IN PARTNERSHIP WITH:
CNRS

Université Paris-Sud (Paris 11)

Activity Report 2011

Project-Team AMIB

Algorithms and Models for Integrative Biology

IN COLLABORATION WITH: Laboratoire d'informatique de l'école polytechnique (LIX)

RESEARCH CENTER
Saclay - Île-de-France

THEME
Computational Biology and Bioinformatics

Table of contents

Project-Team AMIB

Keywords: Protein Structure, Docking, RNA Annotation, Computational Biology, Machine Learning

1. Members

Research Scientists

Mireille Régnier [Team leader, Research Director (DR) Inria, HdR]
Julie Bernauer [Research Associate (CR) Inria]
Yann Ponty [Research Associate (CR) CNRS]

Faculty Members

Patrick Amar [Université Paris-Sud]
Jérôme Azé [Université Paris-Sud]
Sarah Cohen-Boulakia [Université Paris-Sud]
Alain Denise [Université Paris-Sud, HdR]
Christine Froidevaux [Université Paris-Sud, HdR]
Sabine Peres [Université Paris-Sud]
Jean-Marc Steyaert [Ecole Polytechnique, HdR]
Florence d'Alché-Buc [on leave from Université Evry, 50%, HdR]

PhD Students

Jiuqiang Chen [Université Paris-Sud]
Adrien Guilhot-Gaudeffroy [Université Paris -Sud XI, since 01/10/11]
Mahsa Behzadi [Ecole Polytechnique, until 30/04/11]
Daria Iakovishina [Ecole Polytechnique, since 01/11/11]
Feng Lou [Université Paris-Sud]
Philippe Rinaudo [Université Paris-Sud]
Cong Zeng [Université Paris-Sud]
Audrey Sedano [Ecole Polytechnique, 50%]
Thuong Van Du Tran [Ecole Polytechnique, until 30/09/11]
Bo Yang [Université Paris-Sud and Wuhan University]

Post-Doctoral Fellows

Saad Sheikh [until 31/08/11]
Loic Paulevé [Ecole Polytechnique, since 01/10/11]
Alexis Lamiable [Université Paris-Sud, since 01/09/11]

Administrative Assistant

Evelyne Rayssac [Secretary (SAR) Inria]

2. Overall Objectives

2.1. Introduction

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.

2.2. Highlights

In 2011, the VARNA software was highlighted in both the scientific reports of CNRS (*Rapport Scientifique 2010*) and Digiteo foundation. VARNA is currently used by RNA scientists (Cited by 44 research articles since its presentation in Fall of 2009), web servers 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/), SRNATARBASE (http://ccb.bmi.ac.cn/srnatarbase/) and the RFAM database (http://rfam.sanger.ac.uk/), the main source of sequence/structure data for RNA scientist, to display secondary structures. It is also used as an integrated component within JalView, arguably one of the leading sequence alignment editor (http://www.jalview.org/).

3. Scientific Foundations

3.1. RNA and protein structures

3.1.1. RNA

Participants: Julie Bernauer, Alain Denise, Feng Lou, Yann Ponty, Mireille Régnier, Philippe Rinaudo, Jean-Marc Steyaert.

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

3.1.1.1. From RNA structure to function

Recoding 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, and this process involves particular motifs and secondary structures in mRNAs. 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. We also study an other biological process that may involve particular motifs and structures in mRNAs: nonstop mRNA decay (NSD) and no-go mRNA (NGD) decay, that are recently identified mechanisms that control the quality of RNA transcription. This work is currently funded by the ANR (project NGD-NSD, ANR BLANC 2010-2014).

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. Beyond the secondary structure

One of our major challenges is to go beyond secondary 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 (ANR BLANC 2008-2012) in collaboration with PRISM at Versailles University and 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. 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.

3.1.2. PROTEINS

Participants: Jérôme Azé, Julie Bernauer, Adrien Guilhot-Gaudeffroy, Saad Sheikh, Jean-Marc Steyaert, Thuong Van Du Tran.

3.1.2.1. 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 [43]. The recent rounds of the CAPRI challenge show that real progress has been made using new techniques [40]. 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.

3.1.2.2. 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.

3.1.2.3. 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.

3.2. Annotation and Combinatorics

3.2.1. Word counting

Participants: Alain Denise, Daria Iakovishina, Mireille Régnier, 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 structural motifs. 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.

3.2.2. Random generation

Participants: Alain Denise, Yann Ponty.

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 [38]: 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 [3], [19], 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 [42].

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 [13], [29].

3.2.3. Knowledge extraction

Participants: Jérôme Azé, Jiuqiang Chen, Sarah Cohen-Boulakia, Christine Froidevaux.

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. For instance, a commonly accepted hypothesis is that function depends on the order of the motifs present in a genomic sequence. Likewise we must be able to evaluate the quality of the annotation obtained. This necessitates giving an estimate of the reliability of the results. This may use combinatorial tools described above. It 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.

3.2.4. Systems Biology

Participants: Patrick Amar, Mahsa Behzadi, Sarah Cohen-Boulakia, Christine Froidevaux, Loic Paulevé, Sabine Peres, Mireille Régnier, Jean-Marc Steyaert.

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).

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 Lab (UMR CNRS 3145) 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*.

We participate in a research project *eSignal* with INRA (ASAM, collaboration with INRA-BIOS laboratory) that aims at providing unique tools allowing to decipher and model the most proximal layer of biological systems: intracellular biochemical networks. More precisely we are interested in GPCRs (G protein-coupled receptors) trigger complex signalling networks that are involved in a wide array of physio-pathological processes. As such, GPCRs are targeted by almost half of the currently marketed drugs. As systems biology has developed experimental means to generate massive quantities of high quality data, there is a need for computational methods to integrate these data in predictive dynamic models. AMIB group aims at building an innovative pipeline of computational methods covering all the tasks needed to go from the initial data to predictive dynamic models of intracellular signalling mechanism.

A cooperation with an INSERM-INRA team based in Clermont-Ferrand addresses the behaviour of biological systems. A mathematical approach is currently being developed to study stability of some sub-domains, the importance of initial conditions that are to be inferred. This involves data analysis of experimental facts and a comparative analysis. Discrete approaches are relevant here, to cope with the combinatorial explosion of dynamics to explore, and analyze reachability properties within large networks. A software is developed to enhance the scalability of the parameters inference.

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, either as a bitmap (*JPEG, PNG*) or as a vector picture (*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 by 44 research articles since its presentation in Fall of 2009), web servers 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/), SRNATARBASE (http://ccb.bmi.ac.cn/srnatarbase/)and the RFAM database (http://rfam.sanger.ac.uk/), the main source of sequence/structure data for RNA scientist, to display secondary structures. It is also used as an integrated component within JALVIEW, arguably one of the leading sequence alignment editor (http://www.jalview.org/).

4.2. GeneValorization

Participant: Sarah Cohen-Boulakia [correspondant].

High-throughput technologies provide fundamental information concerning thousands of genes. Many of the current research laboratories daily use one or more of these technologies and end-up with lists of genes. Assessing the originality of the results obtained includes being aware of the number of publications available concerning individual or multiple genes and accessing information about these publications. Faced with the exponential growth of publications available and number of genes involved in a study, this task is becoming particularly difficult to achieve. We introduce GENEVALORIZATION, a web-based tool which gives a clear and handful overview of the bibliography available corresponding to the user input formed by (i) a gene list (expressed by gene names or ids from ENTREZGENE) and (ii) a context of study (expressed by keywords). From this input, GENEVALORIZATION provides a matrix containing the number of publications with co-occurrences of gene names and keywords. Graphics are automatically generated to assess the relative importance of genes within various contexts. Links to publications and other databases offering information on genes and keywords are also available. To illustrate how helpful GENEVALORIZATION is, we have considered the gene list of the OncotypeDX prognostic marker test. it is available at http://bioguide-project.net/gv.

4.3. 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 is an 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.

4.4. Cartaj

Participants: Alain Denise [correspondant], Alexis Lamiable.

CARTAJ is a software that automatically predicts the topological family of three-way junctions in RNA molecules, from their secondary structure only. The Cartaj software http://cartaj.lri.fr that implements our method can be used online. It is also meant for being part of RNA modelling softwares and platforms. The methodology and the results of CARTAJ are presented in [16].

5. New Results

5.1. RNA structures

5.1.1. RNA secondary structures: folding, design and evolution

In a collaboration with J. Waldispuhl (McGill, Canada) (Presented at the RECOMB'11 conference [32]), we used weighted grammatical models, introduced by members of the group [2], to perform an efficient exploration of the mutational landscape of RNA. We proposed an adaptive sampling algorithm, where weights were used to compensate an identified bias toward regions of higher GC-content within sampled sequences, thereby allowing for the exploration of more relevant portions of the evolutionary landscape. These adaptive sampling principles can be adapted into a method for the RNA design following similar principles. This constitutes a competitive alternative to local search strategies used by all existing tools for this problem. This work is ongoing as a collaboration with B. Berger group (MIT) and J. Waldispuhl (McGill, Canada), and a manuscript was recently submitted.

5.1.2. RNA knowledge-based potentials and 3D studies

We used the curated database of biologically interesting structures we have set up to perform a statistical analysis and developed knowledge-based potentials. The database server is available at http://csb.stanford.edu/rna.

We obtained RNA knowledge-based potentials that now performs well at different representation levels. They can be used in three well-known Molecular Dynamics (MD) and modeling software suites ENCAD [44], GROMACS (v3 and 4) [45] and MOSAICS [41] and are available for the community. The study we performed on a large number of new decoys showed that our potential outperforms Rosetta RNA scoring function [37] which is the gold standard. We show that not having correction terms for base-stacking and pairing can be of advantage when modelling loops at high resolution. The study was welcomed by the RNA community and published in *RNA Journal* (IF 6.5) [8].

We also refined the mixture model strategy we developed for building knowledge-based potentials. In collaboration with O.Schwander at LIX, we compared different mixture models: Dirichlet Process Mixture models (DPM), Kernel Density Estimation (KDE) models, Expectation Maximization models (MM) with different number of components (including a simplified version based on a post-processing step using K-Means). We showed that the Dirichlet Process Mixtures (DPM) is a good tradeoff despite its longer precomputation time as it provides a smooth potential having relatively few components. This study was presented at the MCMMB'11 conference and was submitted as a journal paper.

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

5.2. Proteins structures

5.2.1. Protein sequence alignment

In comparative protein modeling, the quality of a template model depends heavily on the quality of the initial alignment between a given protein with unknown structure to various template proteins, whose tertiary structure is available in the Protein Data Bank (PDB). Although pairwise sequence alignment has been solved for more than three decades, there remains a large discrepancy between the accuracy of the best sequence alignment between two amino acid sequences, as produced by the Needleman-Wunsch or Smith-Waterman algorithms, and that of the best structural alignment between two protein X-ray structures, as produced by the software DALI, CE, TOPOFIT, etc. To improve the quality of initial alignments in template modeling, one can integrate valuable information from an ensemble of generated suboptimal alignments, that is alignments whose score is below the best possible score. In a collaboration with P. Clote (Boston College/DIGITEO) [26], we presented a novel algorithm to produce suboptimal pairwise alignments.

5.2.2. 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 [9].

We have also proposed an approach that improves the predictions made by HEX, a state-of-the art docking tool developped by INRIA Nancy. We applied Voronoi fingerprint to the output of HEX and learn how to rank them, and we have tested new ranking strategies. The obtained ranking improve the initial ranking of HEX [33], [23].

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.

5.2.3. Transmembrane β-barrels:

We have recently proposed an algorithm [31] 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 [21], [5]. Especially, our PPV, MCC and F-Scores are second only to a very recent algorithm by Freeman and Wimley [39], 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.

In addition, we have developed consensus methods to combine multiple secondary structures into one more reliable solution. Our results show that our technique can be used to combine multiple solutions to produce structures that are more than any of the input structures. These methods are based mainly on social choice theory and known properties of TMB proteins. In addition, we are working on methods for combining information on the super-secondary structures, and using them to augment the supersecondary structure provided by our approach.

5.3. Combinatorics and Annotation

5.3.1. Word counting and random generation

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 long-term research on word enumeration has been realized by the team. An extension to Hidden Markov Models has been realized recently in a collaboration with M. Roytberg (IMPB, Puschino, Russia). It relies on a new concept of overlap graphs that efficiently overcomes the main difficulty - overlapping occurrences - in probabilities computation. This is part of E. Furletova's thesis, to be defended soon. An implementation is available at http://server2.lpm.org.ru/bio. This algorithm provides a significant space improvement over a previous algorithm, AHOPRO developed with our former associate team MIGEC. M. Régnier and S. Sheikh have addressed combinatorial problems on clumps that should allow further space decrease and large deviation results were presented at MCCMB'11.

An other application of word combinatorics has been started this year. During his internship, L. Pei (Paris-Sud 11 U.) provided a pipeline that simulates a random generation of reads and assembles them using MIRA software. This work will be pursued by D. Iakovishina in her thesis. It is a collaboration with MAGNOME at INRIA-BORDEAUX and IOGENE in Moscow.

A previous work [36], published in 2010, generalized Boltzmann samplers to multivariate objects, allowing for the efficient random generation achieving a fixed or approximate composition for context-free languages. However, the performances of such algorithms were only guaranteed in the case of strongly-connected context-free grammars. In a recent collaboration with O. Bodini, H. Tafat and C. Banderier (LIPN, Paris-XIII) we are working on characterizing the distributions arising from simply connected grammars. In a short paper accepted for presentation at the ANALCO'12 conference [24], we showed that: i) a large class of distributions can be reached for the number of occurrence of a single letter, arguably the simplest observable pattern; ii) simple grammars/regular expressions can be built that realize these distributions; iii) Classic Boltzmann samplers remain largely unaffected by this diversity.

Our work on random generation has applications in software testing and model-checking, in a collaboration with the Fortesse group at LRI [13], [29].

5.3.2. RNA combinatorics

Pseudoknots are usually ignored by popular software for RNA prediction. This means that, even under the daring assumptions of an unique and well-defined fold for RNA, coupled with a perfectly accurate energy model, the real structure of RNA will not be recovered perfectly. In a collaboration between AMIB members and S. Janssen (Universität Bielefeld), we investigated the practical implications of such a limitation. We used RNAFOLD, a popular software for the prediction of RNA structure on representative sequences of the RFAM database, which groups known RNA sequences into about 2000 functional families. We observed that 12% of RFAM families exhibited a total absence of overlap between predicted structures and manually-curated structures, derived from experimental or evolutionary data. Combination of RFAM annotations, a survey of literature, and a newly developed predictive method for the presence of a functional pseudoknots, we were able to validate that a large majority of the mispredicted families featured evidence of pseudoknots in the functional conformation. Preliminary results were presented by B. Raman at the Fifth Indo-French Bioinformatics Meeting [34].

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 [19], 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.

Similar decompositions can be used for the design of algorithms that include tractable subclasses of pseudoknots. In [30] Y. Ponty and C. Saule extended a unifying framework introduced by Roytberg and Finkelstein to design ensemble RNA algorithms. This framework uses a family of hypergraphs to describe the conformation space, allowing for a clear separation between the search space, i.e. the set of admissible conformations, and the intended application (Minimal Free-Energy folding, partition function, statistical sampling...). We illustrated the promises of such an approach by explicitly rephrasing three major search spaces within the framework, and introduced an algorithm for computing the moments of any additive feature in the Boltzmann distribution.

By comparing empirical observations with the expected behavior of a model, combinatorial methods can be used to identify an evolutionary pressure weighing on RNA. In a collaboration with P. Clote (Boston College/DIGITEO) [11], we used analytic combinatorics to study the expected distance between both ends of an RNA molecule, or $5'$-$3'$ distance. Postulating a Boltzmann distribution on all secondary structures, we showed that this parameter is bounded by a – typically small – constant value when the sequence length goes to the infinity. Computing this quantity on a database of experimentally-determined secondary structures, we observed that the $5'$-$3'$ distances take larger values than those predicted from the model. Furthermore, quite surprisingly, this quantity was shown to correlate positively with the length. We concluded by hypothesizing that the secondary structure of RNA may be under evolutionary pressure to fold in a modular way, creating independent domains on the exterior face.

5.3.3. 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 this domain, the contribution of AMIB in 2011 has been three folds: First, we have followed our collaboration with Ulf Leser (invited in the AMIB group at LRI during 6 months in 2010) and have worked on the review of 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 in [28].

Additionally, we have worked on a vision of what should be done by workflow systems to make it possible to search, adapt, and reuse scientific workflows, the complete state-of-the-art on this domain has been provided [12]. Second, in close collaboration with oncologists from the Institut Curie and the Children's Hospital of Philadelphia we have worked on the problem of ranking genes of interest associated to a given disease. The software GENEVALORIZATION has been designed and developed in this context and is able to provide a concise view of the literature available associated to a list of genes [10]. A second aspect of this research has been the design of a consensus ranking method, BioConsert, able to make the most (ie underline common points) of a set of established rankings [27]. This last point has been done in close collaboration with Sylvie Hamel invited professor in our group in 2010 (2 months). Third, we have presented a simple logical query language called RL for expressing different kinds of rules, especially well-suited to express association rules for transcriptomic data. In that context the challenge is to find out relationships between genes that reflect observations of how expression level of each gene affects those of others. The conjecture that association rules could be a model for the discovery of gene regulatory networks has already been partially validated. Nevertheless, several different kinds of rules between genes could be useful with respect to some biological objectives and we have designed a framework in which biologists may define their "own customized semantics" for rules with regard to their requirements. We have studied how the RL language behaves with respect to the well-known Armstrong's axioms [22]. The main contribution of this paper is to exhibit a restricted form of RL-queries, yet with a good expressive power, for which Armstrong's axioms are sound. From this result, this sublanguage turns out to have structural and computational properties which have been shown to be very useful in data mining, databases and formal concept analysis.

5.4. Systems Biology

In her thesis, M. Behzadi has developed a know-how on the behaviour of biological systems along a cooperation with an INSERM-INRA team based in Clermont-Ferrand. In the methodology that was developed, one computes the equations' parameters from the experimental data in systems that can be considered at equilibrium. It was proved mathematically that some sub-domains are intrinsically stable and that their behaviour is not much affected by the initial conditions [4] for phospholipids biosynthesis. A review for carbone toxicity can also be found in [15]. Software Analyser software (MPSA) are currently under development by L. Paulevé.

Elementary flux mode is a fundamental concept as well as a useful tool in metabolic pathway analysis. However, when the networks are complex, the determination of elementary flux modes leads to combinatorial explosion of their number which prevents from drawing simple conclusions from their analysis. To deal with this problem, a biclustering method has developed [18] based on the Agglomeration of Common Motifs (ACoM). It was applied to the central carbon metabolism in Bacillus subtilis and to the yeast mitochondrial energy metabolism. It helped to give biological meaning to the different elementary flux modes and to the relatedness between reactions.

Once molecules and complexes participating in the signalling network have been identified, the relations between them (enzymatic reactions, activations, inhibitions, etc.) have to be deduced from experimental and literature data to build the influence graph. Partners INRA-BIOS and AMIB have started the development of a knowledge-based method, which uses the solver SOLAR, developed by NII (Tokyo), that allows automating this data integration task. We have already formalized the knowledge necessary for inferring the signalling network triggered by the FSH receptor, one famous GPCR. Preliminary results of this project ASAM are very encouraging.

6. Partnerships and Cooperations

6.1. Regional Initiatives

6.1.1. Digiteo

Participants: Alain Denise, Daria Iakovishina, Feng Lou, Loic Paulevé, Mireille Régnier, 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.

6.2. National Initiatives

6.2.1. ANR

AMIS-ARN, ANR BLANC 2009-2012: *Graph Algorithms and Automatic Softwares for Interactive RNA Structure Modelling*. This project is being coordinated by AMIB. The two other ivolved groups are from PRISM (Versailles University) and E. Westhof's lab (Strasbourg University). 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. Implementation 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.

6.2.2. 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.).

6.3. International Initiatives

6.3.1. INRIA Associate Teams

6.3.1.1. GNAPI

Title: Geometric and knowledge-based analysis for Nucleic Acid and Protein dynamics and Interactions

INRIA principal investigator: JulieBernauer

International Partner:

Institution: Stanford University School of Medicine (United States)

Laboratory: Computational Structural Biology

Duration: 2009 - 2011

See also: http://www.lix.polytechnique.fr/~bernauer/EA_GNAPI/

Many biological processes of therapeutic interest, such as gene regulation, involve RNA molecules and their interactions with large protein assemblies. Recent high-throughput experiments have yielded insights into mechanisms of these processes but often structural models showing important structural features and interactions are lacking. Using 3D data available for proteins and RNA, we derived knowledge-based potentials to predict protein and nucleic-acid 3D structure. In combination with appropriate geometric representations, we obtained fast and accurate all-atom and coarse-grained predictions of biomolecular structures. We show that we can accurately build knowledge-based potentials from various all-atom and coarse-grained measures. Using this method and an encoding of multi-body contacts through arrangement of circles on a sphere, we obtained a reasonable model of protein structure. We also applied this strategy to assess RNA structures and showed that it is currently one of the best performing potentials for RNA structure evaluation. These results suggest that our knowledge-based models may also be suitable for the study of RNA dynamics and interactions.

6.3.2. Visits of International Scientists

6.3.2.1. Invited researchers (long stays)

Peter Clote

Subject: Digiteo chair

Institution: Boston College (United States of America)

6.3.2.2. Invited researchers (Short stays)

Artem Kasyanov, (IOGene, Moscow), 2 weeks;

Institution:IOGene (Russia (Russian Federation))

M. Levitt, 3 days

Institution: Stanford University (USA)

A. Sim(Stanford), 10 days;

Institution: Stanford University (USA)

6.3.2.3. Internship

Leonid Uroshlev

Subject: Study of reference states for the building of RNA knowledge-based potentials

Institution: Laboratoire Franco-Russe Poncelet (Russia (Russian Federation))

Angela Yen

Subject: A dynamic-programming extension of MC-Fold applicable to Boltzmann equilibrium applications.

Institution: MIT (United States of America)

Anindya Jyoti Roy

Subject: Development of new support vector machines techniques for the analysis of RNA motifs

Institution: IIT Kanpur

6.3.3. Participation In International Programs

Exists a long term collaboration between AMIB and IOGENE, previously NIIGENETIKA, through Liapunov Institute, former MIGEC associate team and Poncelet Institute.

7. Dissemination

7.1. Animation of the scientific community

7.1.1. French Community

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

All team is involved in GDR-BIM (Biology, Computer Science and Mathematics, http://www.gdr-bim.u-psud.fr/). A. Denise is a member of the Scientific Council. J. Azé is the webmaster. Ch. Froidevaux and S. Cohen-Boulakia participate to the subdomain *Knowledge Representation, Ontologies, Data Integration and Grids*.

A. Denise and M. Régnier participate to the subdomain Sequence Analysis and to COMATEGE subgroup of GDR-IM (Informatique Mathématique, http://www.gdr-im.fr/)

Many members participate to ALEA working group (http://algo.inria.fr/AofA/Alea/index.html.

7.1.2. Manifestations

Participants: Jérôme Azé, Yann Ponty, Mireille Régnier, Jean-Marc Steyaert.

Y. Ponty coorganized with E. Fusy (LIX) and G. Schaeffer (LIX) the ALEA'2011 workshop/CNRS spring school at the CIRM center (http://www.lix.polytechnique.fr/alea11).

As part of the "Chemistry Year 2011", AMIB animated the INRIA booth, both at the "Salon des Jeux et de la Culture Mathématique" http://www.cnrs.fr/insmi/spip.php?article323 in May at University UPMC, organized by CIJM http://www.cijm.org/, and at the yearly "Fete de la Science" http://www.fetedelascience-idf.fr/index.php?p=recherche-geographique&a=view&id=154 in Moulon (Saclay). The topic was to illustrate the principles underlying RNA folding algorithms through playing combinatorial games.

AMIB organized at Ecole Polytechnique, jointly with L. Schwartz (La Pitié-Salpétriere- CHU; now with Garches Hospital) a one day meeting *Models for Cell Metabolism.* Y. Ponty and C. Poignard (INRIA-MC2, Bordeaux) gave a talk.

7.1.3. Seminars

7.1.3.1. Amib seminars

M. Levitt (Stanford) gave a Digiteo seminar.

We received in our weekly seminar: J. Selbig (Potsdam U.), O. Lichtarge (Baylor College of Medicine), L. Breuza (SwissProt), A.-L. Thevenin (Tel Aviv U.), P. Clote (Boston College), T.-B. Ho (Japan Advanced Institute of Science and Technology), S. Hornus (INRIA, Nancy), S. Vagner (IGM-Villejuif), C. Medigue (LABGEM, Genoscope, Evry), H. Falentin (INRA-Rennes), L. Berti-Equille (IRD), P. Carbonell (Evry U.), M. Michaut (Toronto U.), O. Lespinet (IGM, Orsay).

7.1.3.2. Other seminars and invited talks

J. Bernauer and A. Sim presented Associate Team GNAPI work at the Berkeley-INRIA-Stanford Meeting in Berkeley (May). J. Bernauer organized the discussion session of the PSB MSMB session in collaboration with S.Flores, X.Huang, S.Shin and R.Zhou.

A. Denise gave invited talks at the Workshop "Optimization and Machine Learning: Theory, Algorithms and Applications" at Metz, and at the conference "Les 20 ans du LABRI" at Bordeaux.

Y. Ponty gave an invited talk at the Dagstuhl seminar *Combinatorial and Algorithmic Aspects of Sequence Processing* (Germany).

S. Sheikh gave seminars at University of Central Florida (Orlando, Gainesville) and Florida International University (Miami).

In a regular partnership with Garches Hospital, AMIB members give seminars for medical staff.

7.1.4. 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", Sophia-Antipolis, May 2011. (http://epigenomique.free.fr/en).

J. Bernauer in collaboration with X. Huang (HKUST), S. Flores (Uppsala Univ.), S. Shin (Seoul National Univ.) and R.Zhou (IBM Watson Research Center) organized the "Multi-scale Modelling of Biosystems: from Molecular to Mesoscale" session of the Pacific Symposium on Biocomputing (Jan 3-7, 2011). See http://psb.stanford.edu/psb11/cfp_msmb.html.

S. Cohen-Boulakia, Ch. Froidevaux (co-head) and Y. Ponty served as members of the program committee for the JOBIM'11 bioinformatics conference (Institut Pasteur). (http://www.pasteur.fr/ip/easysite/pasteur/fr/recherche/communication-scientifique/conferences-et-congres-scientifiques/conferences-service-colloques-institut-pasteur/jobim-2011).

Ch. Froidevaux and J. Azé served as members of the program committee for the EGC'2011. Jerome Azé served as PC member of the international conferences and workshops QIMIE2011 (workshop of PAKDD2011) and DMIN2011. He co-chaired the national workshop QDC2011 (workshop of EGC2011) and was co-editor of a special number of the journal RNTI QDC-EVALECD'11.

S. Cohen-Boulakia is member of the editorial board of the Journal on Data Semantics (Springer) and was PC member of the following international conferences: VLDB2011, ICDE 2011 and SSDBM2011.

A. Denise is a member of the editorial committee of Techniques et Sciences Informatiques (Hermès).

M. Régnier organized with V. Makeev (IOGENE) and M. Gelfand (RAS) the 5th conference MCCMB'11 in Moscow. It was supported by a trilateral fund (France-Germany-Russia) and a Russia-India fund. She was a member of CPM'11 program committee.

A special session *RNA structure: from genomes to nanotechnology* was organized http://www.iscb.org/ismbeccb2011-program/895-special-session-details at ISMB'11 (Vienna, Austria) with a DIGITEO support.

7.1.5. Research Administration

Ch. Froidevaux took part as an external jury member in the hiring committee of a Professor position at INSA-Lyon, and as a member in the hiring committee of CR2 et CR1 positions at INRIA Bordeaux. Ch. Froidevaux served in the DIM (Ile-de-France) Committee. She is a Deputy member of the Scientific Committee of the Bioinformatics Center of the Geneva University. Ch. Froidevaux is head of the Computer Science Department at the University Paris Sud. This department involved research activities in two laboratories: LRI and LIMSI/CHM.

A. Denise serves in the *Comité National de la Recherche Scientifique*, section 7 and CID 43. He is copresident of the *Commission Informatique de Centre (CIC)* of the INRA research center at Jouy-en-Josas. He is member of the scientific committee of the Faculty of Sciences of University of Versailles, and of the scientific commission of the INRIA research center at Saclay. He serves as an expert for expert the *Direction Générale de la Recherche et l'Innovation (DGRI)* of the French Research Ministry. He has been member of the AERES evaluation committees of the I3S research unit (University of Nice Sophia Antipolis/CNRS) and of the INRIA research center at Rennes. He took part as an external jury member in the hiring committee of an Assistant position at the university of Nantes.

Y. Ponty took part as an external jury member in the hiring committee of an Assistant position at IUT Vélizy/Université de Versailles St Quentin. He participated in the LIX/QUALCOMM 2011 postdoc hiring committee.

M. Régnier serves in the Committee of French ANR http://www.agence-nationale-recherche.fr/, and, as a deputy member, in DIGITEO Program Committee.

7.2. Teaching

The Master of Bioinformatics and Biostatistics, which is a joint master between University Paris-Sud and Ecole Polytechnique (http://www.bibs.u-psud.fr), is co-headed by members of the group.

Software Engineering for Bioinformatics, 48h, M2 BIBS (Bioinformatics and BioStatistics), Paris-Sud University/École Polytechnique, France (P. Amar)

Modelling and Simulation of Biological Processes, 24h, M2 BIBS (Bioinformatics and BioStatistics), Paris-Sud University/École Polytechnique, France (P. Amar)

Biological Networks and Systems Biology, 9h, M1 BIBS (Bioinformatics and BioStatistics), Paris-Sud University/École Polytechnique, France (P. Amar)

Programmation Python, 20h, M2 BIBS (Bioinformatics and BioStatistics), Paris-Sud University/École Polytechnique, France (J. Bernauer)

RNAomics and RNA Bioinformatics, 12h, M2 BIBS (Bioinformatics and BioStatistics), Paris-Sud University/École Polytechnique, France (A. Denise)

Theoretical Computer Science, 28h, M2 BIBS (Bioinformatics and BioStatistics), Paris-Sud University/École Polytechnique, France (A. Denise)

Algorithmics and Programming in C, 48h, M1 BIBS (Bioinformatics and BioStatistics), Paris-Sud University/École Polytechnique, France (A. Denise)

Biological Networks and Systems Biology, 6h, M1 BIBS (Bioinformatics and BioStatistics), Paris-Sud University/École Polytechnique, France (A. Denise)

Integration and Analysis of heterogeneous data from the Web, 24h, M2 BIBS (Bioinformatics and BioStatistics), Paris-Sud University, France (J. Aze, S. Cohen Boulakia, Ch. Froidevaux)

Advanced Data Bases and Data Mining, 42h, M2 BIBS (Bioinformatics and BioStatistics), Paris-Sud University, France (S. Cohen Boulakia, Ch. Froidevaux).

Combinatorics, Algorithms, Structure and Models; 28 hours, M2 BIBS (Bioinformatics and BioStatistics), Paris-Sud University, France (Y. Ponty, M. Regnier, J.-M. Steyaert).

Advanced Algorithms and Optimization; 28 hours, M2 BIBS (Bioinformatics and BioStatistics), Paris-Sud University, France (J. M. Steyaert).

Introduction à l'informatique, 40h, L3, École Polytechnique, France. (Y. Ponty)

Modélisation et bioinformatique de l'ARN, 16h, M2 BioInformatique et Modélisation (BIM), UPMC, France. (Y. Ponty)

PhD students and Post-Doctoral fellows gave the following courses.

Langages et Programmation, 40 hours, L1, Ecole Polytechnique, France (S. Sheikh).

M. Régnier gave a 5 hours course "Mots et motifs: combinatoire et asymptotique" at CIRM meeting "Algorithmique et programmation" for CPGE teachers http://www.cirm.univ-mrs.fr/Site_test/?rubrique2&EX=info_rencontre&annee=2011&id_renc=628. She serves in the committee of French Agregation of Mathematics (Computer Science option).

Mahsa Behzadi, A Mathematical Model of Phospholipid Biosynthesis, Ecole Polytechnique, 12/07/2011, Jean-Marc Steyaert

Van Du Tran Thuong, Modeling and Predicting Super-secondary Structures of Transmembrane Beta-barrel Proteins, Ecole Polytechnique, 10/12/2011, Jean-Marc Steyaert.

PhD in progress : Adrien Guilhot-Gaudeffroy, Modeling and scoring of protein-RNA complexes, 01/10/2011, Ch. Froidevaux, J. Azé, J. Bernauer.

PhD in progress: Feng Lou, Etude et conception d'algorithmes de prédiction de la structure secondaire des molécules d'ARN, 01/10/2008, Alain Denise and Peter Clote.

PhD in progress: Daria Iakovishina, A Combinatorial Approach to Assembly Algorithms, 01/11/2011, M. Régnier.

PhD in progress: Jiuqiang Chen, Mining and Integration of heterogeneous data in e-science environments, 01/10/2011, S. Cohen-Boulakia and Ch. Froidevaux.

PhD in progress :Philippe Rinaudo, RNA sequence-structure alignements: a parametrized complexity approach, 01/10/2009, A. Denise.

PhD in progress : Cong Zeng, Identification of Structural Motifs in Messenger RNAs, 01/10/2011, A. Denise.

PhD in progress :Bo Yang, RNA tertiary structure prediction, 01/10/2011, A. Denise.

PhD in progress: Evgenia Furletova, Calculation of statistical significance of biological sequences, 01/12/2008, M. Roytberg (Moscow) and M. Régnier.

A. Denise served as a referee for the PhD dissertations of Florian Sikora (Marne la Vallée), José António Almeida Costa da Cruz (Strasbourg), Emmanuel Bénard (Strasbourg), Anne-Laure Gaillard (Bordeaux) and for Cédric Chauve's *Habilitation à Diriger des Recherches* (Bordeaux 2011). He served as a jury member in the PhD defence of Loïc Magnin's (Paris-Sud), Wael Khemiri (Paris-Sud), Tom Dreyfus (Sophia Antipolis) and in Marie-Hélène Mucchielli's *Habilitation à Diriger des Recherches* (Paris-Sud).

Ch. Froidevaux was a referee for D. Heitzler's PhD (Univ. Tours), S. Benabderrahmane's PhD (LORIA, Nancy) and J Wollbrett's PhD (CIRAD, Montpellier) and served as a jury member in F. Hamdi's PhD defense (Univ. Paris Sud).

Y. Ponty served as a jury member in A. Saffarian's PhD defence at Université Lille 1.

M. Régnier served as a jury member in M. Behzadi's and Van-Du Thuong Tran PhD defence (Ecole Polytechnique) and was a referee for C. Loi's PhD (Ecole Centrale).

8. 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] 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^{o} 40-42, p. 3527-3552 [*DOI :* 10.1016/j.tcs.2010.05.010], http://hal.inria.fr/hal-00483581/en.

[3] 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

[4] M. Behzadi. *Un modèle mathématique de la biosynthèse des phospholipides*, Ecole Polytechnique X, July 2011, http://hal.inria.fr/tel-00650399/en.

[5] T. V. D. Tran. *Modeling and predicting super-secondary structures of transmembrane beta-barrel proteins*, Ecole Polytechnique X, December 2011, http://hal.inria.fr/tel-00647947/en.

Articles in International Peer-Reviewed Journal

[6] J. Allali, C. Saule, Y. d'Aubenton-Carafa, A. Denise, C. Drevet, P. Ferraro, D. Gautheret, C. Herrbach, F. Leclerc, A. De Monte, A. Ouangraoua, M.-F. Sagot, M. Termier, C. Thermes, H. Touzet. *BRASERO: A resource for benchmarking RNA secondary structure comparison algorithms*, in "Advanced in Bioinformatics", 2012, page : to appear, http://hal.inria.fr/hal-00647725/en.

[7] J. Bernauer, S. Flores, X. Huang, S. Shin, R. Zhou. *MULTI-SCALE MODELLING OF BIOSYSTEMS: FROM MOLECULAR TO MESOCALE - Session Introduction.*, in "Pacific Symposium on Biocomputing", 2011, p. 177-80 [*DOI :* 10.1142/9789814335058_0019], http://hal.inria.fr/inria-00542791/en.

[8] J. Bernauer, X. Huang, A. Y. L. Sim, M. Levitt. *Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation.*, in "RNA", June 2011, vol. 17, n^{o} 6, p. 1066-75 [*DOI :* 10.1261/rna.2543711], http://hal.inria.fr/inria-00624999/en.

[9] T. Bourquard, J. Bernauer, J. Azé, A. Poupon. *A collaborative filtering approach for protein-protein docking scoring functions.*, in "PLoS ONE", 2011, vol. 6, n^{o} 4 [*DOI :* 10.1371/journal.pone.0018541], http://hal.inria.fr/inria-00625000/en.

[10] B. Brancotte, A. Biton, I. Bernard-Pierrot, F. Radvanyi, F. Reyal, S. Cohen-Boulakia. *Gene List significance at-a-glance with GeneValorization.*, in "Bioinformatics", April 2011, vol. 27, n^{o} 8, p. 1187-9 [*DOI :* 10.1093/bioinformatics/btr073], http://hal.inria.fr/inria-00627865/en.

[11] P. Clote, Y. Ponty, J.-M. Steyaert. *Expected distance between terminal nucleotides of RNA secondary structures*, in "Journal of Mathematical Biology", 2011, 18, http://hal.inria.fr/inria-00619921/en.

[12] S. Cohen-Boulakia, U. Leser. *Search, adapt, and reuse: the future of scientific workflows*, in "Sigmod Record", 2011 [*DOI :* 10.1145/2034863.2034865], http://hal.inria.fr/inria-00638043/en.

[13] A. Denise, M.-C. Gaudel, S.-D. Gouraud, R. Lassaigne, J. Oudinet, S. Peyronnet. *Coverage-biased random exploration of large models and application to testing*, in "Software Tools for Technology Transfer (STTT)", 2011, http://hal.inria.fr/inria-00560621/en.

[14] S. J. Fleishman, T. A. Whitehead, E.-M. Strauch, J. E. Corn, S. Qin, H.-X. Zhou, J. C. Mitchell, O. N. A. Demerdash, M. Takeda-Shitaka, G. Terashi, I. H. Moal, X. Li, P. A. Bates, M. Zacharias, H. Park, J.-S. Ko, H. Lee, C. Seok, T. Bourquard, J. Bernauer, A. Poupon, J. Azé, S. Soner, S. K. Ovali, P. Ozbek, N. B. Tal, T. Haliloglu, H. Hwang, T. Vreven, B. G. Pierce, Z. Weng, L. Pérez-Cano, C. Pons, J. Fernández-Recio, F. Jiang, F. Yang, X. Gong, L. Cao, X. Xu, B. Liu, P. Wang, C. Li, C. Wang, C. H. Robert, M. Guharoy, S. Liu, Y. Huang, L. Li, D. Guo, Y. Chen, Y. Xiao, N. London, Z. Itzhaki, O. Schueler-Furman, Y. Inbar, V. Patapov, M. Cohen, G. Schreiber, Y. Tsuchiya, E. Kanamori, D. M. Standley, H. Nakamura, K. Kinoshita, C. M. Driggers, R. G. Hall, J. L. Morgan, V. L. Hsu, J. Zhan, Y. Yang, Y. Zhou, P. L. Kastritis, A. M. J. J. Bonvin, W. Zhang, C. J. Camacho, K. P. Kilambi, A. Sircar, J. J. Gray, M. Ohue, N. Uchikoga, Y. Matsuzaki, T. Ishida, Y. Akiyama, R. Khashan, S. Bush, D. Fouches, A. Tropsha, J. Esquivel-Rodríguez, D. Kihara, P. B. Stranges, R. Jacak, B. Kuhlman, S.-Y. Huang, X. Zou, S. J. Wodak, J. Janin, D. Baker. *Community-Wide Assessment of Protein-Interface Modeling Suggests Improvements to Design Methodology.*, in "Journal of Molecular Biology", September 2011, in press [*DOI :* 10.1016/j.jmb.2011.09.031], http://hal.inria.fr/inria-00637848/en.

[15] A. Guais, G. Brand, L. Jacquot, M. Karrer, S. Dukan, G. Grévillot, T. J. Molina, J. Bonte, M. Regnier, L. Schwartz. *Toxicity of Carbon Dioxide: A Review.*, in "Chemical Research in Toxicology", July 2011, epub ahead of print [*DOI :* 10.1021/tx200220r], http://hal.inria.fr/hal-00641044/en.

[16] A. Lamiable, D. Barth, A. Denise, F. Quessette, S. Vial, E. Westhof. *Automated prediction of three-way junction topological families in RNA secondary structures*, in "Computational Biology and Chemistry", November 2011, http://hal.inria.fr/hal-00641738/en.

[17] V. Norris, A. Zemirline, P. Amar, J. N. Audinot, P. Ballet, E. Ben-Jacob, G. Bernot, G. Beslon, A. Cabin, E. Fanchon, J.-L. Giavitto, N. Glade, P. Greussay, Y. Grondin, J. A. Foster, G. Hutzler, J. Jost, F. Kepes, O. Michel, F. Molina, J. Signorini, P. Stano, A. R. Thierry. *Computing with bacterial constituents, cells and populations: from bioputing to bactoputing*, in "Theorie in den Biowissenschaften / Theory in Biosciences", September 2011, vol. 130, n^{o} 3, p. 211-228 [*DOI :* 10.1007/s12064-010-0118-4], http://hal.inria.fr/hal-00643738/en.

[18] S. Pérès, F. Vallée, M. Beurton-Aimar, J.-P. Mazat. *ACoM: A classification method for elementary flux modes based on motif finding*, in "BioSystems", 2011, vol. 103, n^{o} 3, p. 410-419, http://hal.inria.fr/hal-00642137/en.

[19] C. SAULE, M. REGNIER, J.-M. STEYAERT, A. DENISE. *Counting RNA pseudoknotted structures*, in "Journal of Computational Biology", October 2011, vol. 18, n^{o} 10, p. 1339-1351 [*DOI :* 10.1089/CMB.2010.0086], http://hal.inria.fr/inria-00537117/en.

[20] 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", April 2011, vol. 90, n^{o} 2, p. 615-23 [*DOI :* 10.1007/S00253-010-3018-0], http://hal.inria.fr/hal-00643741/en.

[21] S. SHEIKH, P. CHASSIGNET, J.-M. STEYAERT, T. V. D. TRAN. *A graph-theoretic approach for classification and structure prediction of transmembrane beta-barrel proteins*, in "BMC Genomics", 2012, http://hal.inria.fr/hal-00650429/en.

International Conferences with Proceedings

[22] M. AGIER, C. FROIDEVAUX, J.-M. PETIT, Y. RENAUD, J. WIJSEN. *On Armstrong-compliant Logical Query Languages*, in "4th International Workshop on Logic in Databases, (EDBT/ICDT '10 joint conference)", Uppsala, Sweden, G. H. L. FLETCHER, S. STAWORKO (editors), ACM, 2011, p. 33-40, http://hal.inria.fr/hal-00649604/en.

[23] J. AZÉ, T. BOURQUARD, S. HAMEL, A. POUPON, D. RITCHIE. *Using Kendall-Tau Meta-Bagging to Improve Protein-Protein Docking Predictions*, in "PRIB 2011", DELFT, Netherlands, M. LOOG, ET AL. (editors), Marcel Reinders and Dick de Ridder, 2011, p. 284-295, http://hal.inria.fr/inria-00628038/en.

[24] C. BANDERIER, O. BODINI, Y. PONTY, H. TAFAT. *On the diversity of pattern distributions in combinatorial systems*, in "ANALCO'12", Japan, 2012, http://hal.inria.fr/hal-00643598/en.

[25] C. BANDERIER, P. NICODEME. *Constant time estimation of ranking statistics by analytic combinatorics*, in "Statistical Methods for Post-Genomic Data", Paris, France, January 2011, http://hal.inria.fr/hal-00567091/en.

[26] P. CLOTE, L. FENG, A. DENISE. *A new approach to suboptimal pairwise sequence alignment*, in "CompBio 2011: IASTED International Conference on Computational Bioscience", Cambridge, United Kingdom, 2011, http://hal.inria.fr/inria-00594890/en.

[27] S. COHEN-BOULAKIA, A. DENISE, S. HAMEL. *Using medians to generate consensus rankings for biological data*, in "SSDBM 2011: Scientific and Statistical Database Management Conference", Portland, United States, 2011, http://hal.inria.fr/inria-00584690/en.

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

[29] J. OUDINET, A. DENISE, M.-C. GAUDEL, R. LASSAIGNE, S. PEYRONNET. *Uniform Monte-Carlo Model Checking*, in "FASE 2011", Saarbrücken, Germany, 2011, http://hal.inria.fr/hal-00644834/en.

[30] Y. PONTY, C. SAULE. *A Combinatorial Framework for Designing (Pseudoknotted) RNA Algorithms*, in "11th Workshop on Algorithms in Bioinformatics (WABI'11)", Saarbrucken, Germany, 2011, http://hal.inria.fr/inria-00601060/en.

[31] T. V. D. Tran, P. Chassignet, S. Sheikh, J.-M. Steyaert. *Energy-based Classification and Structure Prediction of Transmembrane Beta-Barrel Proteins*, in "1st IEEE International Conference on Computational Advances in Bio and medical Sciences (ICCABS)", Orlando, FL, United States, February 2011, http://hal.inria.fr/inria-00562699/en.

[32] J. Waldispühl, Y. Ponty. *An unbiased adaptive sampling algorithm for the exploration of RNA mutational landscapes under evolutionary pressure*, in "RECOMB", Vancouver, Canada, V. Bafna, S. Sahinalp (editors), Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2011, vol. 6577, p. 501-515 [*DOI :* 10.1007/978-3-642-20036-6_45], http://hal.inria.fr/hal-00546847/en.

National Conferences with Proceeding

[33] T. Bourquard, J. Azé, A. Poupon, D. Ritchie. *Protein-protein docking based on shape complementarity and Voronoi fingerprint*, in "Journées Ouvertes Biologie Informatique Mathématiques", Paris, France, E. Barillot, C. Froidevaux, Eduardo PC. Rocha (editors), Institut Pasteur, July 2011, p. 9-16, http://hal.inria.fr/inria-00613186/en.

Other Publications

[34] S. Janssen, Y. Ponty, B. Raman, S. Sheikh, J.-M. Steyaert, P. Clote. *Investigating the RFAM paradox: The pseudoknot explanation*, 2011, Short abstract, http://hal.inria.fr/hal-00585647/en.

[35] A. Lorenz, Y. Ponty. *Non-redundant random generation algorithms for weighted context-free languages*, http://hal.inria.fr/inria-00607745/en.

References in notes

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.