

# Action de Recherche Amont MASSE de DONNEES

## Scientific Description of the project Modulome

### 1. Goal and context

#### Context

A number of nucleic sequences from a large number of organisms are now available. They form a huge repository of data growing exponentially and produced at a decreasing cost. Indeed, for more than 10 years, the size of genomic banks has doubled every year. Today, it represents more than 50 Giga bytes of raw sequences, and for the next ten years it is expected to be multiplied by 1000.

In contrast, the exploitation of such a quantity of data remains relatively poor. The structure of a genome is usually described as a DNA sequence, which may be organised in several chromosomes, associated to a set of annotations (mainly coding genes for proteins and various kinds of RNA).

With the availability of more than 270 complete genome sequences from Archaea, Bacteria and Eukarya (data from GOLD, [www.genomesonline.org](http://www.genomesonline.org)) and more than 1200 ongoing sequencing projects, more ambitious projects in fundamental biology are starting in the international community: comparing entire genomes will allow to gain new insights into evolutionary, biochemical, genetic, metabolic, and even physiological pathways. For instance, DoE (Department of Energy, USA) has initiated several important projects in his program GTL (Genome to Life) towards a more comprehensive, integrated view of biology at the whole-systems level.

Some years ago, genomes were considered as static objects containing an informative part, the coding sequences, which only represent a few percent of the total genome, and a so-called junk DNA part that was generally free of any annotation. It is now widely acknowledged that genomes must be considered with a more dynamic point of view, involving the study of the many “copy” events that occur during evolution and concern coding genes as well as non coding sequences generally involved in regulation. A wide number of *in silico* studies of repeated sequences have lead to the hypothesis that they play major roles in the structure, the function, the dynamics and the evolution of genomes in Archaea [Blount, 2005; Mojica, 2005], Bacteria [Achaz et al., 2002; Pourcel, 2005] and Eukarya [Achaz, 2000; Friedman, 2001; Kazazian, 2004]. Repeats may also be strictly conserved through the evolution, as revealed by comparisons of human, mouse, rat, chicken and dog genomes [Berejano, 2004]. Complex mechanisms such as chromosome segments duplications, or even whole genome duplications, are thought to occur, explaining genome evolutions [Taylor, 2003; Dujon, 2004]. Also, converging studies on the human and other genomes revealed that variations in the number of occurrences of particular repeats may be an important responsible factor of diseases [Rubinsztein et al., 1995].

At the core of life information, there exists an outstanding opportunity to analyse the genomic structure by deciphering its content in repeated sequences. Indeed, the exhaustive analysis of published genomes has revealed that most of them, especially in Eukarya, have a genomic content consisting in large proportions of repeats. As an example, more than half of the human genome is repeated sequences [International Human Genome Sequencing Consortium, 2001]. Revealing the structure of sequences as an assembly of elementary repeated “domains” is thus a task of utmost importance. The decomposition of biological sequences into domains is already well studied, at the price of heavy computations, at the level of protein sequences [Servant et al., 2002]. This task remains to be done at the level of genomes, and we refer to this exhaustive search of genomic domains, or modules, as deciphering the “modulome”. Such a task is highly combinatorial in nature, since considering repeats instead of sequences themselves considerably increases the quantity of objects to be considered: each position in a sequence may contribute to several overlapping or embedded repeats.

## Goals

Our project, Modulome, aims at providing methods for the **identification, visualization and formal modelling of the structure of genomes in terms of an assembly of nucleotides “modules” (we call them v-modules)** that are repeated along a genome or between several genomes. Combined together, these methods will provide an appropriate methodology for a fruitful production of hypotheses concerning genome organizations. The challenge is to allow the biologist to represent and reason on large genomic sequences in an abstract way, by segmenting them into v-modules and revealing the organization of such modules. It would thus be possible to get a more unified view of genomes and to discover new interesting structures (e.g. promoters or transposons) in genomes.

## Related projects

A very common but unsuited method to locate repeated segments is to produce a bench of Blast comparisons [Altschul et al. 1997] on genomic subsequences and then try to analyse the results in a coherent way. This approach suffers from many drawbacks, including the fact that it is a heuristic approach, that it produces too much irrelevant results, and that it can hardly be exhaustive. During the past few years, several methodological approaches have been proposed to investigate the three main classes of repeats in entire genomes: tandem repeats (consecutive copies of patterns of  $k$ -mers,  $k$  being generally less than 5 but sometimes greater than 1000 in case of microsatellites), duplicated segments (which include genes and chromosome segments duplications) and interspersed repeats (which include transposons). These methodological approaches can be algorithmic and/or visualization-based.

Algorithmic-based methods have been proposed to locate tandem repeats, such as Tandem Repeats Finder [Benson, 1999], as well as tools aiming at identifying known repeats, such as the widely-used RepeatMasker [Smit and Green, unpublished]. In addition to only identifying particular repeats categories, these tools have some restriction on the maximum length of the treated sequences.

Efficient computational tools, such as REPuter [Kurtz et al., 2001], RepeatFinder [Volfovsky, 2001] and FORRepeats [Lefebvre et al., 2003], are able to find both exact and error-prone repeats in sequences as long as eukaryotic chromosomes (up to 100 millions nucleotides). However, these tools do not provide any overview or summary of the repetitive structure of the sequence. Even if they provide a repeat-graph based graphical interface, like REPuter and FORRepeats, it becomes quite difficult to analyse repeats organization because of the huge amount of data.

Recon [Bao et Eddy, 2002] and FragmentGluer [Pevzner et al., 2004] create clusters of repeats using algorithms based on local multiple alignments. However, use of multiple alignments is not always adequate to represent mosaics (i.e. hierarchical structures) of repeats [Horvath et al., 2000].

ADHoRe [Vandepoele, 2002] and DAGchainer [Haas, 2004] use methods aim at mining genes duplications at genome levels. As a consequence, they do not take into account other types of repeats.

Visualization is also a fundamental part of repeat analysis. Various kind of tools have been proposed during the past years, including dotplot [Gibbs and McIntyre, 1970], chaos game [Jeffrey, 1990], percent identity plot [Schwartz et al., 2000], repeat-graph [Kurtz et al., 2001] and arc-diagram [Spell et al., 2003]. However, all these tools apply on single sequence or pairwise genome sequence alignments, and, because they only work at the sequence level, they become difficult to use as the number of repeats increase.

To summarize, currently available methods suffer from several limitations with respect to an in-depth study of genome organization. They have generally been designed to target particular application of genomic repeats analysis. As a consequence, they generally look for heuristically defined repeats (a repeat is essentially defined by the procedure that recognize it), and they do not provide an easy access to a global overview of the repeated structures at a genome level (neither for an analysis of the structure of a single genome, nor for a comparison of genomes). Computing a set of relations between pairs of segments that are “similar” and eventually drawing some statistics from them may be a useful first step. But it is clearly not sufficient to get a comprehensive view of the organization and to infer possible mechanisms governing the genomes structuring.

## Issues

Hard problems concern, first, the specification of a formal, biologically relevant *definition of v-modules* on DNA, and, second, the design of complete and efficient *algorithms enumerating* the solutions. Of course, this cannot be achieved without a firm anchoring on a biological background, that involves in particular taking into account a typology of such v-modules (e.g tandem repeats, transpositions...) and to fix some thresholds (e.g. maximum length or scope of repeats) for a precise specification of each type. Another issue is to *give access* to this new type of information in a way that effectively allows the biologist to draw some hypotheses from them. As in many data mining techniques, it is possible to produce a volume of final results greater than the initial volume of data! Carefully choosing the types of visualization and selection operators that have to be implemented is crucial with respect to a full exploitation of results. Too often, visualization techniques in this domain seem to seek for nice publishable figures. What is needed, in the daily practice of laboratories, is some *clever browsing techniques* allowing each biologist to capture interesting properties from data and to infer hypotheses in the form of theoretical models. This leads to a last type of problem, which concerns the verification of such hypothetical models at the level of whole genomes. As long as models are words or even more complicated patterns such as regular expressions, verification remains feasible. However, for the more expressive combinations needed by descriptions involving sequence repeats, one lacks parsers working both with the right expressiveness and at the scale of complete genomes. Thus, there exists actually a gap between the computation of repeated structures and their exploitation into formal and operational models.

## Contributions of the project

The Modulome project aims at providing methods for the description of genome structures in term of assembly of *v-modules* that are duplicated along a genome or between several genomes. The architecture of these *v-modules* may be relatively complex. In order to provide a global overview of the v-modules, we propose (1) to capture the hierarchical structures of repeats, (2) to model such structures using a formal language, and (3) to search these models within databases of genome sequences.

More specifically, the Modulome project targets the following scientific and technological key issues for genome analysis:

1. Develop highly efficient computational algorithms able:
  - to identify complex repeat structures (*v-modules*) in reasonable time and space,
  - to deal with huge amount of data. For example, processing all the available bacteria genomes require to manage a few hundreds of gigabytes of data, what would imply, as a first estimate, to process more than  $10^{11}$  *v-modules*.
2. Design an original graphical user interface able to give a synthetic view of the *v-modules* organization inside one or several genomes. This *module viewer* must interactively guide biologists to the understanding of the overall genome structure.
3. Create a high throughput parser for capturing *v-modules* detected by the two previous steps on new genomes. Searching for such complex templates require optimized algorithms ensuring fast database scan among a huge amount of data. Based on the approach developed in point 1, the project will lead to a new parser going far beyond the possibilities of Blast analyses.

## Presentation of Teams

The Modulome project associates four research teams:

- 1 **INRIA bioinformatics project Symbiose, Rennes:** it will have the charge of developing the bioinformatics tools. Symbiose has an experience of more than ten years of cooperative research works with biological laboratories. Symbiose is responsible of the bioinformatics platform of Ouest genopole (<http://www.genouest.org>) and transfer its developments on this platform for a wide access to the community. Symbiose has already worked on a particular family of repeats: the Atrep transposons family.
- 2 **Laboratoire d'Etude des Parasites Génétiques (LEPG), Tours:** LEPG members will perform the study, using the modulome facilities, to inventory and characterize various forms of packaged  $\nu$ -modules in the human genome: the pack-*miHsmar1* family. Their occurrence will be then investigated in one close relative, the chimpanzee, and in at least one more distant species, the cheep. LEPG has an expertise in the study of *mariner* transposons. LEPG and Symbiose have deposited another project on the evolutionary genomics of polydnviruses in the framework of a CNRS call.
- 3 **Laboratoire de Microbiologie des Environnements Extrêmes (LM2E), Brest** LM2E will perform the study, using the modulome facilities, to look for and to inventory CRISPR elements within the Archaea domain (Euryarchaeota and Crenarchaeota). The repeated DNA elements will be characterized in the genome of hyperthermophilic archaeal and their mobile genetics elements (plasmid, virus). LM2E is an important actor of Ouest genopole with an original focus on Archaea and hyperthermophilic Bacteria living at temperatures close to 100°C in marine environments. LM2E has already worked with Symbiose on a related family, testing some of the ideas that will be developed in this project.
- 4 **Laboratoire Dynamique du Génome et Evolution, Institut Jacques Monod (LDGE), Paris.** LDGE will perform the study, using the modulome facilities, to look for and to annotate transposable elements in *D. melanogaster*, *D. simulans* and *D. yakuba* genomes. The team main research area concern the impact of transposable elements on the structure and functioning of genomes. LDGE has collaborated with LEPG but works on a different biological model: *Drosophila* and more recently anopheline mosquitoes.

Thus, for validating the Modulome approach, three large scale genomic studies will be carried out:

1. Archaea and their mobile genetic elements, and Bacteria (LM2E, Brest): The objective will be to study the biological significance of a family of regularly spaced repeats derived from foreign genetics elements in hyperthermophilic Archaea genomes using the modulome facilities. This approach will be extended next to the Bacteria domain.
2. Eukarya (LEPG, Tours): *miHsmar1* are 80 bp transposons using the transposase of complete *Hsmar1* to move in mammal genomes. Their mobility as *solo* element is limited by their size. The objective is to verify how bigger structure made of various gDNA fragments flanked by one *miHsmar1* at ends, the pack-*miHsmar1*, jump in chromosomes.
3. Eukarya (LDGE, Paris): LDGE proposes to include tools developed during the modulome project in a general annotation platform of transposable elements. LDGE has already a good practical experience in bioinformatics. Its strategy for *de novo* transposable elements (TE) identification will use combined-evidences derived from the integration of multiple homology-based and various other *de novo* TE identification methods.

These studies have been chosen with respect to the following criteria:

1. to gather teams familiar with bioinformatics and able to work on biologically relevant occurrences of  $\nu$ -modules;
2. to find examples of  $\nu$ -modules in Archaea, Bacteria and Eukarya kingdom;
3. to study different types of  $\nu$ -modules: modules transferred between host and viruses (and plasmids), modules transposed inside a genome and packing of modules;
4. to focus on the fundamental issue of mobile genetic elements and their distribution in complete genomes.

## 2. Project description

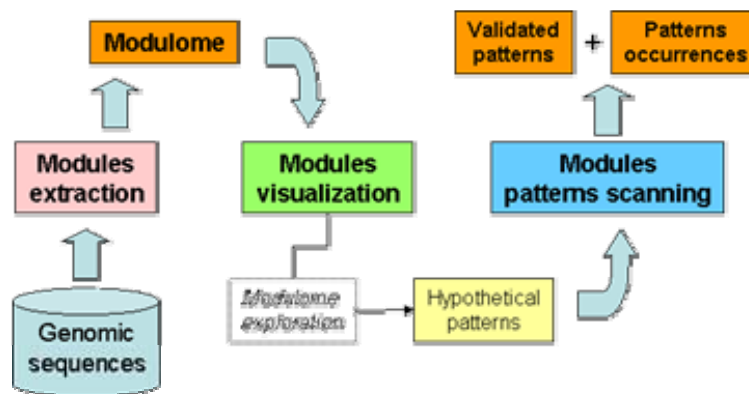
We introduce first the concept of modulome. The three steps of our approach are then described and illustrated on a first application. Finally, case studies are presented for various large scale observations of mobile genetic units and we propose to organize the animation of a group around these topics.

### 2.0 From repeats to modulome

Repeats are ubiquitous in genomes. A number of authors have described several mechanisms that are implied in partial or large scale duplications within genomes. Repeated (multiply occurring) sequences are abundant in eukaryotic genomes, and in some cases, represent most of the genome [Margaret, Damon, 2000]. Many studies show the relationship between a given family of repeat elements and the host genome, but except for phylogeny and some crude statistics, very few studies analyze the relationships and systematical variations between copies of a given family of repeats. Three main classes of repeats occur in entire genomes: tandem repeats (consecutive copies of patterns of  $k$ -mers,  $1 < k \leq 5$  or  $100 < k \leq 2000$ ), duplicated segments (which include genes and chromosome segments duplications) and interspersed repeats (which include transposons).

Brosius and Gould (1992) proposed a 'genomenclature' which provide a comprehensive taxonomy for pseudogenes and other so-called junk DNA. They proposed for this purpose a general terminology. They defined as a 'nuon' any stretch of nucleic acid sequence that may be identifiable by any criterion and propose then several classes of nuons for which they gave a number of examples.

Our goal is to fully assist biologists working on the structure and on the comparison of genomes, with operational tools to explore the world of nuons. The existence of such nuons is revealed by the existence of multiple occurrences of them. Since we try to characterize them as a hierarchically organized set of modules, we will refer to the set as a "modulome" and to the elements of this set, in the spirit of the term forged by Brosius, as  $\nu$ -modules. We propose an ambitious chain of treatment of genomes in order to work in a concrete way on modulomes. The process will include 3 steps (fig. 1):



**Fig 1: Overall structure of the analysis of  $\nu$ -modules in the project**

1. Extraction of  $\nu$ -modules: the first step is looking for a complete representation of modules inside a sequence or a set of sequences. It requires a careful design of data structures and algorithms, since one faces a double combinatorial problem: the number of nucleotides to consider is generally very large (e.g. several Gigabases) and the number of repeats may be quadratic with respect to this number! Our approach is based on powerful indexing techniques that include suffix tree data structures and special purpose architectures for the treatment of such indexes based on reconfigurable devices (FPGA);
2. Visualization of  $\nu$ -modules: biologists are fundamental actors for the interpretation of the structures emerging from the previous step. The second step and challenge is to give them access to the corresponding quantity of information in a sufficiently abstract way and with advanced visualization functionalities. This will lead to a genuine specialized browser on repeated structures.

3. Analysis of v-modules: once modules have been produced in a genome, and some hypotheses have been elaborated on the organization of some modules, one must parse them at the genome level, that is, to check their validity with respect to the whole set of available data. This assumes the existence of a high level language on sequences allowing expressing these models as patterns on v-modules, and of a parser efficiently analysing them on large sequences.

We propose to elaborate and tune such a chain of treatment on carefully chosen applications. The fundamental points of interaction between computer scientists and biologists will be on the concept of v-module and on the types of models that are necessary on genomic sequences.

## 2.1 Extraction of v-modules

A number of representations of large sequences exist, that share the common concern of providing linear operations on such sequences, in space and time. One of the most effective computational representation of a sequence is the suffix-tree data structure [McCreight, 1976; Kurtz, 1999], which is especially well suited for genomes analysis and comparison [Kurtz and Schleiermacher, 1999; Kurtz *et al.*, 2004]. Since it forms a well understood background, and a very flexible data structure, we will base our own approach on such a structure.

Our first work is to better delineate and then formally characterize a natural definition of building block, or v-module, on nucleic acids sequences. Related to the concept of "nuon" introduced by Jurgen Brosius, they are elementary entities that allow to segment these sequences and to describe them at a more abstract level. In order to assess the existence of a v-module, a minimal requirement consists in observing at least two copies of the v-module in the genome. This definition is however not sufficient, because we are not interested in all repeats in such a sequence. The notion of **maximal repeat** (MR) seems an attractive starting point for v-modules, since it only focuses on largest common blocks, without possible left or right extension (note that this definition refers to words, not to their occurrences). Maximal repeats have nice properties: they can be computed in linear time [Gusfield97], their number is limited (at most  $n$  exact maximal repeats in a word of size  $n$ ), and they "assemble" together quite well (given 3 words  $U, V, W$ , if  $UV$  and  $VW$  are maximal repeats, then  $V$  is also a maximal repeat). However, we need to take into account finer biological characteristics.

Figure 2 sketches the approach we propose. Questions to be studied are: 1) How to handle natural variations of a maximal repeat, i.e. approximations of MR including some mutations or indel? 2) What types of parameters are necessary to select the various types of interesting biological repeats? 3) How to manage efficiently huge suffix trees, exceeding the size of the main memory of a computer?

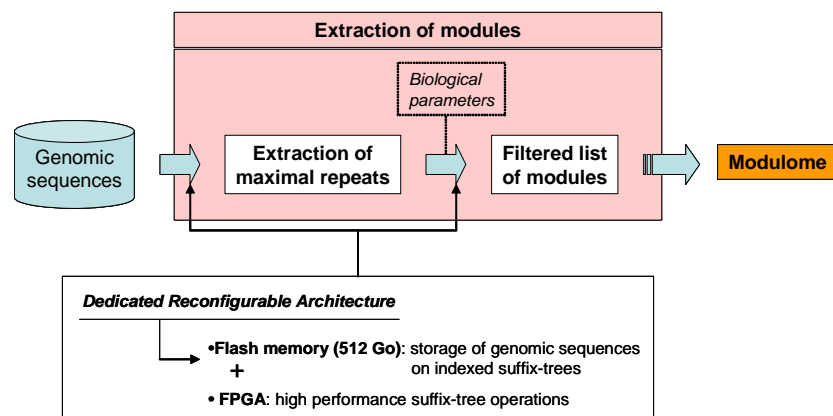


Figure 2: Production of modules from genomic data

Here, we particularly focus on the key issue of managing large suffix trees. The main limitation comes from the large and fast memory it requires to provide efficient accesses. As an example, the "human" suffix tree needs a minimum of 60 Giga bytes of memory. This is far beyond the current main

memories of computers. On the other hand, storage on magnetic disks suffers from very low accesses, especially latency, due to mechanical constraints on the reading head.

An intermediate solution is to use FLASH memory technology. As magnetic disks, FLASH components now reach high density and the information remains when the supply power is down. As DRAM memories, data can be accessed in a fast way. Such a FLASH memory computer has been designed in the ReMIX project (ACI Masses de Données, coordinator D. Lavenier). A prototype is operational since July 2005 and is dedicated to indexing applications. It offers a FLASH memory space of 512 Giga bytes, allowing containing, for example, all the suffix trees of all the available “bacteria”.

It is important to point out that the FLASH memory developed in the ReMIX project is far from being a simple memory extension. It includes specific features dedicated to fast indexing search:

- Zero latency: null data latency can be achieved if the application is able to have a short-term forecasting of the data accesses.
- On-the-fly computation: reconfigurable hardware (FPGA devices) is connected to the FLASH memory output to process data at a very high rate.

These two characteristics can be efficiently used when parsing large suffix trees. Finding repeats is a predictable path along a suffix tree, and processing maximal repeats could be performed on-the-fly by a specific operator implemented into the reconfigurable hardware.

## 2.2 Visualization of $v$ -modules

Visualization of genome sequence data is a widely explored issue in bioinformatics. There are numerous approaches to visualize either a single sequence or multiple sequences comparison at the genome level. The first type of viewers are known as genome browsers, which are available from many sources (see for example [www.acedb.org](http://www.acedb.org), [www.ensembl.org](http://www.ensembl.org) and [www.ncbi.nih.gov/Genomes](http://www.ncbi.nih.gov/Genomes)), whereas the second type of visualizers are dedicated to the analysis of conserved regions based on sequence alignments [Dicks 2000]. Since none of these applications are adapted to the analysis of repetitive sequences, dedicated tools have been developed to target that problem. Among them, there are dotplot [Gibbs and McIntyre, 1970], chaos game [Jeffrey, 1999], percent identity plot [Schwartz et al., 2000], repeat-graph [Kurtz et al., 2001] and arc-diagram [Spell et al., 2003]. However, most of these tools remain only usable on a single sequence or pairwise genome sequence alignments, and, because they only work at the sequence level, they become difficult to use as the number of repeats increases. In addition, they are not able to summarize the hierarchical organization of repetitive sequence structures in an elegant way so that they are interpretable by the end users.

A first key problem of repetitive sequences visualization at the genome level is the huge amount of data to deal with. As an example, there are around 810.000 exact maximal repeat words (size ranges from 1 to 3500 nucleotides), located at more than 15 millions different positions, in the 1.5 million nucleotides chromosome IV of yeast *Saccharomyces cerevisiae*. These figures increase dramatically when considering repeat structures within either a single larger genome or between several genomes.

As stated above, the second key problem lies on the manner to display, in a comprehensive way, the organization of repeat structures.

Our project aims at targeting these two key issues by providing a dedicated visualization system able to display hierarchical summaries of repeat structures at genome level.

Starting from the localization of maximal repeats, we want to use them to get all meaningful segmentations of the sequence. A possible resulting structure on the sequence would be a partition into non overlapping  $v$ -modules. However, we have to take into account two properties of  $v$ -modules.

The first property is that some domains may be included in other ones. Moreover, depending on the accepted distance between the domain word model and its occurrences in the sequence, this natural hierarchical structure on the set of domains is enlarged: if more errors are allowed, larger domains are

found. It is possible to build this hierarchical structure from no-errors repeats at the bottom level to the whole sequence at the top level, thus proposing several levels of summaries of the genome at hand.

The second property is that  $v$ -modules may overlap in the sequence (periodic structures).

A tree data structure (or simply a hierarchy on the set of positions) is not sufficient then. We propose to build instead a pyramidal structure [Diday 86]. This slightly more general structure is possible due to maximal repeats properties. It allows describing overlapping structures, but keeps the requirement of non crossing structures, thus leading to clear visualizations. To our knowledge, there exists just one attempt to use pyramids in the context of the analysis of biological data [Aude & al. 1999].

Our project will propose a graphical browser for in-depth interactive exploration of pyramidal structure of  $v$ -modules. This browser will be able, on the one hand, to handle huge amount of data, and, on the other hand, will provide the required functionalities for such an exploration.

## 2.3 $v$ -modules patterns scanning

Modelling sequences is an important task in bioinformatics since it is used to capture the common properties shared by functionally and/or structurally related sequences. Such models are very useful, not only to summarize and to represent sequence properties, but moreover to scan sequence database looking for new members. Modelling sequences can be achieved in three main ways using either a prototype sequence, or a hidden Markov chain or a formal language.

For modelling purpose, the first approach, proposed by BLAST [Altschul et al., 1997], is very limited since a prototype, i.e. one sequence chosen from a family, does not provide all the properties of that family. The Markovian approach [Durbin et al., 1998] is the most efficient approach to day but suffers from the need to create the Markov model given a multiple sequence alignment of the family members, and results in a model (either a probability matrix or a probabilistic automata) difficult to understand. Actually, the Markov model is a black box predictor mainly designed for database scanning purpose.

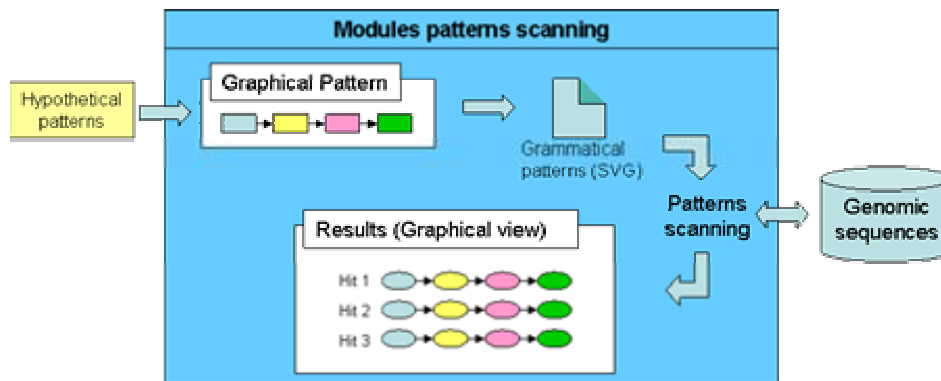
The formal language approach aims at explicitly designing properties of sequences. Regular expressions are widely used, especially for protein sequences [Kucherov and Rusinowitch, 1995; Altschul et al., 1997; Gatiker et al., 2002], since they provide a good balance between syntactical simplicity and good level of expressiveness. When considering nucleotidic sequences, that involves repeats mechanisms such as copy or palindrome, and even more complex structural properties, more expressive formalisms are available, such as those deriving from Definite Clause Grammars (DCG) [Pereira and Warren, 1980], a particular form of context-free grammars. DCG have been used in various works to model DNA sequence features [Searls, 1989; Helgesen and Sibbald, 1993; Leung et al., 2001], as well as to model gene regulation [Collado-Vides, 1992]. Among those formalisms, the pioneering work of David Searls on String Variable Grammars (SVG) is of particular interest [Searls, 1995; Searls, 2002]. SVG introduces the concept of a variable that can be associated to a string during a pattern search. SVGs can be used to model not only DNA/RNA sequence features, but also structural features such as repeats, palindromes, stem-loop or pseudo-knots. For the purpose of the Modulome project, we will provide a  $v$ -modules modelling approach base on SVG, since it provides the expressiveness required for modelling complex structures on DNA.

For databases search purpose, we meet a paradox, which is the less expressive power, the most efficient the matching algorithm. In the context of SVG-based model, which has high expressiveness, we target searching problems of exponential complexity. To our knowledge, the only two tools capable of searching for SVG-based patterns in biological sequences are GenLang [Dong and Searls, 1994] and PatScan [Dsouza et al., 1997]. However, GenLang is no longer maintained and, because of its time complexity, was restricted to the analysis of medium size sequences (several Mbases). PatScan, on the other hand, does not guarantee to find all occurrences of complex patterns and exhibits in some cases a degraded behaviour. Recently, our team has designed STAN (Suffix Tree ANalyser; Nicolas, 2005) a very efficient pattern matching algorithm, capable of searching a subset of SVG-

based models within entire genomes. To achieve high performance, in time and space, STAN uses a suffix tree representation of genomic sequences. However, STAN is not designed to search for models of v-modules.

Our approach in v-modules modelling will be to go far beyond the current capabilities of STAN by providing the biologists with a modelling system combining expressive models, high performance parser and easy-of-use. Regarding expressivity, we will provide an extended SVG formalism, aims at modelling v-modules. Concerning high performance parser, we will design a model parser based on the ReMIX architecture: suffix trees will be indexed in its memory, whereas searching operators will be running on its reconfigurable processors. Regarding ease-of-use, since SVG is a logic programming formalism, it cannot be provided ‘as is’ to the biologist: we will propose a graphical modelling environment allowing (1) the creation of models without any programming skills, (2) the launch of models execution against databases, and (3) the visualization of the results. A prototype of that graphical platform has been designed in the context of SVG-based protein modelling (Durand et al., 2005). We will have to extend that prototype so that (1) it can be used in cooperation with the v-modules visualization tool, (2) it exploits the ReMIX architecture as the database parser, and (3) it allows users to work with entire genomes.

The following figure sketches the overall architecture of the modelling platform.



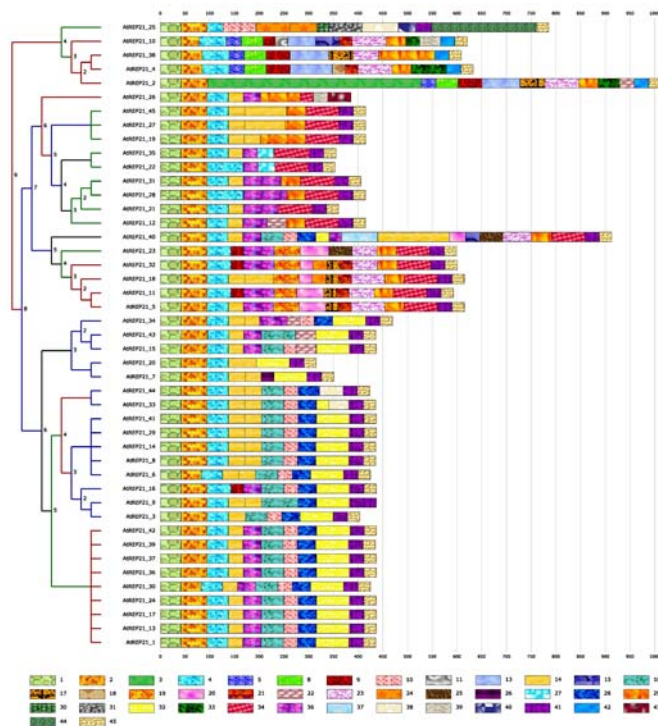
**Figure 3: Modules patterns scanning**

## 2.4 An example of v-module analysis: the AtREP21 transposable element family in the genome of *Arabidopsis thaliana*.

Transposition is a multiplication mechanism capable of transferring DNA segments (transposable element) from one genomic site to another one [Feschotte et al., 2002]. They are characterized and classified into two main classes on the basis of terminal or subterminal remarkable structures (class 1) or of their protein-coding capacity (class 2).

The elements of class 1 move via an RNA intermediate, and show LTR (Long Terminal Repeats) at each extremity. Elements of class 2 move via “cut-and-paste” mechanisms on DNA elements, and are characterized by TIR (Terminal Inverted Repeats). Some transposons, called non-autonomous elements [Wessler et al. 1995; Feschotte, Mouches 2000], do not show any coding capacity and are difficult to characterize: their internal sequences do not have a consensus sequence (Feschotte and Mouches, 2000).

Currently, most studies fail to characterize exhaustively the transposable element family organization. More particularly, at the sequence level, extracting automatically the right “architecture” of such complex repeats is still a challenge. The non-autonomous transposable element family, called AtREP21, is a good example. We study this family in the whole genome of the model plant *Arabidopsis thaliana*. This family is a real illustration of how complex is the internal transposon architecture. As shown on the figure below, it clearly demonstrates the existence of a complex assembly of v-modules, even for non coding sequences. The AtREP21 family shows a significant variation of sequences resulting from insertions, deletions or substitution of domains.



### **Visualization of domains and classification of the AtREP21 family:**

We found 48 AtREP21 sequences in the whole *Arabidopsis thaliana* genome. These elements range from 315 bp to 1012 bp.

The sequences have been first selected based on their terminal repeat. Then we performed a multiple alignment, followed by a specialized v-module detector.

The Modulome approach would first detect a combination of v-modules allowing the biologist to characterize this family of transposable elements.

To identify the elementary modules, a combination of language analysis, alignment and classification tools have been used. This method, mainly based on manual parameter tuning is a clear limitation for large scale analyses. In that specific case, the Modulome approach would shorten the analysis by an immediate detection and visualization of these complex structures.

We now describe three large scale case studies that we will develop during the project, all concerned with mobile genetic elements.

## 2.5 First case study: transposable elements in vertebrates and viruses

The objective of this study will be to characterize the mobility of some exons and genes mediated by Miniature Inverted repeat Transposable Elements (MITEs) in vertebrate genomes and in the genomes of the large double-stranded DNA viruses.

### *State of the Art*

MITEs are a peculiar kind of short repetitive DNA elements that are major components of eukaryotic genomes: examples have been found in many animal phyla including humans, but they are particularly numerous in plants. They consist of short DNA sequences ranging from 80 pb to 500 bp composed of a non-coding AT-rich core flanked by a pair of 10-50 bp long inverted terminal regions (ITR). The presence of ITRs and of a duplicated target site surrounding the elements are a characteristic feature of Class II transposable elements (DNA transposons), which strongly suggests that the particularly high copy number reached by MITEs is mediated by transposition. Although MITEs share characteristic structural features with class II transposable elements, and despite some recent progress, the precise mechanism of their transposition remains elusive. Indeed, the lack of coding potential in MITEs suggests that they in turn parasitize the transposition machinery of larger and less numerous full-length transposons, on which they presumably rely for mobilization.

There is currently very few data concerning the genomic impact of MITEs. A recent study, however, highlighted the impact of the biological activity of a MITE family on the modular evolution of genes and plants genome, by mediating the genomic mobility of exons and favoring genes duplication [Jiang *et al.*, 2004]. These events occur in the way of pack-MITEs, consisting of a fragment (exons) or a full length gene, flanked by two MITEs or by truncated MITEs: the transposition of pack-MITEs would mobilize the two MITE (fragments) and the sequences between. The main difficulty to locate these pack-MITEs comes from the fact that MITEs can be difficult to detect because they lack coding potential and their sequence is prone to accumulate point mutations, deletions and insertions during evolution. The currently available tools are too unreliable to efficiently detect canonical and degenerated MITEs and pack-MITEs structures in the genomes. The systematic approach of the Modulome, that can explore the various types of repeats involved, should circumvent the requirements of the pack-MITEs analyses.

### *Research program*

We will particularly focus our analyses of the pack-MITEs on the *miHsmar1* model. *miHsmar1* are 80 pb-MITEs related to the autonomous *mariner*-like transposon *Hsmar1*. They are found in the human genome where they reach high copy number (~10 000 copies), interspersed evenly in euchromatic regions of every chromosomes. Sequence comparison with the databases revealed that they are probably mobile in human genome [Buisine *et al.*, 2005]. Our preliminary analyses also revealed that these *miHsmar1* might be involved in pack-MITEs. Our first objective will be to inventory the different pack-*miHsmar1* in the human genome. Our second objective will be to establish a repertoire of the occurrences of these structures in vertebrate genomes in which *Hsmar1*-like and *miHsmar1*-like occur, such as the closely related species *Pan troglodytes* and more distantly related vertebrates (*Ovis aries*, *Bos Taurus*, *Mus musculus*, *Rattus norvegicus* and the bat, *Rhinolophus ferrumequinum*).

## 2.6 Second case study: mobile genetic elements in hyperthermophilic archaea

The objective will be to study the biological significance of a family of regularly spaced repeats derived from foreign genetics elements in hyperthermophilic Archaea genomes. Recently, a new peculiar type of repeated element has been detected in a widening circle of microbial hosts (Mojica *et al.*, 2000). After the sequencing of nearly 230 procaryotic genomes, 16 of them representatives of hyperthermophilic archaea lineages, these repeated elements, known as “clustered regularly interspaced short palindromic repeats” (CRISPRs), might represent the most widely distributed family of repeats among prokaryotic genomes. As the CRISPR spacers likely derive from preexisting sequences, a biological function should be predicted (Mojica *et al.*, 2005).

*The origin, phylogeny and biological significance of CRISPRs arises as an item to be elucidated.*

### *State of the Art*

**Archaea: the third domain of life.** Despite a ubiquitous distribution and a diversity that may parallel that of the Bacteria, the Archaea still remain the most unexplored of life’s domains. Archaea are still primarily considered to be extremists, dominating habitats that define the physical limits for biological systems, such as hot or acidic springs, deep-sea hydrothermal vents, or anoxic systems. From an evolutionary viewpoint, it might be justified to give priority to the extremists: some of archaea from hot environments branch close to the root of the archaeal tree (Barns *et al.*, 1996; Takai *et al.*, 1999; Forterre *et al.*, 2002), indicating that thermophiles might have arisen first on the planet.

**Viruses and plasmids, major components of the biosphere and chief actors in evolutionary processes.** Viruses are now considered one of the major components of the biosphere. So, the universal tree of life can thus be considered as immersed into a virtual viral ocean (Bamford, 2003). The importance of viruses and plasmids in prokaryotic evolution becomes increasingly clear as more and more complete microbial genome sequences have become available. The magnitude of lateral gene transfers (LGT) mediated by these mobile genetic elements (MGEs) is now being revealed from complete genome analysis (Logsdon and Faguy, 1999) as from metagenome analysis (Edwards and Rohwer, 2005). Viruses and plasmids are considered to be central players in mobilizing and reorganizing genes and they are now seen as key players in the reshuffling of genetic material, which in combination with mutations and selection, drive evolution. The discovery of unexpected homologies between viruses infecting cells belonging to all domains of life strongly suggests that the last common cellular ancestor, LUCA, were already immersed into an ancestral virosphere. Forterre recently proposed that DNA itself originated in such ancestral virosphere (Forterre, 2002). Plasmids also appear to be very ancient. The existence of many homologous proteins, present in both plasmids and viruses but absent from most cellular genomes, suggests that plasmids are derived from ancient viruses that have lost their capsids (Forterre *et al.*, 1992).

**Mobile Genetic Elements in hyperthermophilic Archaea.** Despite the extreme conditions of their habitats and relative isolation of individual populations, hyperthermophilic Archaea have proven to be a surprisingly rich source of MGE. The recent discovery of an amazing richness and novelty of viruses and plasmids in archaea living in terrestrial hot springs (Rachet *et al.*, 2002; Pranghisvili, 2003) as well as in deep-sea hydrothermal vents (Geslin *et al.*, 2003a, b) that have not equivalent in natural environments, suggests that these MGE probably play a key role in the adaptation of their hyperthermophilic hosts in such extreme environments.

The availability of 16 hyperthermophilic archaeal genome sequences has also revealed many new and diverse transposable genetic elements (TE) (Brügger *et al.*, 2002). The complete genome of *Sulfolobus solfataricus* P2 revealed that this archaeal species contains the most IS of all prokaryotic genomes sequenced so far (Blunt and Grogan, 2005). Other species show examples of complex clusters of interwoven elements (Redder *et al.*, 2001) while some of them exhibit a very high level of truncated IS elements that exacerbate the problem of their detection and that present the disadvantage of requiring substantial computing power because of the lack of suitable bioinformatics tools.

## ***Research program.***

The bioinformatics techniques available for genome exploration, particularly the detection of repeated sequences like CRISPR, are still unsatisfying. It results clearly from a combination of limitations from both of the search algorithms and the diversity represented in sequence databases. Considering CRIPR, this problem should be partly alleviated as more viral and plasmid sequences are sampled and characterized.

We propose to use the viral and plasmidic genome sequences reservoir from hyperthermophilic archaea as a model for validating “the Modulome approach” within the context of CRISPR detection in the archaeal domain.

We will focus on viruses and plasmids present in hyperthermophilic archaea because (i) they branch close to the root of the archaeal tree, (ii) data already available indicate that archaeal viruses are much more diverse than bacteriophages in term of genomic diversity (Prangishvili, 2003) (iii) to our knowledge, we are presently the only laboratory (in collaboration with P. Forterre’s laboratory in Orsay) in the world working actively on viruses and plasmids from marine hyperthermophilic archaea (Priour et al., 2004) (iv) our laboratory possess a great collection of plasmids and viruses of hyperthermophilic archaea, several viral and plasmidic genomes being currently sequenced and analysed in our laboratory which has access to sequencing facilities and expertise in comparative genomics and phylogenomics (Erauso *et al.*, 1996, Gonnet *et al.*, 2005). Our laboratory has good records in genome sequencing and analyses: it has been involved with the Forterre’s group and the Genoscope in the sequencing of the complete genomes of two hyperthermophilic archaea : *Pyrococcus abyssi* (Cohen *et al.*, 2003) and *Thermococcus gammatolerans* (publication in preparation), two strains previously isolated and characterized in our laboratory (Erauso *et al.*, 1993; Jolivet *et al.*, 2003).

## **2.7 Third case study: *de novo* transposable element identification**

### ***State of the Art***

Transposable elements (TE) are mobile, repetitive DNA sequences that constitute a structurally dynamic component of genomes. TE have been found in nearly all eukaryotic organisms studied and represent an important proportion of genome sequences (e.g. 44% of human genome). There is no doubt that modern genomic DNA has evolved in close association with TE. The forces controlling the dynamics of TE spread within a species are poorly understood, as are the systemic effects of the elements on their host genomes. Insertions of individual TE may lead to genome restructuring (e.g., inversions), mutations in genes or changes in gene regulation. Some TE insertions may even have become domesticated to play roles in normal functions of the host. Despite their manifold effects, abundance and ubiquity we understand very little about most aspects of TE biology.

By detailed comparison of the abundance and distribution of TE in entire genomes, we can infer the fundamental biological properties of TE that are shared or that differ among species. However, meaningful inferences about TE biology based on computationally-derived TE annotations can only be done if we are confident about the results of these analyses.

To elevate the quality of TE annotations, we have developed a combined evidence TE annotation pipeline analogous to systems used for gene annotation, by integrating results from multiple homology-based and *de novo* TE identification methods (Quesneville *et al.* 2005). This leading approach differs from standard efforts to annotate TE in genome sequences that rely on the results of a single computational method. Moreover, our system is designed for use with a genome annotation tool, allowing results to be manually curated to produce reliable annotations. Following on the annotation of the euchromatic transposable elements in *D. melanogaster* genome (Release 4) which have been incorporated in the *Drosophila* community database (FlyBase), we are now engaged into several international collaborations to annotate TEs in the *D. melanogaster* peri-centromeric heterochromatin (the *Drosophila* Heterochromatin Genome Project, G. Karpen), *D. simulans* and *D. yakuba* (M. Ashburner and C Bergman), *A. thaliana* (V. Colot, N Buisine).

A good TE annotation relies critically on an expertly assembled reference sequence set, data that currently cannot be obtained in an automatic fashion. The task to assemble such reference set will be most difficult in genomes where only few TE families are known. In these situations, we will need good *de novo* TE detection procedures. In general, the problem of TE discovery remains a major challenge for TE annotation. Various approaches have been developed such as RECON (Bao and Eddy, 2002), TE-HMM (Andrieu *et al*, 2004), PILER (Edgar and Myers, 2005), and LTR\_STRUC (McCarthy and McDonald, 2003). Our recent studies (Quesneville *et al* 2003, Quesneville *et al* 2005) indicate that they suffer from high false positive rate and they are not enough sensitive.

### **Research Project**

We propose a strategy that could hopefully detect “unknown” elements, *i.e.* that operates without *a priori* knowledge over the TEs that are to be found in the sequence. We want to extend our combined-evidence approach to *de novo* TE identification. Our pipeline will use evidences derived from TBLASTX, all-by-all BLASTN, RECON, TE-HMM, PILER, LTR\_STRUC, to get a complementary view of the Modulome tools. We will design our system to visualize a synthesis of all these evidences in the Apollo annotation tool (Lewis *et al.* 2002).

The key problem of *de novo* TE identification is the nested structure of the repeats in a genome. Many TEs are characterised by direct (Long Terminal Repeats, LTR) or indirect (Terminal Inverted Repeat, TIR) repeats at the ends of its sequence. Moreover when it inserts, it induces a small direct repeat, called “target site duplication” (TSD), at its insertion site. The TSD signs the transposition event and is an important feature for TE identification. But these repeat features are often wiped because of the abundant deletion events that affect TE copies in a genome. Moreover these features are also often hidden inside of more complex repeated structures. Indeed, ETs are often found embedded in larger repeats called segmental duplications or in other TEs. The nested structure of these repeats is difficult to visualize and to interpret in order to extract from them true TEs and not mosaics of repeats. The modulome project will provide tools to analyse such structures. Then, it will be possible to model the repeated structure of a TE (LTR, TIR, and TSD) and to search it in genomes.

In order to develop, test, and improve our *de novo* TE detection procedures, we need high quality TE annotations to be used as a benchmark. We think that our *D. melanogaster* TE annotations could serve to further the development and refinement of our TE discovery and annotation methods. Once the pipeline tested, we will use it to annotate TEs in *D. simulans* and *D. yakuba* genomes for which very few TE are currently known.

## **2.8 Scientific animation**

We plan to organize the animation of a scientific group that will work on various points of the project and will contribute to the overall quality of its results. Financing of the team Symbiose in the project includes the functioning of such a group, particularly, the organization of one annual meeting. In Bioinformatics we have contacted or intend to solicit: Lirmm (E. Rivals), SBR Roscoff (O. Collin), Inra Toulouse (T. Farault), Helix (A. Viari), Université de Paris VII (S. Verlan). In Biology, we have contacted or intend to solicit: Unité de génétique moléculaire des levures Département "Structure et dynamisme des génomes" Institut Pasteur (B. Dujon), URA 2171 ABI, Paris (E. Rocha), Laboratoire de Biométrie et Biologie Evolutive Lyon (C. Biéumont), Genoscope Atelier de génomique comparative Evry (C. Médigue), Ecobio Rennes (I. Couée), Institut de Génétique et Microbiologie IFR 4115 Orsay (P. Forterre).

To our knowledge, no equivalent project exists at the European level. Achievements of the project could lead to the preparation of such a European project in the same research field.

### 3. Intended results

We give in this section a detailed work plan for our research program, including a list of deliverables.

---

#### Symbiose, Rennes

---

We will design a dedicated software/hardware platform for the identification, the visualization and the parsing of the structure of genomes in terms of *v*-modules that are repeated along a genome or between several genomes. With respect to computer science, the result will be a better indexing scheme on large sequences, and a parser capable of searching SVG-based models on large sequences. With respect to bioinformatics, we will provide new visualization techniques for analyzing the structure of genomes, as well as new graphical techniques for modelling structures of genomic repeats.

#### Work plan and objectives

Objectives	Works	Main participants	Year 1		Year 2		Year 3	
			01-06	07-12	01-06	07-12	01-06	07-12
Extraction of <i>v</i> -modules	Implementation of data structure	Symbiose participants + contractor engineer	X					
	Implementation of operators	Symbiose participants + contractor engineer		X				
Visualization of <i>v</i> -modules	Prototyping of a first basic viewer <sup>(1)</sup>	Symbiose participants + contractor engineer + 1 <sup>st</sup> Master student	X	X				
	Implementation of a viewer based on reconfigurable hardware	Symbiose participants + contractor engineer + 2 <sup>nd</sup> Master student			X	X		
	Implementation of navigation methods	Symbiose participants + contractor engineer					X	X
Modelling of <i>v</i> -modules	Specification of SVG language	Symbiose participants + 1 <sup>st</sup> Master student	X					
	Implementation of a parser	Symbiose participants + contractor engineer			X	X		
	Implementation of graphical user interface	Symbiose participants + contractor engineer + 3 <sup>rd</sup> Master student					X	X

(1) This viewer will not be based on the reconfigurable hardware.

#### Deliverables

Deliverable N°	Description of the deliverable	Delivery date (month N°)
D1	Specialized data structure for storing genomic sequence on FPGA	6
	SVG-based <i>v</i> -modules modelling language	6
D2	Software for <i>v</i> -modules extraction	12
D3	First basic viewer of <i>v</i> -modules	12
D4	High-performance <i>v</i> -modules viewer based on reconfigurable hardware	24
D5	<i>v</i> -modules models parser	24
D6	Navigation methods for the <i>v</i> -modules viewer	36
D7	Graphical user interface for <i>v</i> -modules modelling	36

---

## LEPG, Tours

---

The results obtained will be used to improve our understanding of the genome plasticity in mammal genomes at two levels. First, the inventory of the *pack-miHsmar1* in several species will allow characterizing the impact of such large mobile structures on genes and genome organization and evolution. Second, the results obtained in human will be used to create a complete inventory of the structures putatively dependent on *Hsmar1* and *miHsmar1* elements for their mobility. This inventory is the first step to design specialized DNA chips dedicated to the analysis of the genome plasticity during human tumor evolution.

### Work plan and objectives

Objectives	Works	Main participants	Year 1		Year 2		Year 3	
			01-06	07-12	01-06	07-12	01-06	07-12
Pack- <i>miHsmar1</i> in human genome	Inventory	LEPG participants + 1st Master 2 student	X					
	Paper writing	LEPG participants		X				
Pack- <i>miHsmar1</i> in other mammal genomes	Inventory	LEPG participants + 2nd Master 2 student			X			
	Paper writing	LEPG participants				X		
Impact of pack- <i>miHsmar1</i> on genome plasticity	Designing chips Some loci Analysis	LEPG participants + 3 <sup>rd</sup> Master 2 student					X	
	Paper writing	LEPG participants						X

### Deliverables

Deliverable N°	Description of the deliverable	Delivery date (month N°)
D1	Report of the first Master 2 student evaluating the efficiency of the modulome for inventory	6
D2	Submission of a first manuscript on <i>pack-miHsmar1</i> in human genome	12-14
D3	Report of the second Master 2 student evaluating the efficiency of the modulome interspecific comparisons	18
D4	Submission of a second manuscript on <i>pack-miHsmar1</i> in mammal genomes	24-26
D5	Report of the third Master 2 student: Designing DNA CHIPS from modulome data	30
D6	Submission of a third manuscript on the evolutionary conservation of <i>pack-miHsmar1</i> between mammal genomes	36-38

---

## LM2E, Brest

---

We foresee that the project will help us to confirm preliminary observations that CRISPRs might arise from viruses and related genetic elements such as plasmids. Moreover, the “Modulome approach” will be powerful to explore the biological significance of the CRISPR elements. It will provide insights into the (genome) composition and structure of environmental archaeal communities and it will enable a deeper understanding of the impact of horizontal gene transfer on microbial

diversity and evolution. CRISPRs spacers will be used to design specialized DNA chips dedicated to the exploration of the vast amount of uncharacterized genetic elements in hot environments.

An additional interest is that this will stimulate future projects in viral metagenomics analyses. Analysis of metagenomes presents substantial computational challenges. The bioinformatics techniques available for virus metagenomics libraries still have severe limitations. Almost all comparisons between metagenomic libraries are currently carried out using sequence-similarity algorithms like BLAST and most sequences present no recognizable similarity to the GenBank database. This problem is exacerbated by the number of genomes in the sample (possibly several millions) and by repeated sequences, such as IS elements, transposons, MITEs, CRISPRs. To circumvent these problems, future analyses of metagenomic sequences should include various word content analyses. The “Modulome approach” is promising in this respect.

## Work plan and objectives

Objectives	Works	Main participants	Year 1	Year 2	Year 3
Creation of a database of MGE sequences	MGE isolation, sequencing and functional annotation	LM2E participants	X		
		LM2E PhD		X	X
Application of Modulome methods on MGE database	v Module annotation	LM2E participants	X	X	
		2nd Master 2 student		X	
		LM2E PhD		X	
EGMs Diversity and biogeography	Design of “CRISPR” DNA chips	LM2E participants			X
		3 <sup>rd</sup> Master 2 student			X
		LM2E PhD			X

First objective relies on genomic DNA sequencing, as reported in the following table. Sequencing efforts will cover the 3 years of the project but for the clarity of the above table, we have only reported the first year.

Year	EGM isolation and sequencing	Functional annotation	v Module annotation	Design of “CRISPR” DNA chips
1	10	10	+ (10V*+ 10P*)	
2	15	10	++ (2V°+15 P°)	+
3	15	15	+++ (2V°+ 15P°)	++

V = virus; P= plasmid; \* = already available at the beginning of the project; ° = under isolation

## Deliverables

Deliverable N°	Description of the deliverable	Delivery date (month N°)
D1	evaluating the efficiency of the modulome tool set for inventory CRISPR element in archaeal genome	6-12
D2	Report of the first Master 2 student evaluating the efficiency of the modulome tool set for inventory conserved and /or degenerated CRISPR elements in procaryotic genome	12-18
D3	submission of a first manuscript on the detection of CRISPR elements in hyperthermophilic Archaea by the Modulome approach	18-20
D4	Report of the second Master 2 student evaluating the efficiency of the modulome to detect others unrelated repeated elements (ISs, MITEs, transposons) in hyperthermophilic prokaryotes Designing DNA CHIPS from modulome data	24-30
D5	Set up of CRISPR-based DNA chips to explore the viral genomic diversity . Submission of a second manuscript on the phylogenetic relationships between CRISPR elements in hyperthermophilic prokaryotes and MGEs and their functional and evolutionary significance	36

---

## LDGE, Paris

---

The Modulome project will provide tools to analyse nested structure of repeats in *Drosophila* genomes. We will model the repeated structure of a TE (LTR, TIR, and TSD) and search it in genomes. In order to develop, test, and improve our *de novo* TE detection procedures, we need high quality TE annotations to be used as a benchmark. We think that our *D. melanogaster* TE annotations could serve to further the development and refinement of our TE discovery and annotation methods. Once the pipeline tested, we will use it to annotate TEs in *D. simulans* and *D. yakuba* genomes for which very few TE are currently known.

### Work plan and objectives

Objectives	Works	Main participants	Year 1	Year 2	Year 3
TE annotation prototyping	Benchmark of methods <sup>(1)</sup>	LDGE participants +	X		
	TE characterization <sup>(2)</sup>	PhD or Post-doc LDGE participants +	X		
	Choice of methods to integrate in the pipeline	PhD or Post-doc LDGE participants +	X		
First annotation	Creation of first pipeline prototype	LDGE participants +		X	
	Annotation of <i>D. yakuba</i> TEs	PhD or Post-doc LDGE participants +		X	
Second annotation	Annotation of <i>D. simulans</i>	PhD or Post-doc LDGE participants +			X

<sup>(1)</sup> Tests of efficiency of the different methods on the *D. melanogaster* annotation used as a benchmark for a blind test.

<sup>(2)</sup> Determination of the key features to be identified by Modulome tools.

### Deliverables

Deliverable N°	Description of the deliverable	Delivery date (month N°)
D1	Methods to integrate in the pipeline, TE models	12
D2	First pipeline prototype	18
D3	Annotation of <i>D. yakuba</i> TEs	24
D4	Annotation of <i>D. simulans</i>	36
D5	Publication of a paper describing the TE identification pipeline	36

## 4. References

- Achaz G, Coissac E, Viari A, Netter P. (2000) Analysis of intrachromosomal duplications in yeast *Saccharomyces cerevisiae*: a possible model for their origin. *Mol Biol Evol*, 17(8), 1268-75.
- Achaz G, Rocha EP, Netter P, Coissac E. (2002) Origin and fate of repeats in bacteria. *Nucleic Acids Res*, 30(13),2987-94.
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25, 3389-3402.
- Aude JC., Diaz-Lazcoz Y., Codani JJ. and Risler JL. (1999) Application of the pyramidal clustering method to biological objects *Comput. Chem.* 23, 303-315. see also the web site <http://195.221.65.10:1234/Pyramids/>
- Bao Z and Eddy SR. (2002) Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res*, 12(8), 1269-1276.
- Bamford, D.H. (2003) Do viruses form lineages across different domains of life ? *Res Microbiol* 154: 231-236.
- Barns, S.M., Delwiche, C.F., Palmer, J.D. and Pace, N.R. (1996). Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *Proc Natl Acad Sci USA* 93: 9188-9193.
- Bejerano G., Pheasant M., Makunin I., Stephen S., Kent WJ., Mattick JS., Haussler D. (2004). Ultraconserved elements in the human genome. *Science*. 304(5675):1321-1325.
- Benson G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27(2):573-80.
- Blount D. and Grogan D. (2005). New insertion sequences of *Sulfolobus*: functional properties and implications for genome evolution in hyperthermophilic archaea. *Mol Microbiol.*, 55(1):312-25.
- Blunt, Z.D. and Grogan, D.W. (2005) New insertion sequences of *Sulfolobus*: functional properties and implications for genome evolution in hyperthermophilic archaea. *Mol Microbiol* 55(1): 312-325.
- Brügger, K., Redder, P., She, Q., Confalonieri, F., Zivanovic, Y. and Garrett, R. (2002) Mobile elements in archaeal genomes. *FEMS Microbiol Lett* 206: 131-141.
- Brosius, J.; Gould, S. J. (1992) On 'genomenclature': a comprehensive (and respectful) taxonomy for pseudogenes and other 'junk DNA.' *Proc. Nat. Acad. Sci.* 89: 10706-10710, PubMed ID : 1279691
- Buisine N et al (2005) Structure, distribution and evolution of the human MITE *mi-Hsmar1*. Submitted
- Cheung J, Estivill X, Khaja R, MacDonald JR, Lau K, Tsui LC, Scherer SW. Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol.* 2003;4(4):R25. PMID: 12702206
- Cheung J, Wilson MD, Zhang J, Khaja R, MacDonald JR, Heng HH, Koop BF, Scherer SW. Recent segmental and gene duplications in the mouse genome. *Genome Biol.* 2003;4(8):R47. Epub 2003 Jul 9.
- Cohen, G.N., Barbe, V., Flament, D., Galperin, M., Heilig, R., Lecompte, O., Poch, O., Prieur, D., Querellou, J., Ripp, R., Thierry, J.C., Van, d.O.J., Weissenbach, J., Zivanovic, Y., and Forterre, P. (2003) An integrated analysis of the genome of the hyperthermophilic archaeon *Pyrococcus abyssi*. *Mol Microbiol* 47: 1495-1512.
- Collado-Vides, J. (1992) Grammatical model of the regulation of gene expression. *Proc. Natl. Acad. Sci. USA*, 89, 9405-9409.
- Dicks J (2000). Graphical tools for comparative genome analysis. *Yeast*. 17:6-15.
- Diday E. 1986 Orders and overlapping clusters by pyramids *Multidimensional Data Analysis Proc.* , ed. J.De Leeuw et al. DSWO Press, Leiden
- Dong, S. and Searls, D. (1994) Gene structure prediction by linguistic methods. *Genomics*, 23, 540-551.
- Dsouza M., Larsen N., Overbeek R. (1997) Searching for patterns in genomic data. *Trends Genet.*, 13(12), 497-498.
- Dujon B, Sherman D, Fischer G, et al. (2004) Genome evolution in yeasts. *Nature*. 430(6995):35-44.
- Durand P, Lavenier D, Leborgne M, Siegel A, Veber P and Nicolas J (2005) Applying complex models on genomic data. *ERCIM News*, no. 60, 47-78.
- Durbin R, Eddy S, Krogh A and Mitchison G (1998) *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press.
- Edwards, R.A and Rohwer, F. (2005) Viral metagenomics. *Nature* 3: 504-510.
- Erauso, G., Reysenbach, A.L., Godfroy, A., Meunier, J.-R., Crump, B., Partensky, F., Baross, J.A., Marteinsson, V., Barbier, G., Pace, N.R., and Prieur, D. (1993) *Pyrococcus abyssi* sp. nov., a new hyperthermophilic archaeon isolated from a deep-sea hydrothermal vent. *Arch Microbiol* 160: 338-349.
- Erauso, G., Marsin, S, Benbouzid-Rollet, N., Baucher, M., Barbeyron, T., Zivanovic, Y., Prieur, D. and Forterre, P. (1996) Sequence of plasmid pGT5 from the archaeon *Pyrococcus abyssi*: evidence of rolling-circle replication in a hyperthermophile. *J Bacteriol* 178: 3232-3237.
- Forterre, P., Benachenhou-Lahfa, N., Confalonieri, F., Duguet, M., Elie, C., and Labedan, B. (1992) The nature of the last universal ancestor and the root of the tree of life, still open questions. *Biosystems* 28: 15-32.

- Forterre, P. (2002) The origin of DNA genomes and DNA replication proteins. *Curr Opin Microbiol* 5: 525-532.
- Forterre, P., Brochier, C. and Philippe, H. (2002) Evolution of the Archaea. *Theor Popul Biol* 61: 409-422.
- Friedman R, Hughes AL (2001) Gene duplication and the structure of eukaryotic genomes. *Genome Res.* 11(3):373-81.
- Gatiker, A., Gasteiger, E. and Bairoch, A. (2002) ScanProsite: a reference implementation of a PROSITE scanning tool. *Applied Bioinformatics*, 1, 107-108.
- Geslin, C., Le Romancer, M., Erauso, G., Gaillard, M., Perrot, G., and Prieur, D. (2003a) PAV1, the first virus-like particle isolated from a hyperthermophilic euryarchaeote, "Pyrococcus abyssi". *J Bacteriol* 185: 3888-3894.
- Geslin, C., Le Romancer, M., Gaillard, M., Erauso, G., and Prieur, D. (2003b) Observation of virus-like particles in high temperature enrichment cultures from deep-sea hydrothermal vents. *Res Microbiol* 154: 303-307.
- Gibbs AJ and McIntyre GA (1970) The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. *Eur J Biochem.* 16(1):1-11.
- Gonnet, M., Erauso, G., Le Romancer, M., Chevèreau, M. and Prieur, D. (2005) Comparative genomics of Thermococcales plasmids isolated from distinct geographic deep-sea hydrothermal vents. Proceedings of the meeting « Thermophiles 2005 », Plaza Surfers Paradise, Gold Coast (Australia), 18th-22nd September.
- Gusfield, D. Algorithms on strings, trees, and sequences. Cambridge University Press, 1997.
- Haas BJ, Delcher AL, Wortman JR, Salzberg SL (2004) DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics.* 20(18):3643-6.
- Helgesen, C. and Sibbald, P.R. (1993) PALM – a pattern language for molecular biology. *ISMB*, 172-180.
- International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature*, 409(6822), 860-921.
- Jeffrey HJ (1999). Chaos game representation of gene structure. *Nucleic Acids Res.* 18(8):2163-2170.
- Jiang N et al. (2004) Pack-MULE transposable elements mediate gene evolution in plants. *Nature* 431: 569-573.
- Jolivet, E., L'Haridon, S., Corre, E., Forterre, P., and Prieur, D. (2003) *Thermococcus gammatolerans* sp. nov., a hyperthermophilic archaeon from a deep-sea hydrothermal vent that resists ionizing radiation. *Int J Syst Evol Microbiol* 53: 847-851.
- Kazazian HH Jr (2004). Mobile elements: drivers of genome evolution. *Science.* 303(5664):1626-32.
- Kucherov, G. and Rusinowitch, M. (1995) Matching a set of strings with variable length don't cares, Lecture Notes in Computer Science, 937, 230-247.
- Kurtz, S. (1999). Reducing the space requirement of suffix trees. *Softw. Pract. Exper.*, 29, 1149–1171.
- Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res*, 29(22), 4633-4642.
- Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C. and Salzberg, S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biology*, 5(2),R12
- Lefebvre A, Lecroq T, Dauchel H and Alexandre J. (2003) FORRepeats: detects repeats on entire chromosomes and between genomes. *Bioinformatics*, 19(3), 319-326.
- Leung, S., Mellish, C., Robertson, D. (2001) Basic Gene Grammars and DNA-ChartParser for language processing of *Escherichia coli* promoter DNA sequences. *Bioinformatics*, 17, 226-236.
- Logsdon, J.M. and Faguy, D.M. (1999) Divergence of the hyperthermophilic Archaea, *Pyrococcus furiosus* and *Pyrococcus horikoshii* inferred from complete genomic sequences. *Genetics* 152: 1299-1305.
- Margaret G.K, Damon R.L, 2000. Transposable elements and host genome evolution. *Trends in ecology and evolution.* 15:95-99.
- McCreight, E. (1976) A space-economical suffix tree construction algorithm. *J. ACM*, 23, 262-272.
- Mojica, F.J., Diez-Villaseñor, C., Soria, E. and Juez, G. (2000) Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Mol Microbiol* 36 (1): 244-246
- Mojica, F.J., Diez-Villaseñor, C., Garcia-Martinez, J. and Soria, E. (2005) Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol* 60: 174-182.
- Nicolas J, Durand P, Ranchy G, Tempel S and Valin A-S. Suffix-Tree ANalyser (STAN): looking for nucleotidic and peptidic patterns in genomes. Submitted to *Bioinformatics*, in revision
- Pereira, F. and Warren, D. (1980) Definite Clause Grammars for language analysis – a survey of the formalism and a comparison with augmented transition networks. *Artif. Intell.*, 13, 231-278.
- Pevzner PA, Tang H and Tesler G. (2004) De novo repeat classification and fragment assembly. *Genome Res*, 14(9), 1786-1796
- Pourcel C, Salvignol G, Vergnaud G. (2005). CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology.* 151:653-63.
- Prangishvili, D. (2003) Evolutionary insights from studies on viruses of hyperthermophilic archaea. *Res Microbiol* 154: 289-294.

- Prieur, D., Erauso, G., Geslin, C., Lucas, S., Gaillard, M., Bideault, A., Mattenet, AC., Rouault, K., Flament, D., Forterre, P. and Le Romancer, M. (2004). Genetic elements of Thermococcales. *Bioch Soc Trans* 32: 184-187.
- Rachel, R., Bettstetter, M., Hedlund, B.P., Haring, M., Kessler, A., Stetter, K.O., and Prangishvili, D. (2002) Remarkable morphological diversity of viruses and virus-like particles in hot terrestrial environments. *Arch Virol* 147: 2419-2429.
- Redder, P., She, Q. and Garrett, R.A. (2001) Non- autonomous mobile elements in the crenarchaeon *Sulfolobus solfataricus*. *J Mol Biol* 306: 1-6.
- Rubinsztein DC, Leggo J, Coetzee GA, Irvine RA, Buckley M, Ferguson-Smith MA. (1995) Sequence variation and size ranges of CAG repeats in the Machado-Joseph disease, spinocerebellar ataxia type 1 and androgen receptor genes. *Hum Mol Genet*, 4(9), 1585-1590.
- Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W (2000). PipMaker--a web server for aligning two genomic DNA sequences. *Genome Res.* 10(4):577-586.
- Searls, D. (1989) Investigating the linguistics of DNA with definite clause grammars. In Lusk, E. and Overbeek, R. (eds), *Logic programming: Proceedings of the North American Conference on Logic Programming*, pages 189-208. MIT Press.
- Searls, D. (1995) String Variable Grammar : a logic grammar formalism for the biological language of DNA. *J. Logic Programming*, 14, 73-102.
- Searls, D. (2002) The language of genes. *Nature*, 420, 211-217.
- Servant F, et al. (6 co-authors). 2002. ProDom: automated clustering of homologous domains. *Brief Bioinform.* 3:246-51.
- Spell R, Brady R and Dietrich F (2003). BARD: a visualization tool for biological sequence analysis. *Proceedings of the 9<sup>th</sup> IEEE Symposium on Information Visualization*, 20-21 October 2003, Seattle (USA), p28.
- Takaï, K. and Horikoshi, K (1999) Genetic diversity of archaeal in deep-sea hydrothermal vent environments. *Genetics* 152: 1285-1297.
- Takemura, M. (2001) Poxviruses and the origin of the eukaryotic nucleus. *J Mol Evol* 52: 419-425.
- Taylor JS, Braasch I, Frickey T, Meyer A, Van de Peer Y. (2003) Genome duplication, a trait shared by 22000 species of ray-finned fish. *Genome Res.* 13(3):382-90.
- Vandepoele K, Saeyns Y, Simillion C, Raes J, Van De Peer Y (2002) The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between *Arabidopsis* and rice. *Genome Res.* 12(11):1792-801.
- Volfovsky N, Haas BJ, Salzberg SL (2001) A clustering method for repeat analysis in DNA sequences. *Genome Biol.* 2(8):RESE