sufficient for large-scale exploration of the interactome.

Another geometric construction was also tested: the Laguerre tessellation. It also allows fast computation without losing the intrinsic properties of the biological objects. Related to the Voronoi construction, it was expected to better represent the physico-chemical properties of the partners. In , we present the comparison between both constructions.

We also worked on introducing a hierarchical analysis of the original complex three-dimensional structures used for learning, obtained by clustering. Using this clustering model we can optimize the scoring functions and get more accurate solutions. This scoring function has been tested on CAPRI scoring ensembles, and an at least acceptable conformation is found in the top 10 ranked solutions in all cases. This work was part of the thesis of Thomas Bourquard, defended in 2009/

A strong emphasis was recently made on the design of efficient complex filters. To achieve this goal, we focused on the use of collaborative filtering methods state of the art machine learning approaches combined with our genetic algorithm. This work has been submitted for publication.

We also decided to extend these techniques to the analysis of protein-nucleic acid complexes. The first preliminary developments and tests were performed by Adrien Guilhot during his M1 internship for two months.

5.2.2. Computational protein design

A. Sedano has studied the inverse folding problem of proteins during her internship supervised by T. Simonson and J.-M. Steyaert. She applied methods of probability analysis, such as those of Ranganathan, Thirumalai or Nussinov to big sets of sequences of the family of domains *PDZ* (at first calculated then natural) [14]. These methods allow to determine what are the correlations between distant mutations in a structure. Later, these correlations should allow to describe in terms of sequence the *signature* of a given structure. She also tried to test these methods by working not on mutations between amino acids but on mutations between classes of amino acids, to facilitate the comparisons between sites along the sequence.

5.2.3. Transmembrane β-barrels

We have recently proposed an algorithm [23] that classifies Transmembrane β-Barrel Proteins (TMB) and predicts their structure. It first uses a simple probabilistic model to filter out the proteins and strands which are not beta-barrel. Then, we build a graph-theoretic model to fold into the super-secondary structure via dynamic programming. This step runs in $O(n^3)$ time for the common up-down topology, and at most $O(n^5)$ for the Greek key motifs, where n is the number of amino acids. Finally a predicted three-dimensional structure is built from the geometric criteria. If the pseudoenergy is insufficient, the protein is classified as a non-TMB protein. We have tested this approach on TMB and non-TMB proteins for classification and structure prediction. We tested classification on a dataset of 14238 proteins including 48 TMB and 14190 non-TMB proteins. Our classification results are very accurate and comparable to other algorithms. Especially, our PPV, MCC and F-Scores are second only to a very recent algorithm by Freeman and Wimley [29], which relies heavily on training data. We also tested the structure prediction on 42 proteins from the TMB and compared to other existing algorithms. The results are comparable to existing algorithms, the accuracy ranges from 85-93%, depending upon the parameter used. This is very promising given that other algorithms rely heavily on homology and training datasets and may be overfitting. Our approach can be further improved by refining the energetic model, especially on turns and loops.

5.3. Combinatorics and Annotation

5.3.1. Word counting, trees and automata

Cis-Regulatory modules (CRMs) of eukaryotic genes often contain multiple binding sites for transcription factors, or clusters. Formally, such sites can be viewed as *words* co-occurring in the DNA sequence. This gives rise to the problem of calculating the statistical significance of the event that multiple sites, recognized by different factors, would be found simultaneously in a text of a fixed length. A new project aims at studying by au- tomata waiting times for promoters in the context of the evolution of promoters sequences. A wide-scale analysis of these waiting times has been recently done by S. Behrens and M. Vingron, of the Max-Planck Institute for Molecular Biology of Berlin; this study is done by a purely mathematical approach, but does important simplifications by assuming that the overlaps of words are negligible. In a collaboration of P. Nicodème with C. Nicaud and S. Behrens (presently at the University of Munster), an automaton approach has been designed that is subject to no restrictive assumptions. The implementation has begun and the results will be compared with Behrens and Vingronâs results.

As a collaboration of P. Nicodème with F. Bassino, LIPN, University Paris-North, and J. Clément, GREYC, University of Caen, an article tackling the most general case of statis- tics of occurrences of a finite pattern has been worked out in its final version. In particular, links between the Aho-Corasick automaton and the correlations of words have been set out in full details. The explicit formulas for the two first moments of the number of occurrences give good hope that the multi- variate limiting distribution could also be obtained in this general case, which would improve upon previous results of Bender and Kochman. Suffix-trees are a major tool of indexation of large sequences, in particular DNA sequences. The main difficulty comes from overlapping occurrences of motifs. This is partially solved by our previous algorithm, AHOPRO. OVGRAPH, developed with our former associate team MIGEC, intending to solve memory problems. We introduced a new concept of overlap graphs to count word occurrences and their probabilities.

Our preprocessing uses a variant of Aho-Corasick automaton and achieves $O(m|\mathcal{H}|)$ time complexity. Our algorithm is implemented for the Bernoulli and the Markov models and provides a significant space improvement in practice. It is available at http://www.impb.ru/index.php?lang=eng.

In a collaboration with M. Ward, Purdue University, USA, P. Nicodème considers second moments of parameters of suffix-trees, and in particular the second moment of the profiles of these trees. M. Régnier and S. Sheikh address combinatorial problems on clumps.

5.3.2. Random walks

The collaboration of P. Nicodème with C. Banderier, LIPN, University Paris North, about Bounded Random Walks led [17]to derive explicit mathematical formulas for the probability that discrete random walks remain under a barrier, when considering a large class of increments. In particular, a constant time heuristics can be applied to biological data to signal exceptional behaviors of ranking of genes expression, which can be used for medical diagnosis. The biological application of the results obtained with C. Banderier shall be done by M. Schulz, previously also at the Max-Planck Institute of Molecular Genetics and presently at the University of Pittsburgh.

5.3.3. Counting pseudoknots

In 2004, Condon and coauthors gave a hierarchical classification of exact RNA structure prediction algorithms according to the generality of structure classes that they handle. In [13], we completed this classification by adding two recent prediction algorithms. More importantly, we precisely quantified the hierarchy by giving closed or asymptotic formulas for the theoretical number of structures of given size n in all the classes but one. This allows to assess the tradeoff between the expressiveness and the computational complexity of RNA structure prediction algorithms. Additionally, using bijections between the structure classes and sole context-free languages, we were able to develop new and efficient algorithms for the random generation of RNA pseudoknotted structures [13].

5.3.4. Random Generation

In [25], we developed a new algorithm for generating uniformly at random words of any regular language $\mathcal{L}$. When using floating point arithmetics, its bit-complexity is $\mathcal{O}(q\log^2 n)$ in space and $\mathcal{O}(qn\log^2 n)$ in time, where n stands for the length of the word, and q stands for the number of states of a finite deterministic automaton of $\mathcal{L}$. Compared to the known best alternatives, our algorithm offers an excellent compromise in terms of space and time complexities.

In a collaboration with M. Termier (IGM-University Paris-Sud XI), we introduced and studied a generalization of the weighted models to general decomposable classes defined for k different types of atoms. For these models we derived efficient algorithms based on the so-called recursive method. Furthermore we gave a heuristic optimization scheme for a natural inverse problem, ie figuring out weights such that targeted frequencies of atoms are obtained *on the average*. These results recently appeared in the *Theoretical Computer Science* journal [6], and provide new foundations and tools for tackling structural bioinformatics problems, such as RNA design.

In a collaboration with O. Bodini (LIP6-University Pierre et Marie Curie) the previous was work was recently extended and lifted into a weighted version of the Boltzmann sampling. We proposed a Newton iteration to figure out suitable weights, solving exactly the inverse problem for which only a heuristic was known [6]. This iteration was coupled with a multi-dimensional rejection scheme which we analyzed as a generalization of the analysis performed in the seminal paper. This gave a $\Theta(n^{2+k/2})/\Theta(n)$ time/memory algorithm for the random generation of words of a given composition while the best known algorithm for this problem had complexities in $\Theta(n^k)/\Theta(n^k)$. This work was presented at the AOFA'10 conference in Vienna [18].

Finally we analyzed the redundancy of sampled sets of weighted objects in a collaboration between Y. Ponty and D. Gardy (PRISM, University Versailles/St-Quentin). More specifically, assuming one knows how to draw weighted objects at random from a finite set, we addressed the four following questions: How many objects does one need to generate before some object is observed twice? How many objects must be generated before each objects is obtained at least once? How many distinct objects does one obtain after drawing k objects? Which proportion of the distribution is covered after drawing k samples? For all these questions, we obtained efficient algorithms and/or asymptotical estimates when the objects are words of a context-free language of a given size n. The results of this study, which were presented at the GASCOM'10 conference in Montreal [24], give direct insight into the statistical property of sets of structures produced by RNA statistical sampling algorithms.

5.3.5. Scientific workflows

We have followed our work on scientific workflows [1] . We have focused on the problems posed by proprietary modules (e.g., unpublished methods) as well as private or confidential data (e.g., unpublished genomes) in scientific workflows. We have thus started to work on the intricate problem of providing answers to provenance queries over executions of a given workflow without revealing private information [20].

5.3.6. Data integration

Recent years have seen a revitalization of Data Integration research in the Life Sciences. But the perception of the problem has changed: While early approaches concentrated on handling schema-dependent queries over heterogeneous and distributed databases, current research emphasizes instances rather than schemas, tries to place the human back into the loop, and intertwines data integration and data analysis. In [19], we have reviewed the past and current state of data integration for the Life Sciences and discussed recent trends in detail, which all pose various challenges for the database community.

6. Contracts and Grants with Industry

6.1. National Initiatives

6.1.1. ANR

RNA-RECOD, ANR BLANC 2006-2010: *Influence of mRNA structures on ribosome accuracy*. Normal decoding could be diverted by sequences and structures on the mRNA and led to recoding. Analysing these variations constitutes a powerful tool to understand the normal curse of action of the translational machinery. The four teams involved in the project develop complementary approaches that have previously allowed the identification of several elements involved in recoding. Very recently, using a cryo-eletromicroscopy approach, we deciphered for the first time the precise role of the pseudoknot in a -1 frameshifting event. The project gathers together several complementary approaches including biochemistry, genetics, molecular and structural biology and bioinformatics. The goal of the study is to i) compare the molecular mechanisms involved in several recoding events (-1 and +1 frameshifting, pyrrolysine incorporation), focusing on the associated structural modifications and ii) identify new recoding sites in genomes.

AMIS-ARN, ANR BLANC 2009-2012: *Graph Algorithms and Automatic Softwares for Interactive RNA Structure Modelling*. We aim to do substantial progress in the problem of automatically or semi-automatically modelling the three-dimensional structure of RNA molecules, given their sequence. By *semi-automatically* we mean developing algorithms and software that can automatically propose (good) solutions, and that can efficiently compute alternative solutions according to some new constraints or some new hypotheses given by the expert modeler. More precisely, we plan to work on the three following points: 1.Development of computational methods for solving some key steps necessary for modelling RNA 3D structures. These methods will rely on new graph algorithms for molecular structures and on biological expertise on sequence-structure relations in RNA molecules. 2.Implemention of these methods in a software suite, PARADISE, which is being developed by one of the partners (E. Westhof's lab, Strasbourg University) and which will be made freely available to the scientific community. 3. Application of these methods in order to model several molecules of interest.

ANR-MAGNUM, ANR BLANC 2010-2014: *Algorithmic methods for the non-uniform random generation: Models and applications*. The central theme of the MAGNUM project is the elaboration of complex discrete models that are of broad applicability in several areas of computer science. A major motivation for the development of such models is the design and analysis of efficient algorithms dedicated to simulation of large discrete systems and random generation of large combinatorial structures. Another important motivation is to revisit the area of average-case complexity theory under the angle of realistic data models. The project proposes to develop the general theory of complex discrete models, devise new algorithms for random generation and simulation, as well as bridge the gap between theoretical analyses and practically meaningful data models. The sophisticated methods developed during the past decades make it possible to enumerate and quantify parameters of a large variety of combinatorial models, including trees, graphs, words and languages, permutations, etc. However these methods are mostly targeted at the analysis of uniform models , where, typically, all words (or graphs or trees) are taken with equal likelihood. The MAGNUM project proposes to depart from this uniformity assumption and develop new classes of models that bear a fair relevance to real-life data, while being, at the same time, still mathematically tractable. Such models are the ones most likely to be connected with efficient algorithms and data structures.

6.1.2. PRES

LRI and INRA-MIG are partners in a one-year regional project AFON: *Annotation FONctionnelle (Functional Annotation)*. The aim of the project is to design semi-automatic methods to help scientists in the task of functional annotation of prokaryotic genomes.

6.1.3. Inria-Inra

AMIB and INRA-TOURS (A. Poupon) are partners in a two years project ASAM. This project aims to help the understanding of signalling pathways involving G protein-coupled receptors (*GPCR*) which are excellent targets in paramacogenomics research. Large amounts of experiments are available in this context while globally interpreting all the experimental data remains a very challenging task for biologists. The aim of ASAM is thus to provide means to semi automatically construct signalling networks of GPCRS. In particular, ASAM aims to base its solution on the design of a knowledge base containing expert rules able to interpret various experimental results and semi automatically construct signalling networks. Interestingly, each piece of the network (a piece of data or a relationship between pieces of data) may be associated with quality information depending on various criteria (a piece of data obtained by various experiments or by experiments of high quality etc.).

7. Other Grants and Activities

7.1. International Initiatives

7.1.1. Digiteo

Participants: Alain Denise, Feng Lou, Balaji Raman, Jean-Marc Steyaert.

P. Clote (Boston College) is a DIGITEO chair. The project deals with RNA properties, with a focus on folding energy distributions and the identification of riboswitches.

7.1.2. Associate Team

The Associate team GNAPI http://www.lix.polytechnique.fr/~bernauer/EA_GNAPI/ (Geometric and knowledge-based analysis for Nucleic Acid and Protein dynamics and Interactions) is a collaboration between the AMIB project team and the Computational Structural Biology lab at Stanford University. The purpose of this team is to develop novel computational structural biology techniques based on geometry and statistics for the accurate modelling of RNA structures. This includes the development of knowledge based potentials for RNA structure but also techniques for using known 3D information to better model the structure and the dynamics of 3D RNA molecules. The Levitt lab at Stanford has expertise on computational structural biology techniques such as molecular modelling and sampling. It also has large computational resources. Our group provides the knowledge of 3D structures, geometry and computer science tools for the study of RNA. This allows the development of new potentials and data analysis strategies for these techniques and the combination of these development with classical analyses.

7.2. Exterior research visitors

7.2.1. Long stays

- Ulf Leser (Humboldt University, Berlin) visited for 6 months.
- S. Hamel (University of Montreal) visited for one month.
- V. Makeev (NIIGenetika, Moscow) visited for one month.

7.2.2. Short stays

- M. Roytberg (IMPB, Moscow) and E.Furletova (PhD. IMPB) visited for two weeks.

- D. Saakian (Institute of Physics, Academia Sinica, Taiwan) visited for ten days.
- Xuhui Huang (HKUST. Hong-Kong), P. Mignary (Stanford University, USA) and A. Sim (PhD, Stanford University) visited for 10 days (GnaPI associate team).
- M. Ward (Purdue University, USA) and B. Ludascher (UC Davis, USA) visited for one week.

8. Dissemination

8.1. Animation of the scientific community

8.1.1. French Bioinformatics

Participants: Patrick Amar, Jérôme Azé, Julie Bernauer, Thomas Bourquard, Sarah Cohen-Boulakia, Alain Denise, Christine Froidevaux, Feng Lou, Pierre Nicodème, Yann Ponty, Mireille Régnier, Cédric Saule, Jean-Marc Steyaert.

All team is involved in GDR-Bim (Biology, Computer Science and Mathematics, http://www.gdr-bim.u-psud.fr/). J. Azé is the webmaster. Ch. Froidevaux and S. Cohen-Boulakia participate to the subdomain *Knowledge Representation, Ontologies, Data Integration and Grids*, A. Denise, P. Nicodème, M. Régnier and C. Saule participate to the subdomain Sequence Analysis and to Comatege subgroup of GDR-Im (Informatique MathÃ©matique, http://www.gdr-im.fr/)

Many members participate to Alea working group (http://algo.inria.fr/AofA/Alea/index.html. Y. Ponty, M. Régnier and C. Saule gave a talk at Alea'2010 (http://newton.univ-mrs.fr/liste_rencontre/Rencontres2010/Renc401/Renc401.html).

8.1.2. Seminars

8.1.2.1. Amib seminars

We received in our weekly seminar: Ambuj Singh (UC Santa Barbara), B. LudÃ¤scher (UC Davis) S. Flores (Stanford), P. Clote (Boston College), A. Sim (stanford), D. Saakian (Academica Sinica, Taiwan), P. Minary (Stanford), X. Huang (HKUST), S. Hamel (U. Montreal).

We also received F. Coste (Inria-Symbiose), A.Lamiable (UVSQ), M. Chabbert (U. Angers), B. Schikowski (Institut Pasteur), G. Boldina (IECB, Bordeaux), A. Mucherino (Inria-Lille), S. Tempel, S. Peres (Sysdiag), A.Mathelier (UPMC), F. Le Bitoux (U. Perpignan), S. Pradalier (Inria-Contraintes).

8.1.2.2. Other seminars

P. Amar was invited to give the seminar *La programmation multiagents et son intÃ©rÃªt pour l'Ã©tude des systÃ¨mes complexes* at the AMMIS lab. University of Rouen on March, 3rd. 2010. He was invited to give the talk *ModÃ©lisation de l'auto assemblage et du comportement de complexes macro molÃ©culaires* at the Laboratoire d'Informatique, Signaux et SystÃ¨mes de Sophia-Antipolis on July, 1st 2010.

S. Cohen-Boulakia and Ch. Froidevaux have been invited to give a talk in the context of the *Scientific day 2010 â Data integration for the Life Sciences* organized by the PPF bio-informatique of Lille, on May 2010 at Institut Pasteur of Lille. S. Cohen-Boulakia has been invited to give a talk on scientific workflows in the context of the ANR BioWIC project, in June 2010, Perpignan.

A. Denise gave an invited talk at Lacim2010, UniversitÃ© du QuÃ©bec Ã MontrÃ©al, http://lacim2010.lacim.uqam.ca/ and at the workshop Computational Models for RNA structures, at McGill University, Montreal, He gave a talk at SÃ©minaire du DIRO, UniversitÃ© de MontrÃ©al.

Y. Ponty gave talks at the Arena'10 (Toulouse), Alea'10 (Luminy) and *Computational models for RNA* (McGill, Canada) workshops. He was invited to participate at the RNA Ontology Consortium meeting (Strasbourg).

C. Saule attended FPSAC'10, ALEA'10, GTSEQ2010 http://www.lirmm.fr/mab/gtseq/ and *Bioinformatics after Next Generation Sequencing* http://line.imb.ac.ru/NGS_workshop_2010/ngs_home.rhtml. He gave talks and presented posters.

Thuong Van Du Tran attended ISMB/ECCB2010 (Ghent,Belgium) and presented a poster.

M. Régnier, J.-M. Steyaert and L. Schwartz organized a one-day meeting on *Cancer as a metabolic disease* at Ecole Polytechnique, on January, 18th, that involved French and Italian researchers.

8.1.3. Program Committee

P. Amar was chairman of the organising committee, and a member of the scientific committee as well, for the conference *Modelling Complex Biological Systems in the context of genomics* http://epigenomique.free.fr/en/.

J. Bernauer is chair of *Multi-resolution Modeling of Biological Macromolecules* session at the Pacific Symposium on Biocomputing 2011 http://psb.stanford.edu/.

S. Cohen-Boulakia is a program committee member of ICDE 2010 http://www.icde2010.org/ (where she won the outstanding reviewer award), SWPM 2010, ISWC (International Workshop on the role of Semantic Web in Provenance Management) and JOBIM 2010, http://www.jobim2010.fr/.

A. Denise is a member of the editorial board of Technique et Science Informatiques (HermÃ¨s).

Ch. Froidevaux is the co-chair of JOBIM 2010, member of the Program Comittee for EDBT 2010 (13th International Conference on Extending Database Technology), IB2010 (International Symposium on Integrative Bioinformatics) and JOBIM 2011.

Y. Ponty is a member of the organizing and program committees of JOBIM2011. He organizes with E. Fusy and G. Schaeffer (CNRS, LIX-Ecole Polytechnique) the 2011 edition of the ALEA meeting, to be held in Spring 2011 at CIRM, Luminy.

M. Régnier organized with V. Makeev (GosNIIGenetika) a 3-days workshop *Bioinformatics after Next Generation Sequencing* http://line.imb.ac.ru/NGS_workshop_2010/ngs_home.rhtml. This workshop was labellized as an event of France-Russie year. It was supported by INRIA, with the participation of five INRIA teams in Computational Sciences for Biology, Medicine and the Environment research theme, CNRS and RFBR http://www.rfbr.ru/.

AMIB organized the LIX Colloquium http://www.lix.polytechnique.fr/bioinfo/colloquium2010/ at fall 2010. This 3-day event, featuring talks and a poster session, was focused on four bioinformatics core-topics: *High-throughput and Omics*, *RNA in silico biology*, *Computational Structural Biology* and *Systems Biology*. It was supported by CNRS and GDR-BIM. There were 110 attendees, including 30 foreigners.

8.1.4. Research Administration

P. Amar served in Evaluation committees of AERES for SYSDIAG (CNRS Montpellier).

A. Denise is a member of the *ComitÃ© National de la Recherche Scientique* (section 7, CID43). He is a member of the Scientific Committee of the "UFR des Sciences" at the University Versailles-St-Quentin-en-Yvelines and INRIA Saclay-Ile-de-France, as well as the Scientific Committee for Mathematics, Computational Biology and Artificial Intelligence at INRA. He serves in one ANR committee (ComitÃ© d'Evaluation et du ComitÃ© de Suivi du programme Masses de DonnÃ©es et Connaissances) and he is the co-chair of the Computation Committee for the INRA center of Jouy-en-Josas. He is a member of LRI laboratory council. He served in Evaluation committees of AERES for TIMC (U. Joseph Fourier and CNRS), LITIS (INSA de Rouen, UniversitÃ©s de Rouen et du Havre), SYSDIAG (CNRS Montpellier) and IML (UniversitÃ© de la MÃ©diterranÃ©e et CNRS). He was a member of recruitment committees (CNRS-Nice University chair, U. Paris-Sud).

Ch. Froidevaux is the head of Computer Science Department at University Paris-Sud XI (UFR des Sciences d'Orsay). She participated to the AERES committee that evaluated IRISA and to the national committee for PES attribution.

M. Régnier serves in the Committee of French ANR http://www.agence-nationale-recherche.fr/. She was a member of recruitment committees (Inria-Ups chair, Inria internal jury).

8.2. Teaching

The Master of Bioinformatics and Biostatistics of University Paris-Sud (http://www.bibs.u-psud.fr) is co-headed by members of the group. Since September 2010, it is a joint Master between University Paris-Sud and Ecole Polytechnique. Most members of the group teach in Master Bibs.

J. Bernauer teaches a L2 course on *Automata* at Marne-la-Vallee University.

Y. Ponty has taught a second year course on Programming languages at Ecole Polytechnique, and teaches two M2-level courses in RNA Bioinformatics/Algorithms at Paris-Sud/XI (Bibs Masters) and University Pierre et Marie Curie (BIM Masters).

M. Régnier deliver a 20 hours master course in bioinformatics at Al Farabi University (Almaty, Kazakhstan). She was the foreign co-advisor of A. Kabdullina's thesis (defence in June 2010) and currently is the foreign co-advisor of A. Isabekova's thesis. She serves in the Committee of French Agregation of Mathematics (Computer Science option).

C. Saule is a teaching assistant at Orsay UFR (*Internet programming, Engineering software, Data bases* and Java). He is also involved in tutoring.

Ph. Rinaudo teaches courses on *Basics in Computer Science* for biologists, *Arithmetics and complexity in Biology* and *Principles of Programming Languages*.

Van Du teaches courses on *Internet programming* and *Algorithm and Complexity* at University Paris-Sud.

A.Denise was a committee member for the HdR of StÃ©phane Vialette (Marne-la-VallÃ©e and for the thesis of Matthieu Josuat-VergÃ¨s (Orsay), Julien David (Marne-la-VallÃ©e), Magali Naville (Orsay). He was referee for the HdR of Cyril Nicaud (Marne-la-VallÃ©e) and for the PhD thesis of Jean-Philippe Doyon (Montreal). Ch. Froidevaux was a committee member for the HDR of Ch. Zimmer (Orsay) and Vincent Frouin (CEA Saclay), for the thesis of Abdeltif Elbyed(Evry), Mouna Essaba (Evry) and referee for Domitille Heitzler (Tours). M. Régnier was member of the committee for S. Carat (Nantes).

9. Bibliography

Major publications by the team in recent years

[1] Z. Bao, S. Cohen-Boulakia, S. Davidson, P. Girard. *PDiffView: Viewing the Difference in Provenance of Workflow Results*, in "PVLDB, Proc. of the 35th Int. Conf. on Very Large Data Bases", 2009, vol. 2, n^{o} 2, p. 1638-1641.

[2] Y. Ponty. *Efficient sampling of RNA secondary structures from the Boltzmann ensemble of low-energy: The boustrophedon method*, in "Journal of Mathematical Biology", Jan 2008, vol. 56, n^{o} 1-2, p. 107–127, http://www.lri.fr/~ponty/docs/Ponty-07-JMB-Boustrophedon.pdf.

Publications of the year

Doctoral Dissertations and Habilitation Theses

[3] C. Saule. *ModÃ¨les combinatoires des structures dâARN avec ou sans pseudonoeuds, application Ã la comparaison de structures*, Laboratoire de Recherche en Informatique (LRI) – Université Paris-XI/Paris Sud, December 2010.

Articles in International Peer-Reviewed Journal

[4] M. Behzadi, A. Demidem, D. Morvan, L. Schwartz, G. Stepien, J.-M. Steyaert. *A Model of Phospholipid Biosynthesis in Tumor in Response to an Anticancer Agent in Vivo*, in "Journal of Integrative Bioinformatics", 2010, vol. 7, n$^{\text{O}}$ 3, 129.

[5] G. Blin, A. Denise, S. Dulucq, C. Herrbach, H. Touzet. *Alignment of RNA structures*, in "IEEE/ACM Transactions on Computational Biology and Bioinformatics", 2010, vol. 7, n$^{\text{O}}$ 2, p. 309 - 322, to appear.

[6] A. Denise, Y. Ponty, M. Termier. *Controlled non uniform random generation of decomposable structures*, in "Journal of Theoretical Computer Science (TCS)", 2010, vol. 411, n$^{\text{O}}$ 40-42, p. 3527-3552 [*DOI :* 10.1016/j.tcs.2010.05.010], http://hal.inria.fr/hal-00483581/en.

[7] S. Flores, J. Bernauer, X. Huang, R. Zhou, S. Shin. *MULTI-RESOLUTION MODELING OF BIOLOGICAL MACROMOLECULES - Session Introduction.*, in "Pacific Symposium on Biocomputing", 2010, vol. 15, p. 201-204 [*DOI :* 10.1142/9789814295291_0022], http://eproceedings.worldscinet.com/9789814295291/preserved-docs/9789814295291_0022.pdf, http://hal.inria.fr/inria-00536410/en.

[8] C. Herrbach, A. Denise, S. Dulucq. *Average complexity of the Jiang-Wang-Zhang pairwise tree alignment algorithm and of a RNA secondary structure alignment algorithm*, in "Theoretical Computer Science", 2010, vol. 411, p. 2423-2432, http://hal.inria.fr/inria-00541269/en.

[9] S. Loriot, F. Cazals, J. Bernauer. *ESBTL: efficient PDB parser and data structure for the structural and geometric analysis of biological macromolecules.*, in "Bioinformatics", Apr 2010, vol. 26, n$^{\text{O}}$ 8, p. 1127-8 [*DOI :* 10.1093/bioinformatics/btq083], http://bioinformatics.oxfordjournals.org/cgi/reprint/26/8/1127.pdf, http://hal.inria.fr/inria-00536404/en.

[10] F. Lou, P. Clote. *Thermodynamics of RNA structures by Wang-Landau sampling*, in "Bioinformatics (ISMB 2010)", Jun 2010, vol. 26, n$^{\text{O}}$ 15, p. i278-i286, http://bioinformatics.oxfordjournals.org/cgi/reprint/btq218?ijkey=pLya7ztyQ7pWjb3&keytype=ref.

[11] T. Moncion, P. Amar, G. Hutzler. *Automatic characterization of emergent phenomena in complex systems*, in "Journal of Biological Physics and Chemistry", 2010, vol. 10, p. 16–23.

[12] S. Rialle, L. Felicori, C. Dias-Lopes, S. Peres, S. E. Atia, A. R. Thierry, P. Amar, F. Molina. *BioNetCAD: design, simulation and experimental validation of synthetic biochemical networks*, in "Bioinformatics", 2010, http://www.ncbi.nlm.nih.gov/pubmed/20628073.

[13] C. Saule, M. Regnier, J.-M. Steyaert, A. Denise. *Counting RNA pseudoknotted structures*, in "Journal of Computational Biology", 2010, page : to appear, http://hal.inria.fr/inria-00537117/en.

[14] M. Schmidt Am Busch, A. Sedano, T. Simonson. *Computational protein design: validation and possible relevance as a tool for homology searching and fold recognition*, in "PLoS One", 2010, vol. 5, n$^{\text{O}}$ 5, e10410.

[15] N. Seghezzi, P. Amar, B. Koebmann, P. R. Jensen, M.-J. Virolle. *The construction of a library of synthetic promoters revealed some specific features of strong Streptomyces promoters*, in "Applied Microbiology and Biotechnology", 2010, To appear in Applied Microbiology and Biotechnology.

[16] L. WEI, H. ZEXI, S. TAO, A. DENISE, F. XIANG-DONG, Z. YI. *Regulation of splicing enhancer activities by RNA secondary structures*, in "FEBS-Letters", 11 2010, vol. 584, p. 4401-4407 [*DOI :* 10.1016/J.FEBSLET.2010.09.039], http://hal.inria.fr/inria-00542625/en/.

International Peer-Reviewed Conference/Proceedings

[17] C. BANDERIER, P. NICODÈME. *Bounded Discrete Walks*, in "Proc. AofA 2010", DMTCS, 2010, p. 35–48, Vienna, June 2010.

[18] O. BODINI, Y. PONTY. *Multi-dimensional Boltzmann Sampling of Languages*, in "AOFA'10", Autriche Vienne, AM, 2010, p. 49–64, 12pp, http://hal.inria.fr/hal-00450763/en.

[19] S. COHEN-BOULAKIA, U. LESER. *Next Generation Data Integration for the Life Sciences*, in "IEEE International Conference on Data Engineering (ICDE)", Hannover Allemagne, 04 2011, http://hal.inria.fr/inria-00542359/en/.

[20] S. DAVIDSON, S. KHANNA, S. ROY, S. COHEN-BOULAKIA, Z. BAO, A. EYAL. *Privacy Issues in Scientific Workflow Provenance*, in "Proc. of the 1st Int. Workshop on Workflow Approaches to New Data-centric Science (SIGMOD Workshop)", Indianapolis États-Unis, 06 2010, http://hal.inria.fr/inria-00542339/en/.

[21] J. OUDINET, A. DENISE, M.-C. GAUDEL, R. LASSAIGNE, S. PEYRONNET. *Uniform Monte-Carlo Model-Checking*, in "Proc. FASE 2011: Fundamental Approaches to Software Engineering", Saarbrücken, 2011, to appear.

[22] Y. PONTY, J. WALDISPÜHL. *An unbiased adaptive sampling algorithm for the exploration of RNA mutational landscapes under evolutionary pressure*, in "Proceedings of RECOMB'11 RECOMB", Vancouver Canada, 2011, 0000, http://hal.archives-ouvertes.fr/hal-00546847/en/.

[23] VAN DU. TRAN, P. CHASSIGNET, J.-M. STEYAERT. *Prediction of permuted super-secondary structures in beta-barrel proteins*, in "SAC'11: ACM Symposium on Applied Computing Proceedings", 2011, to appear.

Workshops without Proceedings

[24] D. GARDY, Y. PONTY. *Weighted random generation of context-free languages: Analysis of collisions in random urn occupancy models*, in "GASCOM'10", Montréal Canada, LACIM, UQAM, 09 2010, G.: Mathematics of Computing/G.2: DISCRETE MATHEMATICS/G.2.1: Combinatorics/G.2.1.2: Generating functions, F.: Theory of Computation/F.2: ANALYSIS OF ALGORITHMS AND PROBLEM COMPLEXITY/F.2.2: Nonnumerical Algorithms and Problems/F.2.2.1: Computations on discrete structures, J.: Computer Applications/J.3: LIFE AND MEDICAL SCIENCES/J.3.0: Biology and genetics, http://hal.inria.fr/inria-00543150/en/.

[25] J. OUDINET, A. DENISE, M.-C. GAUDEL. *A new dichotomic algorithm for the uniform random generation of words in regular languages*, in "GASCom 2010", Montreal Canada, 2010, http://hal.inria.fr/inria-00542683/en/.

Scientific Books (or Scientific Book chapters)

[26] P. AMAR, F. KÃ©PÃ¨S, V. NORRIS. *Proceedings of the Evry Spring School on Modelling Complex Biological Systems in the context of genomics*, EDP Sciences, 2010.

References in notes

[27] R. Das, D. Baker. *Automated de novo prediction of native-like RNA tertiary structures.*, in "Proc Natl Acad Sci U S A", 2007, vol. 104, n^{o} 37, p. 14664-9, http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=17726102.

[28] Y. Ding, C. Chan, C. Lawrence. *RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble*, in "RNA", 2005, vol. 11, p. 1157–1166.

[29] T. C. J. Freeman, W. C. Wimley. *A highly accurate statistical approach for the prediction of transmembrane beta-barrels.*, in "Bioinformatics", 2010, vol. 26, n^{o} 16, p. 1965-74.

[30] M. F. Lensink, S. J. Wodak. *Docking and scoring protein interactions: CAPRI 2009.*, in "Proteins", 2010, http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=20806235.

[31] P. Minary, M. Levitt. *Conformational optimization with natural degrees of freedom: a novel stochastic chain closure algorithm.*, in "J Comput Biol", 2010, vol. 17, n^{o} 8, p. 993-1010, http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=20726792.

[32] M. . Parisien, F. Major. *The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data*, in "Nature", 2008, vol. 452, n^{o} 7183, p. 51–55, http://www.nature.com/nature/journal/v452/n7183/full/nature06684.html.

[33] D. W. Ritchie. *Recent progress and future directions in protein-protein docking.*, in "Curr Protein Pept Sci", 2008, vol. 9, n^{o} 1, p. 1-15, http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=18336319.

[34] C. M. Summa, M. Levitt. *Near-native structure refinement using in vacuo energy minimization.*, in "Proc Natl Acad Sci U S A", 2007, vol. 104, n^{o} 9, p. 3177-82, http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=17360625.

[35] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, H. J. Berendsen. *GROMACS: fast, flexible, and free.*, in "J Comput Chem", 2005, vol. 26, n^{o} 16, p. 1701-18, http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=16211538.