INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Team AMIB

Algorithms and Models for Integrative Biology

Saclay - Île-de-France

Theme : Computational Biology and Bioinformatics

Table of contents

AMIB *is a joined team with* LIX*, Ecole Polytechnique and* LRI*, Paris-Sud XI University. The team has been created on May the 1st, 2009 and is under evaluation.*

1. Team

Research Scientist

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

Faculty Member

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 Student

Zahira Aslaoui [Université Paris -Sud XI, since 01/10/09]
Thomas Bourquard [Université Paris -Sud XI]
Mahassine Djelloul [Université Paris -Sud XI]
Feng Lou [Université Paris -Sud XI]
Bastien Rance [Université Paris -Sud XI, until 30/09/09]
Philippe Rinaudo [Université Paris -Sud XI, since 01/10/09]
Cédric Saule [Université Paris -Sud XI]
Thuong Van Du Tran [Ecole Polytechnique]

Post-Doctoral Fellow

Balaji Raman [Ecole Polytechnique]
Thomas Moncion [Université Paris -Sud XI]

Visiting Scientist

Julie Bernauer [ABS-Sophia, since September 1st]

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. 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.

Most problems in computational biology are NP-hard as soon as all known and *reasonable* biological information is taken into account. For instance, structural biology is concerned with 3D structures of complex molecules. Prediction, comparison and design are, in fact, three optimisation problems where these structures are classically represented by graphs and they are known to be NP-complete. A fruitful strategy consists in designing models that maintain the biological relevance while being simple enough to be computationnally tractable. The representation chosen determines the data structures and algorithms classes to be used. The challenge is to develop formal models, along with efficient algorithms, or heuristics, to deal with them. The various biological problems described above raise different computer science issues. To tackle them, the project members rely on a common methodology for which our group has a significant experience. Indeed, many of them can be expressed with classical combinatorial objects such as graphs, trees, words and grammars.

3.1.2. RNA

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

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

3.1.2.1. 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 is funded by the ANR (project RNA-RECOD, ANR BLANC 2006-2010). 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*.

It has also 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.

Another mid-term challenge is the design of molecules that enhance or repress such events.

3.1.2.2. Structural tertiary motifs

Single strand RNA folds to a stable and compact structure. This folding leads to a secondary structure that is an intermediate structure level for RNA, between the single sequence and the full structure (tertiary structure). It is based on pairing between complementary bases (A-U and C-G). A recent classification , the *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. Other kinds of interactions are also taken into account, such as *stacking*, and phosphodiester bonds along the sequence. 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. Indeed, it was shown that taking these motifs into account can lead to improve significantly the 3D prediction methods. We develop graph algorithms for extracting tertiary motifs from RNA structures, and for predicting the tertiary structure from the sequence [2]. This project, in collaboration with two groups from University of Strasbourg and University of Versailles, is funded by the ANR (project AMIS-ARN, ANR BLANC 2009-2012).

3.1.3. PROTEINS

Participants: Jérôme Azé, Julie Bernauer, Thomas Bourquard, Thomas Simonson, Thuong Van Du Tran.

3.1.3.1. Docking and evolutionary algorithms

The function of many proteins depends on their interaction with one or many partners. 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. 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.

3.1.3.2. 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. A second problem of interest is *inverse protein folding* or *computational protein design* (CPD), that is the prediction of amino-acid sequences that adopt a particular target tertiary structure. 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.

Molecular dynamics (MD) simulations use numerical methods to study the motion of atoms, by far too complex for analytical studies. They were used by BIOC for extensive computational engineering of aaRS, aminoacyl-tRNA synthetases. For computational protein design, and structure prediction as well, a possible modelling considers the protein *backbone* and *sidechains*. This backbone structure may be known by high-resolution methods. High-quality models for sidechains interactions with solvent have been designed. There is a finite number of possible positions for sidechains, that may be memorized in a *rotamer* library. A fitness or *energy* function that relies on atomistic and physical-chemical criteria is associated to each conformation. Therefore, one may search the set of possible sequences to optimize stability criteria.

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.

3.1.3.3. 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. Combinatorics and Enumeration

Participants: Alain Denise, Pierre Nicodème, Yann Ponty, Mireille Régnier, Cédric Saule, 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, and consensus models of recoding events. 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.

A starting project considers an algorithm of desambiguisation of automata, that uses the powerful techniques developed by Cyril Nicaud (IGM-Marne-la-Vallée University) to generate random automata; An other appealing problem is the random walk problem, considered as a modelization of ranked genes expression that could be used for medical diagnosis. In the mathematical setting, we want to know the probability that a random bridge of length n with increments $X_i = (+d, -c)$ exits of a strip $-H \leq y \leq H$. The increments have expectation zero and it is possible to assume that they are independent, later on conditionning the walk to come back to zero at time n. If the increments X_n are bounded, the limit of the walk as n tends to infinity is a Brownian bridge, the statistics of which is well known; however, practically, on one hand the value of d may be large, and on the other we are in the range of large deviations for small p-values. For these reasons, it is necessary to consider the discrete case. Banderier and Flajolet provided in 2002 a large account on discrete random walks, although they do not consider the heights of the walks. A collaboration has begun with Cyril Banderier (LIPN, University Paris-North) on the subject; Nicolas Broutin (INRIA-ALGORITHMS) and Thomas Feierl (joining INRIA-ALGORITHMS on Dec. 1st) should join this collaboration. The bioinformatics aspects will be considered by Marcel Shulz (Max-Planck Institut Berlin-Dahlem).

Analytical methods 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. For these more complex models, an experimental approach (*i.e.* a computational generation of random sequences) is still necessary. 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.

3.2.1. Knowledge extraction

Participants: Jéroôme Azé, Sarah Cohen-Boulakia, Christine Froidevaux, Bastien Rance, 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.

4. Software

4.1. VARNA

Participants: Yann Ponty [correspondant], Alain Denise.

VARNA [5] is a new 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 drawing using either an interactive point and click approach, within a web server or through command-line arguments.

In November 2009, VARNA is currently used by RNA scientists and websites such as the NESTEDALIGN web server (http://nestedalign.lri.fr/), the IRESITE database (http://iresite.org/), and the TFOLD webserver (http://tfold.ibisc.univ-evry.fr/TFold/). It is a free software, released under the terms of the GPLv3.0 license and available at http://varna.lri.fr.

4.2. SeSiMcMc

Participants: Mireille Régnier [correspondant], Vsevolod Makeev [Associate Team MIGEC], Ivan Kulakovsky [Associate Team MIGEC].

This software, freely available at http://favorov.imb.ac.ru/SeSiMCMC/ is designed to extract motifs and assess their relevance. This assessment relies on a pvalue computation, realized by AHOPRO (2007). OVGRAPH [22] improvement over AHOPRO should be introduced into SESIMCMC this year. One will use AHOPROMPV for separating the noise and the signal in ChIP data. An extension that takes into account ChipSeq data, ChipMunk is currently being designed.

5. New Results

5.1. RNA structures

5.1.1. Counting pseudoknots

In a recent work published in 2004, Condon analyzed 5 recent algorithms that predict secondary structures with pseudoknots. Relying on rewriting rules, she characterized the classes of pseudoknots that may be predicted. A collaborative work [15], [25] between LIX and LRI provides an alternative combinatorial characterization by graphs, from which enumeration follows, and, additionnally, studies a new class. In the long term, one expects to add biological constraints to these combinatorial definitions.

5.1.2. RNA fold and Rfam accuracy

Canonical secondary structures of RNA are those without lonely base pairs. Secondary structure prediction algorithms such as RNAFOLD, etc., claim to have greater accuracy in folding structures without lonely base pairs than with isolated pairs. B. Raman and P. Clote, (Relative Accuracy of RNAfold to Rfam Consensus for Canonical Secondary Structures), validate this claim: RNAFOLD improves accuracy in canonical structures prediction. This is assessed by extensive experiments using RNA sequences obtained from RNA database RFAM. The accuracy of the RNAFOLD algorithm is evaluated with respect to the consensus secondary structure of each and every RNA family in the RFAM database. This paper also points out that for certain families in the RFAM database the consensus secondary structure is inaccurate.

5.1.3. Riboswitches

Towards predicting the structure of a riboswitch, the first step is to extract from the genome sequence the complete RNA sequence, that is, both the aptamer and the expression platform of the riboswitch. To predict the structure after a target molecule binds to the aptamer of the riboswitch, it is also necessary to know the sequence and in turn the structure of the expression platform: then only we could identify the subsequences of the RNA involved in an alternate, stable riboswitch structure. The second step is to predict the secondary structure with the extracted RNA sequence such that the elements of the expected riboswitch family appears in the folded secondary structure. For example, in the aptamer portion of a TPP riboswitch there is a *thi*-box element, whose structure, and a significant portion of the sequence as well, is conserved in Prokaryotes and in some Eukaryotes). To achieve this, it is desirable to have a database containing the correct secondary structures of known riboswitches. The RFAM database has a collection of riboswitch sequences with the consensus structure, and the sequences corresponds to just the aptamer portion. We developed a computational pipeline for generating accurate secondary structures for all TPP riboswitch entires in the RFAM database. In thiswork, we use the software tools in pipeline to achieve the following: (a) retrieve sequences from genome banks corresponding to TPP riboswitch entries in RFAM, (b) locate the aptamer portion in the retrieved sequence, and (c) fold sequences to predict secondary structures that are accurate compared to the conserved structure in known TPP riboswitches.

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. This year we tried another construction: 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 [12], we present the comparison between both constructions. In the recent years, we also worked on introducing a hierarchical structure 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 has been submitted for publication. It is part of the thesis of Thomas Bourquard [1].

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: the classic problem of the fold recognition consists in predicting the threedimensional structure of a protein from its sequence of amino acids, using the modelling by homology. An additional approach consists in inverting this problem, and in raising the inverse folding problem: identify the most favorable sequences corresponding to a 3D structure, or given *fold* [7], [9]. main question is to map the millions of protein sequences extracted from the genomes onto the tens of thousand known 3D structures. 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). 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

Our algorithm [16] predicts first a super-secondary structure by dynamic programming. This step runs in $\mathcal{O}(n^3)$ for the common up-down topology, and at most $\mathcal{O}(n^5)$ 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. The method has been tested on transmembrane β-barrel proteins and it reaches comparable efficiency with respect to previous approaches. It can be further improved by refining the energetic model, especially on turns and loops. The structural model may be also refined since additional structural constraints may simplify the problem. The prediction accuracy, for the class of known β-barrel transmembrane proteins, evaluated as the percentage of well-labelled residues, reaches 70-85%. The number of strands is correctly predicted, whereas the shear number, the second main geometric characteristic for a β-barrel, is relatively suitable. The method is being used to carry out screening experimentations on proteomic databases, eg. the PARAMECIUM bank, in a collaboration with Ph. Dessen (Institut Gustave Roussy).

5.3. Annotation

5.3.1. Combinatorics

5.3.1.1. Word counting and trie profiles

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. The main difficulty comes from overlapping occurrences of motifs. This is partially solved by our previous algorithm, AHOPRO. OVGRAPH [6] and , developed with our associate team MIGEC, intends to solve memory problems. We introduced a new concept of overlap graphs to count word occurrences and their probabilities. The concept led to a recursive equation that differs from the classical one based on the finite automaton accepting the proper set of texts. In case of many occurrences, our approach yields the same order of time and space complexity as the approach based on minimized automaton. OVGRAPH algorithm relies on traversals of a graph, whose set of vertices is associated with the overlaps of words from a set $\mathcal{H}$. Edges define two oriented subgraphs that can be interpreted as equivalence relations on words of $\mathcal{H}$. Let P be the set of equivalence classes and S be the set of other vertices. The run time for the Bernoulli model is $O(np|S| + |\mathcal{H}|)$. In a Markov model of order K, additional space complexity is $O(pm|V|^K)$ and additional time complexity is $O(npm|V|^K)$. Our preprocessing uses a variant of Aho-Corasick automaton and achieves $O(m|\mathcal{H}|)$ time complexity. Our algorithm is implemented for the Bernoulli model and provides a significant space improvement in practice.

A new problem addressed by MPV, developed with J. Bourdon (LINA-Nantes and INRIA-SYMBIOSE) and MIGEC, is the significance assessment for motifs clusters. The classical method to study a set of motifs (defined, for instance, by their *Position Weight Matrices*, PWM), computes a significance score for each motif in the sequence set to be studied and then choses (arbitrarily) a threshold to select the most significant motifs (10 top motifs, motifs with a pvalue smaller than 5%,...). Such a type of choice makes very difficult to keep under control the number of false positive induced by this selection. We have developed a method, that relies on generating functions, that allows to computes a significance criterium for the selection. Therefore, it provides the number of false positive. Such an information is beyond the scope of other methods that correct the pvalues for multiple tests: Bonferroni,Benjamini-Hochberg,...A prototype is available on line http://www.lina.sciences.univ-nantes.fr/bioatlanstic/MPV/.

Some related theoretical aspects have been considered by P. Nicodème. The non-reduced case of words statistics is considered, where words of the searched motif may be factors of other words of the motif. This is a joint work with Frédérique Bassino (LIPN, University Paris-North) and Julien Clément (GREYC, University of Caen); an article about this matter has been submitted to the Journal *Transaction on Algorithms*. Since DNA is a text sequence, it is ubiquitous to present the importance of analysis of suffix-trees. This latter analysis is often coupled with the analysis of tries. A joint work of P. Nicodème with Gahyun Park (University of Wisconsin), Hsien-Kuei Hwang (Academia Sinica, Taiwan) and Wojciech Szpankowski (University of Purdue) about Profiles of Tries has been published in the SIAM Journal on Computing [8].

5.3.1.2. Random Generation

The random generation of combinatorial objects is a 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. To that purpose we introduced a weighted random generation, which we previously implemented within the `GenRGenS` software http://www.lri.fr/~genrgens/. The weighted distributions induced by our generation generalizes both Markov models for genomic sequences and the Boltzmann distribution used by state-of-the-art methods for RNA folding.

In this collaboration between two of the team members and M. Termier (IGM-University Paris-Sud XI), we introduced and studied a generalization of the weighted models to general decomposable classes, defined using different types of atoms $\mathcal{Z} = \{\mathcal{Z}_1, ..., \mathcal{Z}_{|\mathcal{Z}|}\}$. We addressed the random generation of such structures with respect to a size n and a targeted distribution in k of its *distinguished* atoms. We consider two variations on this problem. In the first alternative, the targeted distribution is given by k real numbers $\mu_1, ..., \mu_k$ such that $0 < \mu_i < 1$ for all i and $\mu_1 + \cdots + \mu_k \leq 1$. We aim to generate random structures among the whole set of structures of a given size n, in such a way that the *expected* frequency of any distinguished atom $\mathcal{Z}_i$ equals μ_i. We address this problem by weighting the atoms with a k-tuple π of real-valued weights, inducing a weighted distribution over the set of structures of size n. We first adapt the classical recursive random generation scheme into an algorithm taking $O(n^{1+o(1)} + mn \log n)$ arithmetic operations to draw m structures from the π-weighted distribution. Secondly, we address the analytical computation of weights such that the targeted frequencies are achieved asymptotically, i. e. for large values of n. We derive systems of functional equations whose resolution gives an explicit relationship between π and N. Lastly, we give an algorithm in $O(kn^4)$ for the inverse problem, *i.e.* computing the frequencies associated with a given k-tuple π of weights, and an optimized version in $O(kn^2)$ in the case of context-free languages. This allows for a heuristic resolution of the weights/frequencies relationship suitable for complex specifications. In the second alternative, the targeted distribution is given by k natural numbers $n_1, ..., n_k$ such that $n_1 + \cdots + n_k + r = n$ where $r \geq 0$ is the number of undistinguished atoms. The structures must be generated uniformly among the set of structures of size n that contain *exactly* n_i atoms $\mathcal{Z}_i$ $(1 \leq i \leq k)$. We give a $O(r^2 \prod_{i=1}^{k} n_i^2 + mnk \log n)$ algorithm for generating m structures, which simplifies into a $O(r \prod_{i=1}^{k} n_i + mn)$ for regular specifications.

These results provide new foundations and tools for tackling structural bioinformatics problems, such as RNA design. They are described in a manuscript [23] submitted to *Theoretical Computer Science*.

5.3.1.3. Score function for SNK

Recent work by Forslund and Sonnhammer has investigated to which extent the hypothesis that protein function should follow largely from domain architecture can be true. They have shown that domain functional interplay may not follow directly from the properties of the domains in isolation, and suggested that it could be interesting to take into account conservation of sequential order of the domains. To achieve this, we have proposed a new method [3], called SNK (Sequential Nuggets of Knowledge) http://www.lri.fr/~rance/SNK/, which systematically analyses domain combinations and outlines characteristic patterns potentially associated with targeted properties, such as sets of GO terms or membership to some taxonomic group. We are currently applying this method to discover new associations in some proteins families. Also, we are defining a robust probability model on the variables involved in the sequential association rules to highlight their relevance.

5.3.2. Ontology and provenance

5.3.2.1. Ontology mapping

Identifying correspondences between concepts of two ontologies has become a crucial task for genome annotation. We have proposed O'BROWSER [14], a semi-automatic method to solve that issue in the case of two functional hierarchies. O'BROWSER is based on a classical ontology mapping architecture, but strongly uses expertise on the underlying domain. First, experts are asked to validate obvious correspondences discovered by O'BROWSER and to identify functional groups of concepts in the ontologies. Then, they are requested to validate the correspondences given by the combination of results found in the automatic steps of

our system. These steps consist in matchers designed to fit the characteristics of the ontologies. Especially, we have introduced a new instance-based matcher which uses homology relationships between proteins. We also proposed an original notion of adaptive weighting for combining the different matchers. O'BROWSER has been used to map concepts of SUBTILIST to concepts of FUNCAT, two functional hierarchies.

5.3.2.2. Browsing biomedical datasources

One of the most popular ways to access public biological data is using portals, like ENTREZ NCBI. Data entries are inspected in turn and cross-references between entries follow. However, this navigational process is so time-consuming and difficult to reproduce that it does not allow scientists to explore all the alternative paths available (even though these paths may provide new information). BIOBROWSING [13] is a tool providing scientists with data obtained when all the possible paths between NCBI sources have been followed (source paths generation is done by BIOGUIDE). Querying is done on-the-fly (no warehousing). BIOBROWSING has a module able to update automatically the schema used by its query engine to consider the new sources and links which appear in ENTREZ. Finally, profiles can be defined as a way of focusing the results on userâs specific interests.

5.3.2.3. Differencing two workflows

In this context, we have studied the problem of differencing two workflow runs with the same specification. Our contributions [10] are three-fold: (i) while in general this problem is NP-hard, we have proposed to consider a natural restriction of graph structures (series-parallel graph overlaid with well-nested forking and looping) general enough to capture workflows encountered in practice; (ii) for this model of workflows, we have presented efficient, polynomial-time algorithms for differencing workflow runs [18],[11]; (iii) we have developed a prototype [4] and conducted experimental results demonstrating the scalability of our approach.

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.

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.

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) has started a new activity on a DIGITEO chair about RNA properties, in particular concerned with folding energy distributions and the identification of riboswitches.

7.1.2. Associate Team

MIGEC, Mathématiques et Informatique en GEnomique Comparative (Mathematics and Computer Science in Comparative Genomics), is an associate team with NII-GENETIKA (Moscow, Russia). The goal of this cooperation is the development of analytic and statistical criteria in order to extract and analyze complex motifs in sequences and to use these criteria on entire genome sequences as well. This includes the development of methods for complex motifs and combined motifs identification in the genomes, analytic and numerical approaches to asess the statistical significance of candidates and an experimental verification of putative motifs. Our main application is the analysis of regulatory regions in eucaryote organisms, such as the man, the mouse and insects. A special attention is paid to promoter sequences and to CpG islands in genes that control the tissue differentiation and tumorogenesis. In this project, AMIB members bring their skills and tools in pattern matching algorithms and (probabilistic) combinatorial enumeration. Such results are complementary to the genome analysis technology developped at NII-GENETIKA, that includes genomic databases organisation, databases creation for functionnally important regions and data integration from different sources in biology and bioinformatics. This associate team takes place in a long history of collaboration between Moscow and Inria groups, that also includes biologists from Berkeley.

7.2. Exterior research visitors

Professor D. Frishman and S. Neuman (MIPS, Munchen) visited AMIB during two days and one week, respectively. Professor V.Makeev (NII-GENETIKA and MIGEC) did a one week visit, and E. Furletova (MIGEC PHD student) visited three times during two weeks. Professor M. Ward (Purdue University) did a one week visit and A. Sim (Stanford University, Associate team GNAPI) did a two weeks visit. Professor R. Backofen (Heidelberg) did a two day visit. Professor N. Leontis (Bowling Green State University) did a one week visit.

8. Dissemination

8.1. Scientific Community Involvment

8.1.1. French Bioinformatics

Participants: Patrick Amar, Jérôme Azé, 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). A. Denise has been the head of this GDR since 2006, Ch. Froidevaux was in charge of subdomain *Knowledge Representation, Ontologies, Data Integration and Grids* and J. Azé is the webmaster.

The *Programme PluriFormation, PPF* Bioinformatics and Biomathematics, headed by Ch. Froidevaux, gathers teams of computer scientists, mathematicians, and biologists from the University of Paris Sud-XI interested in bioinformatics and biostatistics. All the team is involved and participated in the final workshop at Tours, (September, 14th-15th).

8.1.2. Seminars

8.1.2.1. Amib seminar

Our seminar is held three times a month. This fall, we welcomed a seminar by B. Behzahdi (Google Research), A. Sim (Stanford University), S. Neuman (MIPS, Munich), M. Ward (Purdue University), N. Leontis (Bowling Green State University), F. Leclerc (Nancy University).

8.1.2.2. Other seminars

P. Amar was invited to give the talk *Modelling self assembly and behaviour of molecular complexes* at the Workshop on “MAS in Biology at the meso or macroscopic scales” in Paris on June, 23rd.

J. Bernauer was invited to give a talk on "Computational Structural Biology: Periodic Triangulations for Molecular Dynamics" at the Workshop "Subdivide and tile: Triangulating spaces for understanding the world", organized in Leiden (Netherlands), 16-20 November, 2009. See http://www.lorentzcenter.nl/lc/web/2009/357/info.php3?wsid=357. J. Bernauer is attending the "Fourth CAPRI Evaluation Meeting in Barcelona, 9-11 December, 2009.

Y. Ponty was invited to give the talk *RNA as a combinatorial object: Asymptotics of RNA Shapes*" at the bioinformatics seminar (hosted by R. Backofen) of the Technical university of Freiburg on November, 27th.

Thuong Van Du Tran attended MCCMB’09 (Moscow, Russia) and ISMB/ECCB2009 (Stockholm, Sweden) and presented posters.

8.1.3. Program Committee

P. Amar was a program committee member and scientific committee member of the conference *Modelling Complex Biological Systems in the context of genomics*.

J.Bernauer is chair of *Multi-resolution Modeling of Biological Macromolecules* session at the Pacific Symposium on Biocomputing 2010

S. Cohen-Boulakia was a program committee member of international conferences or workshops SSDBM2009, DILS 2009, SWPM-2009 (First Int. Workshop on the role of Semantic Web in Provenance Management, co-located with Iswc-2009), ICDE 2010 (general track and demo track) and of national conferences BDA2009, JOBIM2010.

Ch. Froidevaux was a program committee member of EDBT2010, IB2010, DILS2009, IEEE CBMS2009 (Computer-Based Medical Systems-special track on Computational Proteomics-), Third Int. Workshop on "Biomedical and Bioinformatics’ Challenges to Computer Science" co-located with ICCS (2009 et 2010) and of national conferences, EGC2009, EGC2010, JOBIM2009.

Ch. Froidevaux and S. Cohen-Boulakia organized workshop *Metadata, Ontologies and Quality of Annotation*, MOQA (september, 27th).

M. Régnier is a program committee member of RECOMB workshop on Regulatory Genomics and co-organized MCCMB’09 in Moscow.

8.1.4. Research Administration

A. Denise serves in the National Committee of Scientific Research: section 7, Sciences et Technologies de l’Information (Computer Science, Control, Signal and Communication) and interdisciplinary commission 43 (Modélisation de systèmes biologiques, bioinformatique).

Ch. Froidevaux has been the head of Computer Science Department at University Paris-Sud XI (UFR des Sciences d’Orsay) since January, 15th. She participated to the AERES committee that evaluated INRIA Lille Nord-Europe CRI.

M. Régnier serves in the Committee of French ANR http://www.agence-nationale-recherche.fr/.

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. From September 2010, it will become a joint Master between University Paris-Sud and Ecole Polytechnique. Most members of the group teach in the Master.

M. Régnier has been invited by Al Farabi University (Almaty, Kazakhstan) to deliver a 20 hours master course in bioinformatics. She serves in the Committee of French Agregation of Mathematics (Computer Science option).

J. Bernauer teaches at AgroParisTech, Paris, MAP3 (3h) and at University of Nice - Sophia-Antipolis, Master of Science in Computational Biology; *Algorithmic Problems in Computational Structural Biology* (9h).

C. Saule is a teaching assistant at Orsay UFR (*Internet programming, Engineering software, Data bases* and JAVA). Philippe Rinaudo is a teaching assistant for *Programmation principles and languages* (Master 2 CCI) and *Algorithmics and complexity in biology* (Master 1 BIBS). Van Du Tran teaches a course on *Algorithm and Complexity* and a course on JAVA in L3 at Orsay.

9. Bibliography

Year Publications

Doctoral Dissertations and Habilitation Theses

[1] T. BOURQUARD. *Exploitation des algorithmes génétiques pour la prédiction de structures de complexes protéine-protéine*, Laboratoire de Recherche en Informatique (LRI) – Université Paris-XI/Paris Sud, December 2009, Ph. D. Thesis.

[2] M. DJELLOUL. *Algorithmes de graphes pour la recherche de motifs récurrents dans les structures tertiaires d'ARN*, Laboratoire de Recherche en Informatique (LRI) – Université Paris-XI/Paris Sud, December 2009, Ph. D. Thesis.

[3] B. RANCE. *Fouille et intégration de données biologiques hétérogènes*, Laboratoire de Recherche en Informatique (LRI) – Université Paris-XI/Paris Sud, September 2009, Ph. D. Thesis.

Articles in International Peer-Reviewed Journal

[4] 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", vol. 2, n^{o} 2, 2009, p. 1638-1641 US .

[5] K. DARTY, A. DENISE, Y. PONTY. *VARNA: Interactive drawing and editing of the RNA secondary structure*, in "Bioinformatics", vol. 25, n^{o} 15, Apr 2009, p. 1974–1975, http://bioinformatics.oxfordjournals.org/cgi/content/abstract/btp250.

[6] A. IVASHCHENKO, G. BOLDINA, A. TURMAGAMBETOVA, M. RÉGNIER. *Using profiles based on hydropathy properties to define essential regions for splicing*, in "International Journal of Biological Sciences", vol. 5, 2009, 10 p., http://hal.inria.fr/inria-00429780/en/KZ.

[7] A. LOPES, M. SCHMIDT AM BUSCH, T. SIMONSON. *Computational design of protein:ligand binding: modifying the specificity of asparaginyl-tRNA synthetase*, in "J. Comp. Chem.", vol. in press, 2009, 0000.

[8] G. PARK, H.-K. HWANG, P. NICODÈME, W. SZPANKOWSKI. *Profile of Tries*, in "SIAM journal on Computing", vol. 38, n° 5, 2009, p. 1821-1880 US TW .

[9] M. SCHMIDT AM BUSCH, D. MIGNON, T. SIMONSON. *Computational protein design as a tool for fold recognition*, in "Proteins", vol. 77, 2009, p. 139–158.

Invited Conferences

[10] S. DAVIDSON, Y. CHEN, P. SUN, S. COHEN-BOULAKIA. *On User Views in Scientific Workflow Systems (Invited Paper)*, in "Proc. of the the First Int. Workshop on the role of Semantic Web in Provenance Management (ISWC 2009 Workshop)", 2009 US .

International Peer-Reviewed Conference/Proceedings

[11] Z. BAO, S. COHEN-BOULAKIA, S. DAVIDSON, A. EYAL, S. KHANNA. *Differencing Provenance in Scientific Workflows*, in "Proc. of the 25th Int. Conf. on Data Engineering (ICDE), IEEE", 2009, p. 808-819 US .

[12] T. BOURQUARD, J. BERNAUER, J. AZÉ, A. POUPON. *Comparing Voronoi and Laguerre tessellations in the protein-protein docking context*, in "Sixth annual International Symposium on Voronoi Diagrams, Danemark Copenhagen", F. Anton and J. Andreas Bærentzen - Technical University of Denmark, 2009, http://hal.inria.fr/inria-00429618/en/.

[13] S. COHEN-BOULAKIA, K. MASINI. *BioBrowsing: Making the Most of the Data Available in Entrez*, in "21st Int. Conf. in Scientific and Statistical Database Management (SSDBM), LNCS 5566, Springer", 2009, p. 283-291.

[14] B. RANCE, J.-F. GIBRAT, C. FROIDEVAUX. *An adaptive combination of matchers: application to the mapping of biological ontologies for genome annotation*, in "Data Integration in the Life Sciences, DILS 2009", N. W. PATON, P. MISSIER, C. HEDELER (editors), Lecture Notes in Computer Science, vol. 5647, Springer, 2009, p. 113-126.

[15] C. SAULE, A. DENISE. *Counting RNA pseudoknotted structures*, in "Proceedings of ISMB/ECCB, Stockholm", June 2009.

[16] V. D. TRAN, P. CHASSIGNET, J.-M. STEYAERT. *Prediction of super-secondary structure in alpha-helical and beta-barrel transmembrane proteins*, in "Highlights from the Fifth International Society for Computational Biology (ISCB) Student Council Symposium", vol. 10, n° Suppl 13, 2009, O3, http://www.biomedcentral.com/1471-2105/10/S13/O3.

Scientific Books (or Scientific Book chapters)

[17] P. AMAR, F. KÉPES, V. NORRIS, G. BERNOT. *Proceedings of the Nice 2009 spring school on Modelling Complex Biological Systems in the context of genomics*, EDP Sciences, 2009.

[18] S. COHEN-BOULAKIA, W. C. TAN. *Provenance in Scientific Databases*, in "Encyclopedia of Database Systems", Springer US, 2009, p. 2202-2207 US .

[19] Z. LACROIX, C. R. KOTHARI, P. MORK, R. RIFAIEH, M. WILKINSON, J. FREIRE, S. COHEN-BOULAKIA. *Biological Resource Discovery*, in "Encyclopedia of Database Systems", Springer US, 2009, p. 220-223 US CA .

[20] Z. LACROIX, C. R. KOTHARI, P. MORK, M. WILKINSON, S. COHEN-BOULAKIA. *Biological Metadata Management*, in "Encyclopedia of Database Systems", Springer US, 2009, p. 215-219 US CA .

[21] V. NORRIS, P. AMAR, M. AIMAR, P. BALLET, A.-F. BATTO, G. BARLOVATZ, G. BERNOT, G. BESLON, A. CABIN, S. CHEVALIER, A. DELAUNE, J.-M. DELOSME, E. FANCHON, H. GAO, N. GLADE, Y. GRONDIN, D. HERNANDEZ-VERDUN, L. JANNIERE, F. KÉPES, C. LANGE, G. LEGENT, C. LOUTELIER-BOURHIS, F. MOLINA, N. ORANGE, D. RAINE, C. RIPOLL, M. THELLIER, A. THIERRY, P. TRACQUI, A. ZEMIRLINE. *Hyperstructures 2008-2009*, in "Modelling Complex Biological Systems in the Context of Genomics", EDP Sciences, 2009, p. 71–84.

[22] M. RÉGNIER, Z. KIRAKOSSIAN, E. FURLETOVA, M. ROYTBERG. *A Word Counting Graph*, in "London Algorithmics 2008: Theory and Practice (Texts in Algorithmics)", J. CHAN, J. W. DAYKIN, M. SOHEL RAHMAN (editors), London College Publications, 06 2009, p. 10–43, http://hal.archives-ouvertes.fr/inria-00437147/en/AMRU.

Other Publications

[23] A. DENISE, Y. PONTY, M. TERMIER. *Controlled non-uniform random generation of decomposable structures*, 2009, Submitted to Theoretical Computer Science.

[24] 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*, 2009, Submitted to Theoretical Computer Science.

[25] C. SAULE, M. RÉGNIER, J.-M. STEYAERT, A. DENISE. *Counting RNA pseudoknots*, 2009, submitted to SFCA/FPSAC'10 : San Francisco, USA, 2010 and to workshop "Algorithmique, combinatoire du texte et applications en bio-informatique", Montpellier, janvier 2010.